Escuela Politécnica Superior de Linares

Master en Ingeniería de Telecomunicación

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior de Linares

Trabajo Fin de Master

TÉCNICAS DE SEGMENTACIÓN Y SEPARACIÓN DE FUENTES SONORAS APLICADAS A SEÑALES MUSICALES MONOCANAL

Alumno: Juan de la Torre Cruz

Tutor: Prof. D. Pedro Vera Candeas
Depto.: Ingeniería de Telecomunicación
Tutor: Prof. D. Francisco Jesús Cañadas Quesada
Depto.: Ingeniería de Telecomunicación

Febrero, 2017

[AGRADECIMIENTOS]

A mis padres por su paciencia y apoyo, por haberme inculcado los valores del sacrificio y la superación, por haber realizado un gran esfuerzo económico todos estos años. Por haberme dado todo, aunque no tuviesen nada, por haber confiado siempre en mí y por demostrarme que siempre estarán ahí.

A Teresa y Antonio por haberme tratado como a un hijo, preocupándose por el trascurso de mí camino. Por haberme enseñado el verdadero significado de la vida, demostrando el cariño y la fuerza que son capaces de dar. Por enseñarme y darme la oportunidad de conocer, a una madre y a un padre que siempre lo han dado todo.

A Jesús David Roldán Sánchez por haberse cruzado en mi camino, por haber confiado siempre en mí. Por haberme dado la motivación necesaria para conseguir ser lo que soy. Por haberme dado siempre la ayuda que necesitaba. Por darme esas fuerzas que tanto necesitaba. Por haberme tratado como a un hermano, y lo más importante por enseñarme que las cosas que hagamos, los objetivos y las metas que nos propongamos se deben de hacer con el corazón.

A la Escuela Politécnica Superior de Linares y al profesorado por haberme proporcionado los conocimientos en telemática, tecnologías de comunicación y en imagen y sonido.

A mis tutores de este Trabajo Fin de Master, Pedro Vera Candeas y Francisco Jesús Cañada Quesada, por los consejos y la ayuda proporcionada y por haberme ofrecido la oportunidad de participar en varios de sus proyectos como un compañero más.

ÍNDICE

ABSTRACT

This master thesis can be seen as the natural continuation of my grade thesis (SEPARATION AND TRANSCRIPTION OF MUSICAL SIGNALS IN THE FRAMEWORK OF FLAMENCO) in which a system capable of separating any musical signal corresponding to the Fandangos in its three fundamental sound sources (harmonic, percussive and vocal) and to perform the transcription of each one of the sources.

In this master thesis we seek to create a separation system that allows us to obtain the three fundamental sources (harmonic, percussive and vocal) of any musical signal. As a novelty to the grade thesis already developed seeks to generalize the system for any musical genre. In this way an exhaustive study of the algorithms necessary to perform the segmentation and separation of any musical signal will have to be carried out.

The generalized idea of the master thesis is to perform a study of different separation and segmentation algorithms. Thus, we will start with the less efficient algorithms until we arrive at the most efficient algorithm that is the one that allows us to obtain better objective results, in terms of segmentation and separation. Master thesis will be done using the MATLAB environment for its development, optimization and subsequent evaluation of the final system.

Throughout the memory will be detailing the mathematical, physical and musical foundations that have made possible the implementation of the system, giving a generalized approach to musical theory, which will allow to prepare the musical linguistic environment appropriate for this master thesis.

1 RESUMEN

La segmentación y separación de fuentes sonoras aplicadas a señales musicales monocanal es un tema que se encuentra actualmente en auge, en el que investigadores como Zafar Rafii [1], Ozerov [2], Durrieu [3] y Wei Li [4] han dedicado muchos años de investigación para mejorar los resultados obtenidos.

Este Trabajo Fin de Master se puede ver como la continuación natural de mi Trabajo Fin de Grado (SEPARACIÓN Y TRANSCRIPCIÓN DE SEÑALES MUSICALES EN EL ÁMBITO DEL FLAMENCO) en el cual se implementó un sistema capaz de separar cualquier señal musical correspondiente a los Fandangos (palo específico del Flamenco) en sus tres fuentes sonoras fundamentales (armónica, percusiva y vocal) y realizar la transcripción de cada una de las fuentes.

En este Trabajo Fin de Master se busca crear un sistema de separación que nos permita obtener las tres fuentes fundamentales (armónica, percusiva y vocal) de cualquier señal musical. Como novedad al Trabajo Fin de Grado ya desarrollado se busca generalizar el sistema para cualquier género musical. De esta forma se tendrá que llevar a cabo un estudio exhaustivo de los algoritmos necesarios para realizar la segmentación y separación de cualquier señal musical.

Se realizarán distintos sistemas de forma automática y un sistema manual, por lo que se podrá realizar un estudio exhaustivo de los resultados objetivos que permitirá determinar el grado de pérdidas que conlleva realizar un sistema inteligente.

A lo largo de la memoria se irán detallando los fundamentos matemáticos, físicos y musicales que han hecho posible la implementación del sistema, dando un enfoque generalizado sobre la teoría musical, que permitirá preparar el ambiente lingüístico musical apropiado para este trabajo fin de master.

La idea generalizada del Trabajo Fin de Master, es realizar un estudio de diferentes algoritmos de separación y segmentación. De tal forma que se empezara con los algoritmos menos eficientes hasta llegar al algoritmo más eficiente que es el que nos permite obtener mejores resultados objetivos, en términos de segmentación y separación. El trabajo Fin de Master se realizará utilizando el entorno MATLAB para su desarrollo, optimización y posterior evaluación del sistema final.

2 INTRODUCCIÓN

Sin duda, La música es una de las expresiones más fabulosas del ser humano ya que logra transmitir de manera inmediata diferentes sensaciones que otras formas de arte quizás no pueden. La música es un complejo sistema de sonidos, melodías y ritmos que el hombre ha ido descubriendo y elaborando para obtener una infinidad de posibilidades diferentes. Se estima que la música cuenta con gran importancia para el ser humano ya que le permite expresar miedos, alegrías, sentimientos muy profundos y de diverso tipo. La música permite canalizar esos sentimientos y hacer que la persona aliviane sus penas o haga crecer su alegría dependiendo del caso.

Tal como sucede con muchas otras formas de expresión cultural, la música es una manera que tiene el ser humano para expresarse y representar a través de ella diferentes sensaciones, ideas y pensamientos. Así, la música es de vital importancia no sólo por su belleza y valor estético (ambos dos elementos de suma relevancia en lo que respecta al acervo cultural de una comunidad o de una civilización), sino también como soporte a partir del cual el ser humano se puede comunicar con otros y también consigo mismo (ya que la música puede ser disfrutada tanto social como individualmente).

El sistema auditivo humano es capaz de procesar y discriminar los sonidos presentes a su alrededor. Si entendemos la escena auditiva como el conjunto de sonidos concurrentes en el tiempo, que están presentes en el entorno de un punto del espacio, y que pueden ser percibidos por el oído humano, el ser humano puede realizar un análisis de esta escena y establecer una prioridad en función del sonido que más interés le cause. Una persona que escucha un concierto puede identificar el instrumento que lleva la melodía y seguirla mentalmente, o bien, puede elegir los instrumentos acompañantes obviando la melodía, incluso puede centrar su atención sólo en un tipo de instrumento y seguir las notas tocadas, por ejemplo, por la batería.

La tecnología avanza a un ritmo que es difícil de alcanzar, siempre surge un proyecto capaz de desarrollar un servicio que cubra una necesidad o simplemente nos haga la vida más confortable. Para eso está la tecnología hecha por el hombre, para el hombre. De ahí la razón de que se pueda pensar que, si el oído humano es capaz de realizar una separación, de una señal musical, en sus elementos fundamentales (melodía, armonía y ritmo), ¿Por qué no implementar una tecnología que sea capaz de hacerlo mediante una computadora?

El trabajo Fin de Master está asociado a un proyecto de investigación perteneciente al grupo de investigación Tratamiento de Señales y Sistemas de Telecomunicación (TIC-188). Donde se realiza un estudio de diferentes algoritmos de segmentación y separación en señales musicales monocanal.

Por otro lado, el trabajo Fin de Master se puede ver como la continuación natural del Trabajo Fin de Grado (SEPARACIÓN Y TRANSCRIPCIÓN DE SEÑALES MUSICALES EN EL ÁMBITO DEL FLAMENCO). De tal forma que se busca mejorar los resultados obtenidos en términos de segmentación y separación, además de generalizar las bases de datos utilizadas para cualquier género musical. En la figura 2.1 se muestra un esquema general del sistema (segmentación-separación) llevado a cabo. El esquema general está formado por dos etapas bien diferenciadas (segmentación-separación). La primera etapa, segmentador automático, se encargará de detectar la zona instrumental y la zona vocal, obteniendo de esta etapa los tramos temporales donde la zona instrumental está presente. Una vez detectados estos tramos temporales y almacenados en un fichero, la segunda etapa del sistema, separador de señales musicales, utilizara dicho fichero para extraer las características musicales de la zona instrumental y poder realizar una separación de la señal musical en sus tres elementos fundamentales (melodía, armonía y ritmo).

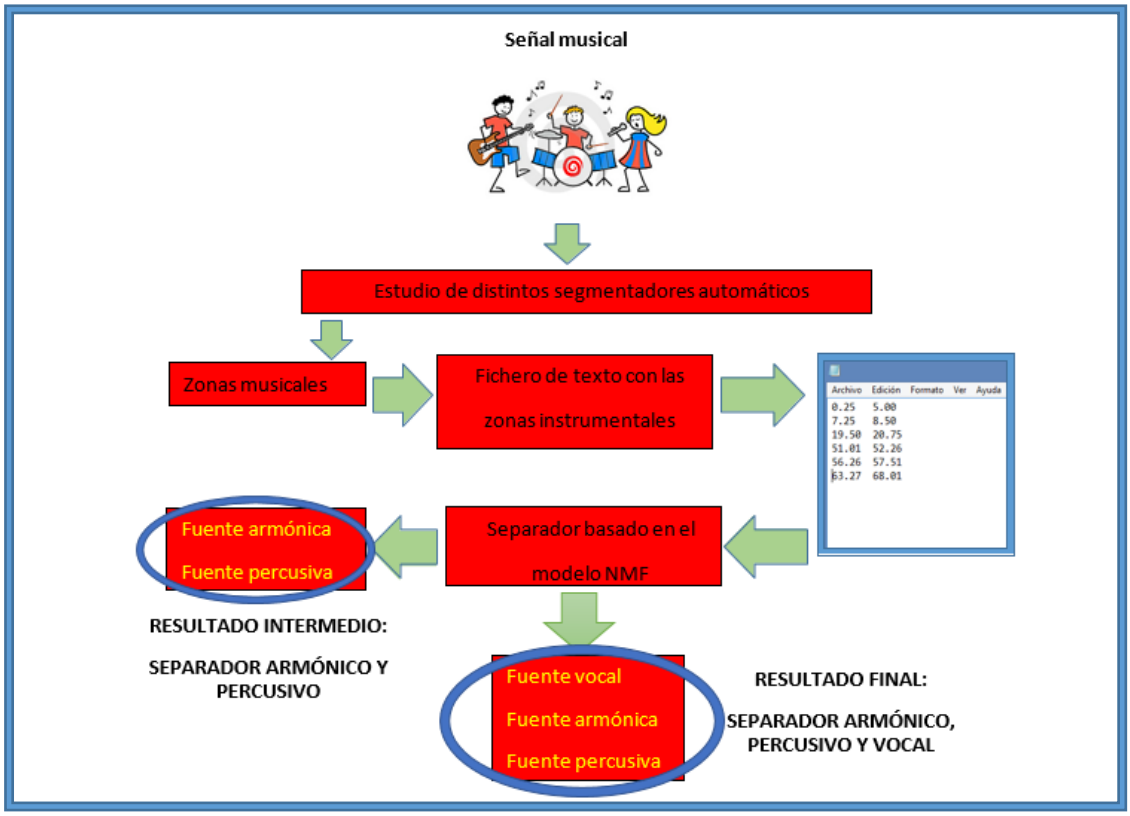

Figura 2.1: Esquema general del Trabajo Fin de Master

2.1 MOTIVACIÓN

Dentro del patrimonio histórico artístico de cualquier cultura, la música es uno de sus pilares fundamentales. Es por ello que no vamos a concretar en ningún género musical, sino que vamos a generalizar el sistema a realizar para cualquier género musical, ya que al fin y al cabo la música es algo que une a todas las culturas. Se podría decir que la música es el lenguaje común de todas las civilizaciones. Un grupo de personas que hablan idiomas distintos no pueden llegar a entenderse, sin embargo, un grupo de personas que escuchan una determinada canción pueden llegar a sentir la misma sensación. Es por ello, que se ha pensado en llevar a cabo un sistema que no se especialice en ningún género musical, sino que se realice de forma genérica para todos los géneros musicales.

La música, como toda manifestación artística, es un producto cultural. El fin de este arte es suscitar una experiencia estética en el oyente, y expresar sentimientos, circunstancias, pensamientos o ideas. La música es un estímulo que afecta el campo perceptivo del individuo; así, el flujo sonoro puede cumplir con variadas funciones (entretenimiento, comunicación, ambientación, etc.).

Este trabajo Fin de Master ofrece la posibilidad de disponer de los elementos principales de una señal musical (melodía, armonía y ritmo) por separado y de forma aislada, con el fin de proporcionar la posibilidad de modelar dichas fuentes sin necesidad de tener previamente los ficheros originales captados de forma individual por un micrófono por cada fuente (fuente vocal, fuente armónica y fuente percusiva).

Esto permite dentro del patrimonio artístico, correspondiente a la música, poder disponer de cualquiera de las fuentes sonoras correspondientes a una de las canciones que fueron un pilar histórico en el mundo de la música, sin la necesidad de haber realizado una grabación en un escenario sobredeterminado [5] (cuando el número de micrófonos es al menos el número de fuentes).

Dicho sistema nos permite explotar al máximo las posibilidades que nos ofrece la música en sus diferentes modalidades, ya que podemos disponer de las distintas fuentes sonoras que componen la canción. En definitiva, un sistema que nos permita escuchar, y sentir cada una de las partes que han conseguido componer esa canción que dejó huella en nuestra historia.

El gran atractivo de este Trabajo Fin de Master se centra en conseguir realizar un sistema, como el que se ha mencionado anteriormente, de forma totalmente inteligente, es decir, realizar un sistema que sea capaz de funcionar de forma automática sin necesidad del ser humano, plasmando la teoría de como el hombre es capaz de distinguir entre la melodía, la armonía y el ritmo. Finalmente se pretende obtener el resultado de separación para el caso de tener anotado las zonas cantadas de las zonas instrumentales, para determinar la mejora de usar un bloque automático de segmentación.

Para la realización de este Trabajo Fin de Master se aprovechan los conceptos estudiados en el Grado de Ingeniería de Tecnologías de Telecomunicación y en el Master en Ingeniería de Telecomunicación, como procesado de la señal, nociones sobre acústica, codificación, transmisión de señales, teoría de la comunicación y programación mediante MATLAB.

2.2 ORGANIZACIÓN DE LA MEMORIA

Los diferentes capítulos indicados en el índice marcan la organización del Trabajo Fin de Master a desarrollar y se van a detallar a continuación:

Punto 2: En el punto 2.3 se realiza una introducción a la teoría musical relacionada con el Trabajo Fin de Master que permite preparar el ambiente lingüístico musical, con el fin de facilitar el entendimiento de la implementación del Trabajo Fin de Master, que al fin y al cabo se basa en las propiedades de la música. Además, se realiza una revisión del estado del arte de las investigaciones científicas en el campo de la segmentación automática y la separación de señales musicales, indicando que se puede encontrar actualmente y partiendo de estos conocimientos.

Punto 3: Detalla los objetivos que enmarcan el proyecto y que deben ser asumidos y llevados a cabo para la correcta finalización del sistema a implementar.

Punto 4: En esta etapa se narra con detalle la implementación de dicho sistema, marcando las tareas requeridas, pautas de trabajo y problemas solventados. El diseño de este sistema consta de dos etapas claramente diferenciadas, la segmentación automática y la separación. A lo largo de este punto se detallarán cada uno de los pasos llevados a cabos para la realización de cada sistema y como se llega al sistema que vamos a considerar como óptimo.

Punto 5: Una vez desarrollado el sistema, en este capítulo se analizarán los resultados obtenidos. Se trata de una evaluación exhaustiva del proyecto realizado, comprobándose que métodos ofrecen una mejor segmentación y separación, y estudiando sus ventajas e inconvenientes.

Punto 6: En esta sección se hará una recopilación global de lo más relevante del trabajo desarrollado y de los resultados obtenidos sacando las conclusiones más certeras del proyecto.

Punto 7: En este punto se abren diferentes líneas futuras para seguir trabajando sobre este aspecto, donde se aportan ideas de mejoras o de extensión.

Punto 8: En este capítulo se incluyen una serie de anexos que facilitan por un lado el entendimiento del Trabajo Fin de Master realizado, y por otro lado se realiza un manual

de usuario sobre la plataforma Matlab para guiar al usuario final sobre el uso del sistema implementado y explicar su funcionamiento.

Punto 9: Incluye la bibliografía de la información usada a lo largo del desarrollo del TFM.

2.3 FÍSICA DEL SONIDO

Desde un punto de vista físico, el sonido es una vibración que se propaga en un medio elástico (sólido, líquido o gaseoso), generalmente el aire. También se podría definir como la sensación producida en el oído por la vibración de las partículas que se desplazan (en forma de onda sonora) a través de un medio elástico que las propaga.

Para que se produzca un sonido se requiere la existencia de un cuerpo vibrante llamado "foco" (una cuerda tensa, una varilla, una lengüeta, etc.) y del medio elástico transmisor de esas vibraciones, las cuales se propagan a través de dicho medio constituyendo la onda sonora.

2.3.1 Modelado del sonido

Un ser humano percibe un sonido cuando el tímpano del oído es puesto en un movimiento característico denominado vibración. Esta vibración es causada por una fuente de sonido que hace pequeños cambios de presión en el aire para propagarse. Cuando un patrón de vibración se repite en intervalos de tiempo iguales se conoce como movimiento periódico. El intervalo de tiempo en el que el patrón de movimiento se repite es llamado período y es denotado por la letra griega tau (τ).

Uno de los movimientos periódicos más simples de visualizar es el de un péndulo. Una característica del mismo es que puede ser representado como la proyección del movimiento de un círculo uniforme sobre un diámetro del círculo, como en la figura 2.2.

A esta proyección se le conoce como movimiento armónico simple y también es llamado movimiento sinusoidal porque puede ser representado por una función trigonométrica llamada seno.

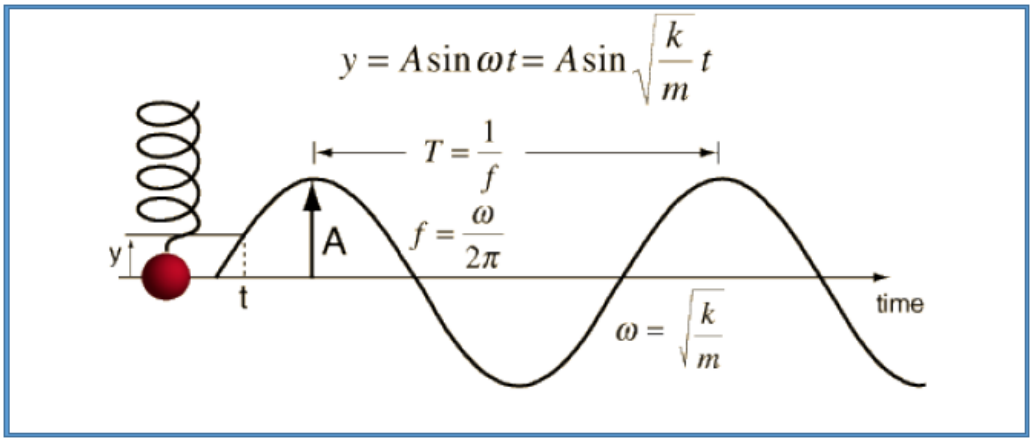

Figura 2.2: Movimiento armónico simple

2.3.2 Percepción del sonido

Una cantidad que es utilizada con mayor frecuencia que el periodo, es la frecuencia:

$$f = \frac{1}{\tau}\ Hz \tag{1}$$

La frecuencia representa el número de repeticiones de un patrón por unidad de tiempo y se mide en hercios (Hz). La razón por la que se prefiere utilizar el término frecuencia se debe a que un aumento de frecuencia es percibido por el humano como un aumento en la agudeza o altura del sonido. Una persona normal puede percibir frecuencias entre 20 y 15,000 Hz.

Un sonido armónico simple con características constantes (frecuencia, amplitud y fase) es denominado como tono puro. La música no está hecha de tonos puros. Sin embargo, para llegar a comprender la manera en la que el ser humano percibe el sonido musical, es recomendable partir de los efectos que se acontecen dentro del oído producidos por tonos puros. En el oído interno se encuentra la membrana basilar, que tiene una longitud de 34 milímetros aproximadamente y alrededor de 30,000 unidades receptoras llamados cilios.

Los cilios son unos pelos minúsculos que entran en movimiento cada vez que una parte de la membrana vibra. Cuanto los cilios actúan, envían pequeñas señales eléctricas al nervio auditivo.

Colectivamente, todo esto forma parte del sonido en nuestro cerebro. Las vibraciones en la membrana basilar están divididas de acuerdo a las frecuencias que componen el sonido que se percibe.

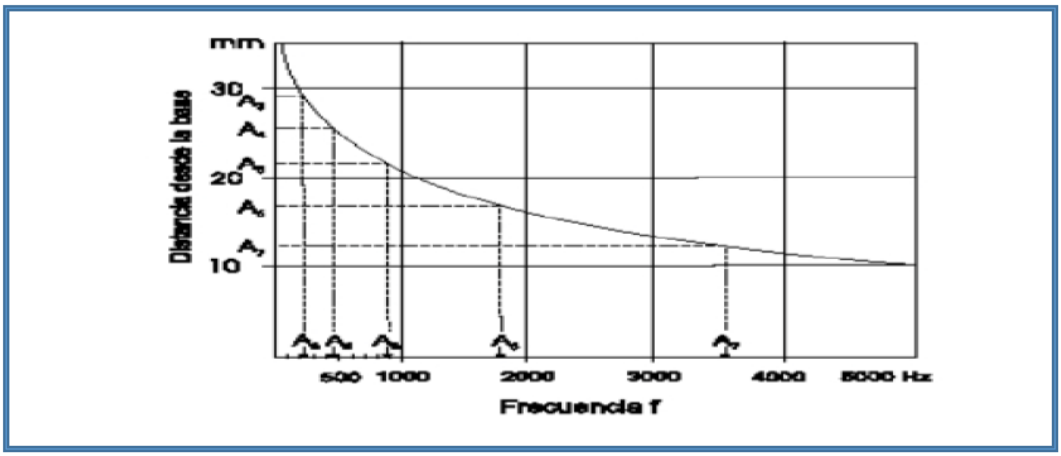

Figura 2.3: Sensibilidad de la membrana basilar

En la figura 2.3, se muestra una gráfica posición (en mm) en la membrana basilar versus frecuencia. Se observa que casi dos tercios de la membrana perciben un rango de frecuencias de hasta 4000 Hz y el tercio restante, la gran porción de frecuencias restantes entre 4000 y 16,000 Hz. Las líneas discontinuas en la frecuencia parten de A3 es decir de 220 Hz y se van duplicando para A4, A5, etc. Como consecuencia, en la región de la membrana basilar la duplicación de frecuencia representa un desplazamiento constante sobre la sección de la membrana que percibe la frecuencia (eje vertical). Una relación como la anterior se describe como logarítmica.

La percepción que tiene el oído humano respecto a pequeñas variaciones de una frecuencia dada depende de la frecuencia. Por ejemplo, para un tono de 2000 Hz, la detección de algún cambio se da a partir de una diferencia de al menos 10 Hz, eso es un 0.5%. Pero para un tono de 100 Hz, la detección empieza ante una variación de al menos 3 Hz, que es un 3%. De lo anterior se hace la importante conclusión que la sensibilidad del oído a cambios en frecuencia es mayor a bajas frecuencias.

2.3.3 Propiedades del sonido

Existen 6 propiedades o atributos que pueden describir lo que un ser humano experimenta al escuchar un sonido. Estas propiedades se verán en primer lugar en un ambiente general para el sonido y en el caso que sea necesario serán enfocadas en términos musicales.

2.3.3.1 Tono o frecuencia

Se refiere a la rapidez de ocurrencia de las vibraciones que componen un tono. Cuando la frecuencia aumenta, también lo hace el tono. La frecuencia se mide en Hertz (Hz) o número de oscilaciones o ciclos por segundo. Cuanto mayor sea su frecuencia, más aguda o "alta" será la nota musical. La altura es una propiedad subjetiva de un sonido por la que puede compararse con otro en términos de "alto o "bajo". En la Tabla 2.1 podemos observar los rangos de frecuencias de las posibles notas generadas por los instrumentos más comunes.

Familia musical	Instrumento musical	Rango de frecuencias (Hz)
Vocal	Soprano	250-1000
	Contralto	200-700
	Baritono	110-425
	Bajo	80-350
Viento	Flautín	630-5000
	Flauta	250-2500
	Oboe	250-1500
	Clarinete (Bb)	125-2000
	Clarinete (Eb)	200-2000
	Fagot	55-575
	Trompeta	90-1000
	Saxofón Alto	125-900
	Saxofón Tenor	110-630
	Saxofón Soprano	225-1000
Metal	Trompeta	170-1000
	Trombon Tenor	80-600
	Trombon Bajo	63-400
	Tuba	45-375
Cuerda	Violin	200-3500
	Viola	125-1000
	Chelo	63-630
	Guitarra	80-630
	Piano	28-4100
Organo	Organo	20-7000
Percusión	Celesta	260-3500
	Timbales	90-180
	Carrillón	63-180
	Xilófono	700-3500

Tabla 2.1: Rangos de frecuencia de las posibles notas generadas por los instrumentos más comunes.

2.3.3.2 Sonoridad

La sonoridad es un atributo de percepción auditiva que permite ordenar los sonidos en una escala desde los más bajos a los más altos en intensidad. La sonoridad no sólo depende de la potencia de un sonido, sino que también depende de su duración y la estructura en tiempo y frecuencia del mismo. Por lo tanto, indica cuan fuerte o suave se escucha un sonido, y se refiere a la amplitud de la onda vibratoria. La cantidad de compresión que sucede en las moléculas de aire, cuando el sonido viaje a través de ellas, no implica la rapidez a la que se desplazan, sino cuanta energía está almacenada en ellas. Sería una propiedad subjetiva y está relacionada con la intensidad (propiedad objetiva), que es el flujo medio de energía por unidad de área perpendicular a la dirección de

propagación. Por lo tanto, la distancia a la que se puedo oír un sonido depende de su intensidad.

Para establecer la sonoridad en frecuencia se establecen unas curvas, denominadas isofónicas, de igual sonoridad con la referencia en 1KHz. Estas curvas son variantes con la frecuencia, por ello dos sonidos con igual potencia, pero distintas frecuencias no tendrán la misma sonoridad, puesto que es un atributo perceptual y no físico. La unidad de medida de sonoridad es el fonio o fon. El umbral de silencio es un claro ejemplo de curva isofónica, debido a que la sonoridad para 1KHz en el umbral del silencio equivale a 3 fonios. En la figura 2.4 se muestran las curvas isofónicas a partir del umbral del silencio.

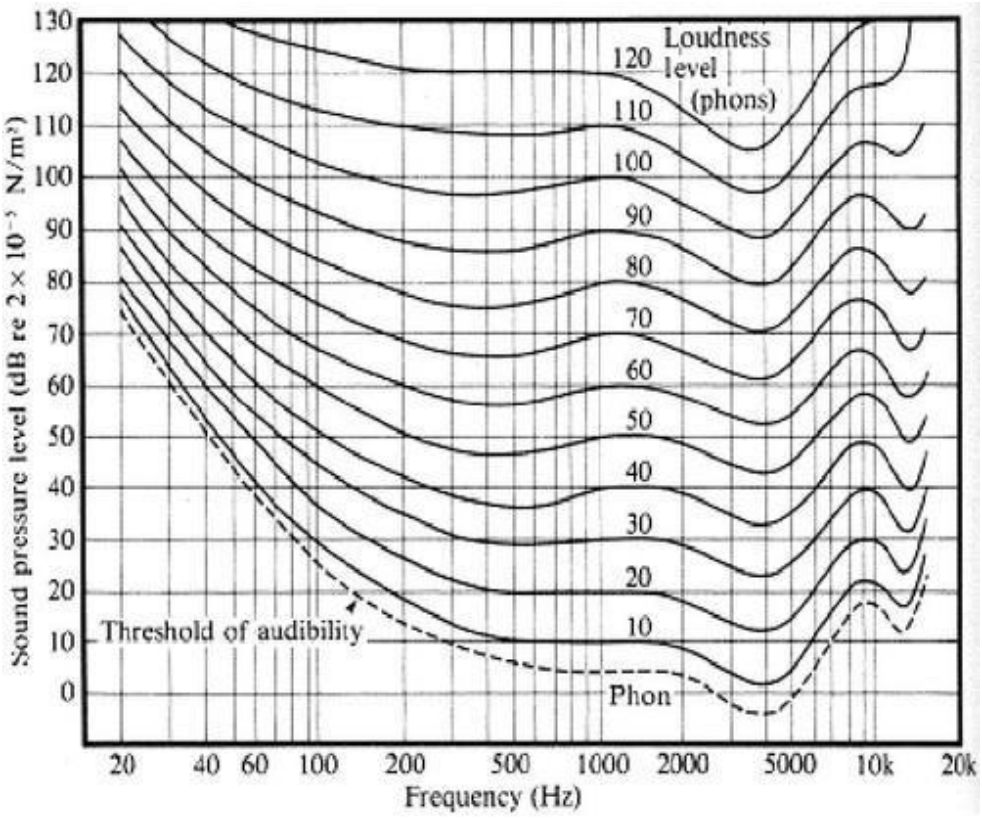

Figura 2.4: Curvas isofónicas

Un concepto musical muy relacionado con la sonoridad es la dinámica musical. La dinámica hace referencia al volumen de un sonido o nota. En la tabla 2.2 se muestra la relación entre sonoridad y dinámica.

Sonoridad de interpretación	Notation	Indicaciones de dinámica
Muy suave	*pp*	pianissimo
Suave	*p*	piano
Medio suave	*mp*	mezzopiano
Medio fuerte	*mf*	mezzoforte
Fuerte	*f*	forte
Muy fuerte	*ff*	fortissimo
Forzando	*sfz*	sforzando
Incremento gradual	<	crescendo
Decremento gradual	>	decrescendo

Tabla 2.2: Relación entre sonoridad y dinámica

2.3.3.3 Timbre

Se refiere a la calidad del sonido. Tanto como los instrumentos musicales y la voz humana producen sonidos que son ricos en vibraciones de diferente frecuencia y que son percibidos al mismo tiempo. La combinación de todas esas vibraciones, con diferentes amplitudes también, es lo que se percibe como timbre.

Una forma sencilla de comprender este concepto es imaginar una flauta y la voz de una persona emitiendo un tono de la misma frecuencia. Fácilmente puede notarse la diferencia entre ambos sonidos. Colectivamente, el cerebro escucha lo anterior como un cambio en el color del sonido.

El timbre es uno de los atributos del sonido más interesantes y complejos de comprender. Pero es en el timbre donde radica la naturaleza del diseño digital del sonido, pues tiene una gran facilidad para cambiar el color y el timbre del sonido rápidamente. Para hacer lo anterior en el mundo acústico, se tendría que cambiar algo del instrumento para cambiar el timbre.

Gracias al diseño digital del sonido, la mayor parte de la música que actualmente escuchamos ha pasado por un proceso digital de refinación. Lo que la electrónica puede hacer, que es tan impresionante, es cambiar fácilmente los timbres sólo manipulando ciertos parámetros.

2.3.3.4 Percepción de duración

La duración no es un elemento fijo, sino algo que se percibe. Un ser humano puede escuchar un sonido, proveniente de cualquier lado, que dura unas pocas milésimas de segundo, o bien, varios minutos. Cuando se habla de lo que se está experimentando, cualquiera puede decir que percibió un tiempo lento o un tiempo rápido. Si se escucha una

canción, ¿cómo se determina si la canción tiene un ritmo lento o rápido? Para ello, existe un punto estándar relativo, como referencia del tiempo, basado en los latidos del corazón.

La duración es el intervalo de tiempo en el que un evento o nota está activo en la señal de audio. Este parámetro está asociado a los términos onset y offset, que indica el instante de tiempo en el que el evento o nota comienza y deja de estar activo respectivamente.

2.3.3.5 Envolvente o articulación

Muestra la forma del sonido en el dominio del tiempo basada en la forma en que se interpreta un instrumento musical, incluyendo a la voz humana. La articulación se refiere a lo que sucede en los primeros milisegundos de un sonido, y a la cantidad de energía que se aplica a la nota. Por ejemplo, con un violín se puede aplicar una gran cantidad de energía al tocar una nota, si se ejerce una presión inicial mayor con el arco sobre la cuerda. Luego, se libera la mayor parte de la presión haciendo que el sonido se estabilice hasta que suavemente desaparece.

Con un piano, la articulación es diferente, pues inicialmente se escucha un martillazo, y luego, la nota que suavemente desaparece. Así, el piano tiene una articulación percusiva, que gráficamente, representa un impulso inmediato.

2.3.3.6 Difusión

Se refiere a capacidad del cerebro para localizar espacialmente la fuente de un sonido. El cerebro tiene la capacidad de procesar simultáneamente múltiples sonidos de diferentes fuentes y, ubicar espacialmente los mismos consciente o inconscientemente. Por ejemplo, si se está en la calle hablando con un amigo, uno le pone mayor atención a la conversación. Cuando se llega a una intersección, inconscientemente se procesan sonidos de ciertos automóviles provenientes de ciertas direcciones. Y en ése momento, se está teniendo la conversación, se están ubicando las fuentes de sonido de los coches, y se está determinando si los mismos se están acercando o alejando.

2.4 TEORÍA MUSICAL

En esta sección de la memoria del Trabajo Fin de Master se detallarán las características fundamentales con respecto a la teoría fundamental y los conceptos apropiados que nos permitirán adentrarnos en este maravilloso arte.

La música es el arte de combinar los sonidos sucesiva y simultáneamente, para transmitir o evocar sentimientos. Es un arte libre, donde se representan los sentimientos con sonidos, bajo diferentes sistemas de composición. Cada sistema de composición va a determinar un estilo diferente dentro de la música.

2.4.1 Elementos fundamentales de la música

La **melodía** es una sucesión coherente de sonidos y silencios que se desenvuelve en una secuencia lineal y que tiene una identidad y significado propio dentro de un entorno sonoro particular. La melodía parte de una base conceptualmente horizontal, con eventos sucesivos en el tiempo y no vertical, incluye cambios de alturas y duraciones, y en general incluye patrones interactivos de cambio y calidad. Por lo tanto, las melodías son las que cantamos o tarareamos cuando un tema nos gusta. No podemos cantar más de una nota a la vez. La melodía es la forma de combinar los sonidos, pero sucesivamente. De ahí que a muchos instrumentos se los llame melódicos, por ejemplo, una flauta, un saxo, un clarinete o cualquier instrumento de viento, porque ellos no pueden hacer sonar más de una nota a la vez.

La **armonía** bajo una concepción vertical de la sonoridad, y cuya unidad básica es el acorde, regula la concordancia entre sonidos que suenan simultáneamente y su enlace con sonidos vecinos.

Hay que tener en cuenta que usando melodías solamente, los temas sonarían “vacíos”. A la larga necesitaríamos algo que nos haga de base, y que nos dé la sensación de estar junto a otros músicos acompañándonos. La armonía es la forma de combinar sonidos en forma simultánea. Cada compositor la usará para crear diferentes climas. Puede transmitir desde estados de melancolía, tristeza, o tensión, hasta estados de alegría, calma, relajación, etc. Los instrumentos llamados armónicos, como el piano o la guitarra, son los que pueden tocar más de una nota a la vez.

Dentro de este elemento musical podemos dar hincapié al concepto de **acorde**. Un acorde en música es una combinación de dos o más notas sonando de manera simultánea

o casi simultánea [6]. Los acordes más frecuentes son las triadas, 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. Concretamente, un acorde consonante es la combinación de notas que producen un sonido agradable para la mayoría de las personas, mientras que uno disonante provocará una sensación molesta o desagradable.

En la Figura 2.5 se puede apreciar claramente la diferencia fundamental entre la melodía y la armonía, ya que en el primer caso no se permiten sonidos que suenan de forma simultánea, sin embargo, en el caso de los armónicos si presentan sonidos que suenan simultáneamente.

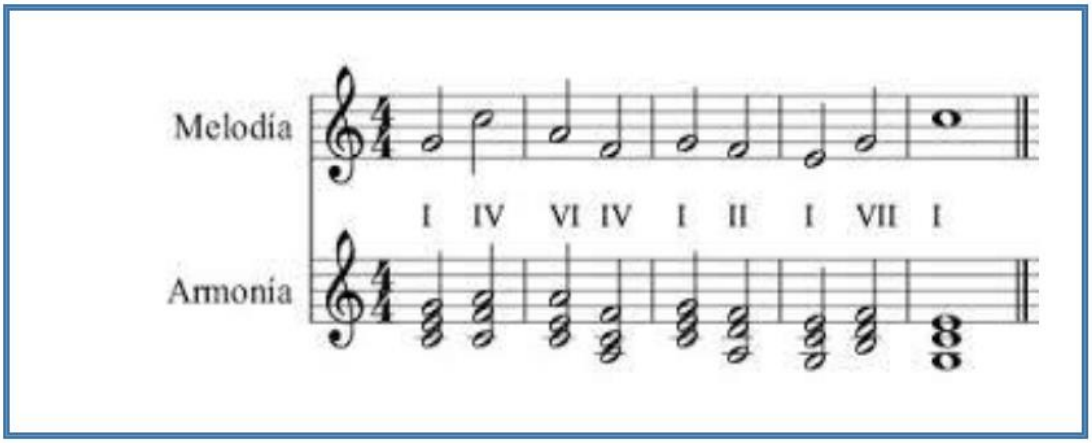

Figura 2.5: Diferencia fundamental entre melodía y armonía

El **ritmo** es la división regular del tiempo. El ritmo está relacionado con cualquier movimiento que se repite con regularidad en el tiempo, en la música se lo divide por medio de la combinación de sonidos y silencios de distinta duración. Cuando estamos escuchando música, es muy común que marquemos golpes de manera intuitiva con el pie o con la mano. A cada golpe lo llamamos tiempo o pulso, y serían las unidades en que se dividen los diferentes ritmos. El ritmo es el pulso o el tiempo a intervalos constantes y regulares. Hay ritmos rápidos, como el rock and roll, o lentos, como las baladas, y podemos diferenciarlos básicamente entre los que son binarios, y los que son ternarios, como el vals. El ritmo está relacionado con el término **"Beat"** con el que trabajaremos a lo largo del Trabajo Fin de Master. El Beat es una unidad básica que se emplea para medir el tiempo. Se trata de una sucesión constante de pulsaciones que se repiten dividiendo el tiempo en partes iguales. Este elemento es en general regular, aunque en algunas obras se presenta el caso de que sea irregular. Asimismo, puede acelerarse o retardarse, es decir, puede variar a lo largo de una pieza musical en función de los cambios de tempo de la misma.

Básicamente se utiliza para marcar el ritmo, es decir, el latido de la música y la utilizamos para comparar la duración de las notas y los silencios.

Los **matices** son la intención, el color o dinámica que se da a la música. Son las diferentes grabaciones que se puede dar a un sonido o frase musical. Son las dinámicas que se aplican para enriquecer el hecho musical. Los matices pueden ser de dos clases: Dinámicos, que tienen que ver con la intensidad de los sonidos y agógicos, relacionados con las duraciones o el tempo de los sonidos. Básicamente los matices caracterizan al músico, podríamos decir que son la característica esencial para poder determinar entre dos cantantes.

Los elementos de la música están directamente relacionados con las cualidades del sonido, ya mencionadas en la memoria del Trabajo Fin de Master, como se puede apreciar en la Tabla 2.3.

ELEMENTOS DE LA MÚSICA	CUALIDADES DEL SONIDO
Melodía	Tono
Armonía	Tono y duración
Ritmo	Duración
Matices	Intensidad y duración

Tabla 2.3: Relación entre los elementos de la música y las cualidades del sonido

Existe un símil muy utilizado en el desarrollo de las clases de teoría musical, conocido como el símil del “edificio musical”. Como se puede apreciar en la Figura 2.6, el ritmo formaría la base de la música, el ático (una única vivienda simulando la presencia de un único sonido) corresponde a la melodía, los apartamentos intermedios (varias viviendas por piso, simulando la presencia de varios sonidos simultáneos) corresponden a la armonía, y por último la fachada (simulando el estilo de cada músico) corresponde a los matices.

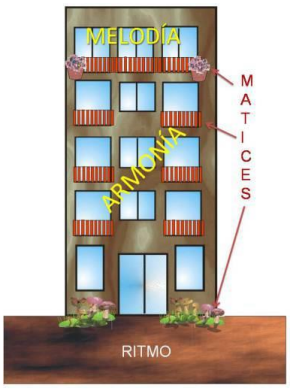

Figura 2.6: Símil del edificio musical

2.4.2 Notación simbólica tradicional

Los sonidos que componen cualquier melodía o composición musical se pueden representar en una partitura como muestra la Figura 2.7. La partitura está formada por una notación simbólica de la misma manera que las letras representan el habla humana. La representación de los sonidos con las notas no es suficiente para describir una composición, es necesario incluir información sobre otros aspectos como la velocidad de interpretación, la intensidad y matices de interpretación para que la melodía quede completamente descrita.

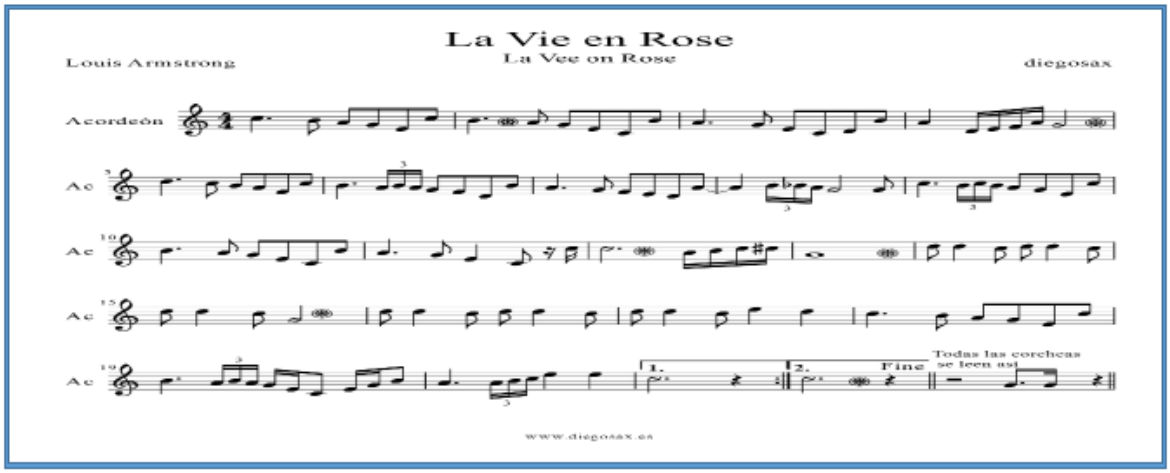

Figura 2.7: Notación simbólica sobre la partitura de la obra "La Vie en Rose" compuesta por Louis Armstrong.

Los elementos básicos que conforman la notación simbólica tradicional son los que se describen a continuación:

Evento (nota): unidad básica en música que representa un sonido con un determinado pitch o frecuencia fundamental (fo) que se encuentra activo durante un determinado intervalo temporal. Las notas musicales se ordenan dentro de las escalas musicales. En la teoría musical occidental se suele dividir cada escala en doce notas (como muestra la figura 2.8): siete notas sin alteración y cinco notas con alteración (símbolos que modifican la entonación), todas ellas ordenadas de grave a agudo. Estas divisiones de doce notas también se denominan escala cromática. En ellas el intervalo entre dos notas adyacentes se denomina semitono.

- Según el sistema de notación latina, las siete notas sin alteración son: Do - Re – Mi - Fa - Sol - La - Si. Las notas con alteración son: (Do# ó Re, Re# ó Mib, Fa# ó Solb, Sol# ó Lab, La# ó Sib) y se generan incrementando o disminuyendo un semitono las anteriores.

- Según el sistema de notación inglesa o denominación literal, las siete notas musicales sin alteración se representan mediante una letra desde la A hasta la G (A,B,C,D,E,F, y G). La nota latina Do es la equivalente a la nota inglesa C, siguiendo esa equivalencia, la relación entre los dos sistemas se puede obtener de manera directa. Por analogía, en el sistema inglés hay cinco notas con alteración que son las siguientes: (o Ab, A# o Bb o C# o Db, D# o Eb, F# o Gb, G#).

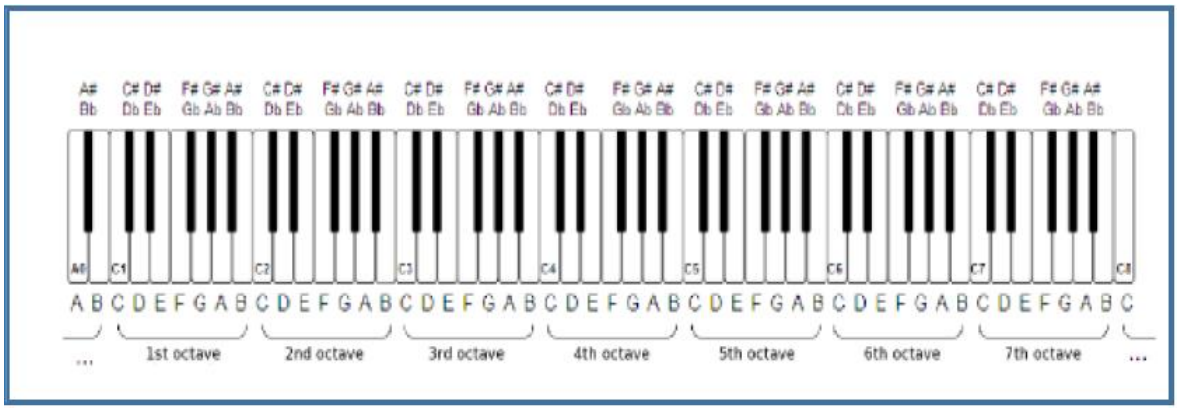

Figura 2.8: Teclado de piano con 88 teclas incluyendo desde la nota A0 hasta la C8. Cada escala se compone de 7 teclas blancas (C, D, E, F, G, A, B) y 5 teclas negras (C# o Db, D# o Eb, F# o Gb, G# o Ab, A# o Bb).

Se denomina **octava** al rango de frecuencias entre dos notas que están separadas por una relación 2:1. La octava a la que pertenece una nota se representa mediante un número que sigue a la letra correspondiente (la nota C4 es la nota C de la escala musical 4 como se puede observar en la Figura 2.8).

Una **escala** es una serie consecutiva de notas que forman una progresión entre una nota y su octava. La escala puede ir hacia arriba o hacia abajo, subiendo o bajando una octava. Cualquier escala se puede distinguir de las demás por su diseño de escalones o grados, es decir, por el modo en que las notas dividen la distancia representado por la octava. El intervalo entre la nota de una tecla blanca y la de la tecla negra siguiente es un semitono. Dos semitonos son igual a un tono.

El **pentagrama** son cinco líneas horizontales y cuatro espacios que se usa para escribir sobre él los símbolos musicales. Por convenio se sigue la misma dirección de izquierda a derecha que en la escritura del lenguaje natural, por tanto, el pentagrama es una línea de tiempo, las notas situadas más a la izquierda se tocan antes que las que se encuentren a su derecha. Cuando el pentagrama se usa para instrumentos armónicos la posición del símbolo indica el pitch o frecuencia fundamental del sonido, cuando se trata

de pentagramas para percusión, cada posición (en el eje vertical) corresponde a un instrumento distinto.

La **clave** es un símbolo musical que se usa para indicar el pitch 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. Su nombre designa a la nota que se sitúe en la línea del pentagrama sobre la que se escriba cada clave.

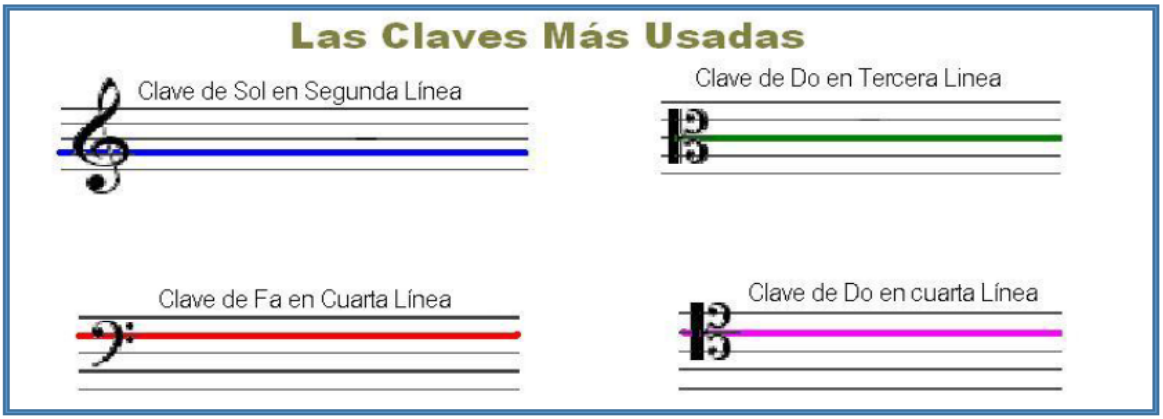

Figura 2.9: Pentagramas con las claves más utilizadas.

La **duración** de cada nota se representa mediante figuras musicales. Cada figura tiene una duración relativa respecto a las demás, por ejemplo, si como referencia de duración se establece la figura negra, el resto de figuras tienen las siguientes relaciones: 20 redonda (4), blanca (2), negra (1), corchea (1/2), semicorchea (1/4), fusa (1/8) y semifusa (1/16), ver Figura 2.10.

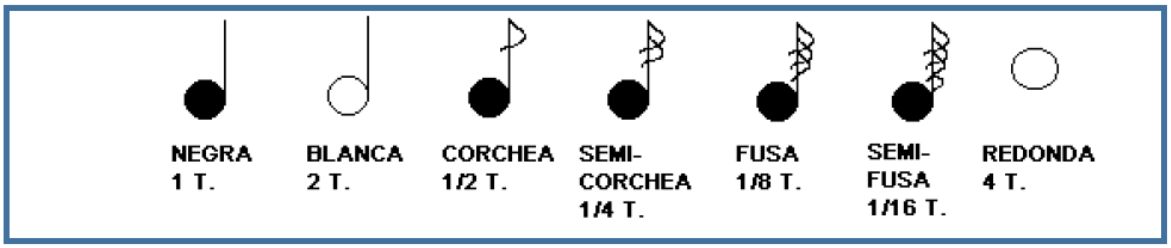

Figura 2.10: Representación de la duración mediante las figuras musicales.

El **pentagrama** se encuentra dividido, en la línea temporal, por unas líneas verticales que dividen la pieza en fragmentos, cada uno de ellos se denomina **compás** y se usan para organizar y hacer más cómoda la lectura de la partitura. Una división más gruesa y doble indica el comienzo y el final de la pieza musical. En ocasiones los compases se marcan con números para facilitar su localización, comenzando con el número 1 para el

primer compás y continuar de manera creciente. Los compases son una herramienta para marcar el ritmo de la música. Al principio del pentagrama se indica mediante dos números en forma de fracción el ritmo del compás (ver figura 2.10). El numerador indica el número de tiempos que tendrá el compás y el denominador indica la unidad de tiempo, es decir, la figura que ocupa un tiempo completo del compás. Por ejemplo, un compás $\frac{4}{4}$ indica que la unidad de tiempo es la negra y que cada compás se compone de 4 negras. Los principales tipos de compas se muestran en la Figura 2.11.

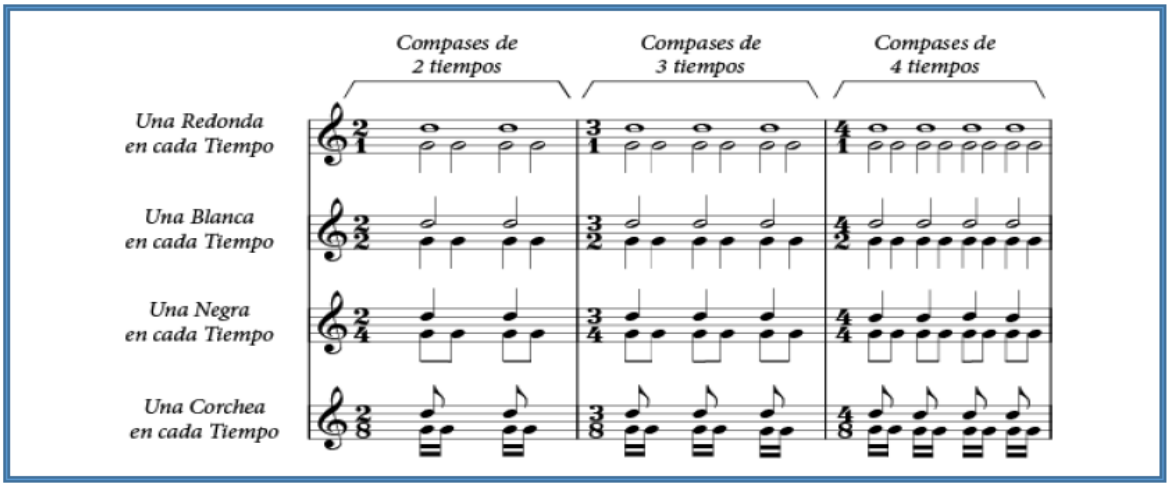

Figura 2.11: Clasificación de los principales tipos de compas.

2.4.3 Parámetros de las notas o eventos musicales

En el contexto del análisis musical cada evento o nota musical puede caracterizarse por tres parámetros básicos.

El **Pitch** es un atributo perceptual que permite ordenar los sonidos en una escala de frecuencia logarítmica, el pitch representa la nota asociada a un evento musical. Un sonido presenta un determinado pitch "si la frecuencia de una sinusoide de amplitud arbitraria puede ser relacionada con el sonido percibido". Esto indica que un sonido tendrá un pitch sólo si ambos son relacionados entre sí por el sistema auditivo humano. El concepto físico asociado al pitch se denomina frecuencia fundamental f0 que se define únicamente para señales periódicas o cuasi-periódicas. El Pitch es una sensación auditiva en el que un oyente asigna tonos musicales a posiciones relativas en una escala musical basada principalmente en la frecuencia de vibración. El Pitch está estrechamente relacionado con la frecuencia, pero los dos no son equivalentes. La frecuencia es un valor objetivo, mientras que el Pitch es un valor subjetivo.

El **Onset** se define como el instante de tiempo en el que se inicia una nota musical.

La **Duración** es el intervalo temporal durante el cual una nota musical se encuentra activa en la señal de audio. Comienza en el instante de onset.

2.4.4 Ordenación de las notas o eventos musicales

En la música occidental, los eventos musicales (notas) se ordenan en una escala logarítmica de manera similar al comportamiento logarítmico del oído humano. Por ello, la frecuencia fundamental, fo (Hz) de cada evento se puede representar como en la ecuación (2) mediante el uso de la escala musical templada (equal-tempered), suponiendo un teclado de piano estándar con 88 teclas.

$$fo(Hz) = fo_{ref} \times 2^{\frac{n}{12}}, \quad n = -48, \dots .,0, \dots ,39 \tag{2}$$

Donde n es el número de semitonos entre la frecuencia de referencia (fo_{ref}), y la frecuencia deseada fo.

El protocolo MIDI (Musical Instrument Digital Interface) es un estándar comúnmente usado para representación de información musical en dispositivos de audio. Este protocolo ofrece una relación entre Hertzios y la numeración MIDI (números naturales). Para traducir una frecuencia fo a notación MIDI se emplea la ecuación (3):

$$N_{MIDI} = \left[69 + 12 \times log2\left(\frac{fo(Hz)}{440}\right)\right], \tag{3}$$

En sentido opuesto, para convertir de la notación MIDI a hercios se emplea la ecuación (4):

$$fo(Hz) = 440 \times 2^{\frac{N_{MIDI}-69}{12}}, \tag{4}$$

Por convenio, la nota MIDI 69 se ha asignado a la nota A4 con $f0$ = 440 Hz. En la Tabla 2.4, se relacionan las frecuencias f0 en hercios a cada nota MIDI para todas las notas musicales de las nueve octavas. Se puede ver que cubren un rango de 7900 Hz. Cada columna representa una escala musical y cada fila es un evento de la escala. En cada celda se muestran, en primer lugar, la frecuencia fundamental (Hz) y, en segundo lugar, entre paréntesis el número de nota MIDI correspondiente.

Notas musicales (Hz - MIDI)									
Nota / Octava	0	1	2	3	4	5	6	7	8
C	16.35 (12)	32.70 (24)	65.41 (36)	130.8 (48)	261.6 (60)	523.3 (72)	1047.0 (84)	2093.0 (96)	4186.0 (108)
C#	17.32 (13)	34.65 (25)	69.30 (37)	138.6 (49)	277.2 (61)	554.4 (73)	1109.0 (85)	2217.0 (97)	4435.0 (109)
D	18.35 (14)	36.71 (26)	73.42 (38)	146.8 (50)	293.7 (62)	587.3 (74)	1175.0 (86)	2349.0 (98)	4699.0 (110)
D#	19.45 (15)	38.89 (27)	77.78 (39)	155.6 (51)	311.1 (63)	622.3 (75)	1245.0 (87)	2489.0 (99)	4978.0 (111)
E	20.60 (16)	41.20 (28)	82.41 (40)	164.8 (52)	329.6 (64)	659.3 (76)	1319.0 (88)	2637.0 (100)	5274.0 (112)
F	21.83 (17)	43.65 (29)	87.31 (41)	174.6 (53)	349.2 (65)	698.5 (77)	1397.0 (89)	2794.0 (101)	5588.0 (113)
F#	23.12 (18)	46.25 (30)	92.50 (42)	185.0 (54)	370.0 (66)	740.0 (78)	1480.0 (90)	2960.0 (102)	5920.0 (114)
G	24.50 (19)	49.00 (31)	98.00 (43)	196.0 (55)	392.0 (67)	784.0 (79)	1568.0 (91)	3136.0 (103)	6272.0 (115)
G#	25.96 (20)	51.91 (32)	103.80 (44)	207.7 (56)	415.3 (68)	830.6 (80)	1661.0 (92)	3322.0 (104)	6645.0 (116)
A	27.50 (21)	55.00 (33)	110.0 (45)	220.0 (57)	440.0 (69)	880.0 (81)	1760.0 (93)	3520.0 (105)	7040.0 (117)
A#	29.14 (22)	58.27 (34)	116.50 (46)	233.1 (58)	466.2 (70)	932.3 (82)	1865.0 (94)	3729.0 (106)	7459.0 (118)
B	30.87 (23)	61.74 (35)	123.50 (47)	246.9 (59)	493.9 (71)	987.8 (83)	1976.0 (95)	3951.0 (107)	7902.0 (119)

Tabla 2.4: Relación entre todas las notas musicales y escalas con la nomenclatura MIDI

2.4.5 Clasificación de los sonidos

Los distintos sonidos presentes en el mundo real se pueden clasificar según diversos criterios, como son el número de canales, grado de periodicidad, número de eventos simultáneos o número de timbres musicales.

2.4.5.1 Sonidos armónicos

Un sonido armónico, (periódico o cuasi-periódico), está compuesto por una serie de sinusoides de frecuencias armónicamente relacionadas entre sí. Esto implica que los sonidos armónicos presenten una estructura espectral donde las componentes frecuenciales están espaciadas regularmente cada $fo(Hz)$.

Un sonido inarmónico no es periódico ni cuasi-periódico. En música hay dos grandes grupos de sonidos inarmónicos: los sonidos producidos con maza (ej. xilófono,

vibráfono) y los producidos con batería (ej. tambor, bombo, caja). El primer grupo se caracteriza por presentar una pequeña periodicidad en el dominio del tiempo, la cual es suficiente para generar claramente un pitch en frecuencia. Por otro lado, los instrumentos del segundo grupo generan sonidos sin ninguna periodicidad en el tiempo, por tanto, no presentan ningún pitch en frecuencia.

En relación al Trabajo Fin de Master los sonidos percusivos (producidos por una batería) son claramente sonidos inarmónicos. Por otro lado, los sonidos producidos por la guitarra son claramente sonidos armónicos. Sin embargo, los sonidos producidos por la voz del cantante no son plenamente armónicos, pero tiene un comportamiento bastante similar (cuasi-armónicos).

En la Tabla 2.5 podemos observar una serie de instrumentos que producen o no sonidos armónicos.

Sonido producido	Familia de instrumentos	Instrumentos
Armónicos o cuasi-armónicos	Instrumentos de cuerda	Piano, guitarra, violin, chelo, etc.
	Lengüeta	Clarinete, saxofón, oboe, fagot
	Viento-metal	Trompeta, trombón, tuba, trompa
	Flautas	Flauta, flautín, etc.
	Órganos de tubos	Órganos de tubos
	Voz humana (cantada)	Cuerdas vocales
No armónicos	Percusión de maza	Marimba, xilófono, vibráfono, carrillón
	Baterías	Baterías y platillos

Tabla 2.5: Clasificación de algunos de los instrumentos de la música occidental como armónicos o no armónicos.

2.4.5.2 Señal monoaural, estéreo y multicanal

Los sonidos monoaurales (también llamados mono) derivan de señales monocanal. Estas señales son las más complicadas de analizar desde el punto de vista del procesado de señal, puesto que no aportan ningún tipo de información espacial de las fuentes. Las señales mono están compuestas directamente por la suma de todas las fuentes de sonido presentes en la grabación.

Los sonidos estereofónicos, derivan en señales de audio de dos canales, son aquellos que son grabados con, al menos, dos micrófonos independientes. Las señales estéreo aportan información espacial de la localización relativa de las fuentes de sonido.

Los sonidos multicanal, también conocidos como sonido surround, son aquellos generados a partir de, al menos, cuatro canales de audio independientes. Por ejemplo, los

sistemas de reproducción surround 5.1 están compuestos por cinco altavoces. Los sistemas 5.1 tienen canales izquierdos y derechos, y además un canal central para conversaciones de las películas o el vocalista en el caso de la música y para efectos especiales y sonido surround. Un canal adicional subwoofer añade sonido de muy baja frecuencia para fuentes musicales y efectos especiales en películas.

2.4.5.3 Sonidos monofónicos y polifónicos

Las piezas musicales pueden ser monofónicas, teniendo un solo instrumento que toque sólo una nota en cada instante temporal (ej. solo de trompeta), o polifónicas si hay uno o más instrumentos que toquen más de una nota en cada instante temporal (ej. Piano o una orquesta).

2.4.5.4 Sonidos monotímbricos y multitímbricos

Dependiendo del número de instrumentos, una interpretación musical puede ser considerada monotímbrica o multitímbrica. El concepto de monotímbrico se relaciona con las piezas musicales que son interpretadas por un instrumento con un único timbre o envolvente espectral. Por otro lado, el concepto de multitímbrico se relaciona con las piezas musicales en las que participan dos o más instrumentos, cada uno de ellos con un timbre distinto.

2.5 ESTADO DEL ARTE

En este apartado, se va a realizar una breve revisión de cómo se encuentra el mundo del segmentador automático y las diversas técnicas relacionadas con la separación de fuentes sonoras para señales musicales monocanal.

2.5.1 Segmentadores automáticos.

Nos adentramos en el mundo de la segmentación, en el cual nunca ha venido investigado por sí solo, sino que siempre ha venido acompañado por otras facetas, que justificaban su investigación. Aunque la segmentación se puede considerar como una rama en la cual se puede hacer una gran profundización, en cuanto a técnicas presentes para su elaboración, siempre tiene que venir acompañado de otro tema que pueda justificar su aplicación final. En el caso del Trabajo Fin de Master la segmentación va acompañada de la separación de las fuentes presentes en una pista musical, y es que sin el sistema de segmentación los resultados tanto subjetivos, como objetivos en cuanto a la obtención de las distintas fuentes esperadas, no serían tan buenos.

A lo largo de las distintas investigaciones que se han llevado a cabo en el ámbito de la segmentación, no existía un objetico único de segmentación, sino que este siempre iba ligado a otra aplicación, como por ejemplo a la transcripción de señales musicales.

La segmentación automática siempre ha jugado un papel crucial en la transcripción de la melodía de una señal musical (fuente vocal), es por eso que en la mayoría de los trabajos la segmentación automática y la transcripción de la melodía están ligados. Dada la grabación de audio de una pieza musical, la tarea de conseguir una correcta transcripción melódica implica una extracción automática de la línea melódica principal (segmentador automático).

Esta extracción automática de la línea melódica principal, es necesaria llevarla a cabo debido a que en la mayoría de grabaciones musicales dos o más notas, procedentes de diferentes instrumentos (por ejemplo: voz, guitarra o bajo), pueden sonar de forma simultánea.

Para realizar una transcripción de la línea melódica extraída debemos tener una clara definición de lo que en realidad es la melodía principal. Como se indica en [7], el término melodía es un concepto basado en musicología, y se pueden encontrar diferentes definiciones para el término melodía en función del contexto [8], [9]. Con el fin de fijar un

marco claro en el contexto de la MIR (Music information retrieval) se ha adoptado la definición propuesta por [7] "… la melodía es la línea principal de la señal musical que el oyente podría silbar o tararear, siendo la esencia de la señal musical en comparación con el resto de las señales presentes".

Se han propuesto muchos métodos para la extracción de melodía. De todos ellos el grupo más grande es el que se podría determinar cómo métodos basados en prominencias, que se derivan de la estimación del tono de relevancia con el tiempo y en una posterior aplicación de unas reglas de seguimiento o transición para extraer la línea de la melodía sin separarla del resto del audio [10], [11], [12], [13]. Tales sistemas siguen una estructura común. Primero se realiza una representación del espectrograma de la señal. Este espectrograma permite obtener una representación tiempo-frecuencia de la función de relevancia. Los picos de dicha función son considerados como candidatos potenciales para la extracción de la frecuencia fundamental fo de la línea melódica. Finalmente, la línea melódica compuesta por las frecuencias fundamentales se calcula utilizando diferentes modelos de selección de pico o de seguimiento.

En relación a los segmentadores automáticos existen sistemas en la transcripción de línea melódica (fuente vocal) que detectan los tramos donde está presente la melodía y donde está presente el acompañamiento musical. Una revisión detallada de tales sistemas se proporciona en [7].

Otro conjunto de enfoques intenta identificar la melodía por separado del resto del audio utilizando técnicas de separación basadas en la caracterización del timbre [14], [15]. Estas técnicas utilizan dos modelos de timbre, uno para la melodía (normalmente voz humana cantada) y el otro para el acompañamiento musical.

Los algoritmos que explotan la información espacial en estéreo también han sido implementados. En [16] se aplica un modelo basado en estos algoritmos para estimar el movimiento de cada fuente, y mediante la utilización de un modelo de producción (fuente/filtro) se permite identificar y separar la melodía (es lo que se conoce como segmentación automática).

Finalmente, en [17] se propone un enfoque basado puramente en análisis de datos. En este modelo se analiza continuamente el espectro en cortos intervalos de tiempo para realizar un entrenamiento que permita clasificar entre zonas instrumentales y zonas melódicas (vocales).

Como se refleja en el estado del arte no existe por si solo ningún artículo dedicado especialmente a la segmentación, sino que todos están liados a otro fin, como puede ser la transcripción de la melodía.

Teniendo esto en consideración, la mayor dificultad para llevar a cabo la transcripción de la línea melódica de forma automática es determinar dónde están presentes las zonas vocales y las instrumentales, en definitiva, desarrollar un segmentador de forma automática. Esto tiene su mayor dificultad en la presencia de las fuentes instrumentales dentro de las zonas vocales. La mayoría de los estudios llevados a cabo sobre la investigación de los segmentadores automáticos llevan siempre el mismo esquema.

Primero comienzan a extraer un conjunto de características de la señal de audio a analizar. Seguidamente utilizan estas características para llevar a cabo una clasificación de las tramas utilizando un umbral o un clasificador estadístico. Según el documento [18], entre las distintas características MFCCs (Mel Frequency Cepstral Coefficients) y sus derivadas son las características más apropiadas para detectar la voz del cantante [19,20] y llevar a cabo una correcta segmentación. Otras características se basan en el estudio de la envolvente espectral, como puede ser en los algoritmos LPC (Linear Predictive Coding), su variante PLP [21] [22] y Warped Linear Coefficients [23]. Estas características son utilizadas para la extracción de la parte vocal, pero tradicionalmente no consiguen describir las características de la fuente vocal en presencia de fuentes instrumentales en la misma zona. También se han utilizado para este fin las siguientes características: Energía [24], Coeficientes armónicos [25] y Vibrato [26]. En cuanto a los clasificadores estadísticos que más se han utilizado para la clasificación de las anteriores características y así determinar las zonas vocales e instrumentales son: GMM [27] y HMM [28].

Para realizar la evaluación del segmentador automático implementado en este TFM nos basaremos de los resultados obtenidos por Justin Salomon y Emilía Gomez en [29], y los obtenidos por Ramona [30], Vembu y Baumann [31] y Bernhard en [32]. En cada caso se utiliza una base de datos específica que se comentara en la sección de resultados. En ambos casos los investigadores involucrados consiguen resultados de segmentación bastante satisfactorios. En la sección de resultados destinados al segmentador se realizará una comparación con los investigadores mencionados.

2.5.2 Estado del arte de los algoritmos de separación de fuentes sonoras

En este caso nos adentramos en un mundo donde todo está bien definido y se han llevado a cabo una gran elaboración de algoritmos dedicados a esta faceta.

2.5.2.1 Introducción

El ser humano, en una situación en la que recibe ondas sonoras generadas por diferentes fuentes, como puede ser una conversación en un ambiente ruidoso, o un concierto de música, es capaz de discriminar la fuente sonora a la que quiere prestar atención aislando el resto a pesar de continuar oyéndolas. Esta tarea de discriminar, siguiendo determinados criterios o patrones, es la que pretende llegar a conseguir la separación de fuentes de audio. La capacidad del oído humano de realizar esta discriminación se denomina Análisis de la escena auditiva (Auditory Scene Analisys, ASA) [33].

La separación de fuentes es una poderosa herramienta para el análisis melódico, rítmico y/o armónico de señales musicales. La investigación en separación de fuentes de audio se aborda en dos sentidos, el escenario sobredeterminado [5] (cuando el número de micrófonos es al menos el número de fuentes) y el escenario infradeterminado [34] [35] (generalmente para señales estéreo o monoaurales).

En grabaciones de música en estudio se puede realizar una captura de audio en la que se obtenga incluso un canal por instrumento de forma aislada. Sin embargo, en grabaciones de música en directo, el número de micrófonos empleados es comúnmente menor que el número de fuentes. En el mejor de los casos se puede realizar una grabación con micrófonos cercanos a cada instrumento, pero aun así el campo sonoro del resto de instrumentos provoca interferencias en el sonido grabado [36]. En estas circunstancias se necesitan fuentes de información adicionales para obtener una separación de fuentes de una calidad aceptable. Así, información musical tal como la partitura alineada [37] o las propiedades tímbricas del instrumento [38] se usan para mejorar el potencial de la separación. Es preciso notar que los algoritmos de separación se implementan incluso en situaciones adversas como con retardo nulo [37] y en procesado paralelo [40]. También es destacable que la separación se puede reducir a extraer la parte armónica y percusiva de los sonidos [39] para realizar un estudio armónico y rítmico más exacto que el obtenido mediante la señal original.

2.5.2.2 Definición

La separación de fuentes sonoras (Sound Source Separation, SSS) pretende recuperar las señales de audio que han sido generadas por varias fuentes y se encuentran mezcladas en una (señal monoaural) o varias señales (señal multicanal). La separación de fuentes no sólo consta de la detección de las notas y su asignación a una de las fuentes, también hay que sintetizar la señal correspondiente a cada una de ellas.

La SSS se puede dividir en dos grandes grupos: la separación de fuentes instrumentales [41] y la extracción de la señal de voz (singing voice) [42]. En la actualidad hay menos trabajos en el campo de la extracción de la señal de voz, aunque hay muchas posibles aplicaciones como el reconocimiento y alineamiento de la letra [43] o la identificación del cantante [44].

Atendiendo a otro criterio, la SSS puede clasificarse en función de la información previa con la que cuente, siendo posibles múltiples combinaciones con distintos tipos de información. La separación que no cuenta con ningún tipo de información se denomina Separación de Fuentes a Ciegas (Blind Source Separation, BSS). A día de hoy, esta separación aporta resultados prometedores en el caso de separación de señales multicanal [5], pero no es así en el caso de señales monoaurales [45]. La BSS ha sido estudiada durante años para las señales monoaurales, pero nunca ha conseguido dar resultados convincentes. La comunidad científica ha apostado por agregar distintas fuentes de información al sistema de separación para que sea capaz de obtener señales separadas de mayor calidad. Esto hace que ya no se trate de BSS, sino de separación de fuentes informadas (Informed Source Separation, ISS). La ISS es muy amplia en cuanto a la tipología y morfología de la información que se aporta al sistema. La información puede ser temporal, como en el caso de información de simbólica de score (Score-Informed Source Separation) [9], o espectral, como en el caso de uso de modelos espectrales de las fuentes [10].

2.5.2.3 Separación de fuentes monoaural

Los primeros pasos en la separación de fuentes de sonido se dieron en el campo de la separación de señales de voz [46, 47, 48]. Sin embargo, en los últimos años los grandes avances en el procesado digital de señal de audio han hecho que las técnicas de separación de fuentes musicales avancen notablemente.

El caso de la separación musical, en algunos aspectos, se puede ver como un reto mayor que el de la señal de voz. Los instrumentos musicales tienen una mayor variedad de mecanismos de producción de sonido distintos, mientras que la generación de voz cuenta con el mismo mecanismo en todos los individuos. Además, las características espectrales y temporales del sonido de los instrumentos musicales presentan, también, una gran variabilidad.

Las propuestas relativas a la separación de fuentes de audio en señales monoaurales se pueden clasificar en tres categorías: métodos basados en modelos, métodos de aprendizaje no supervisado y métodos psicoacústicos.

A pesar de esta clasificación, la mayoría de los algoritmos implementan características de varios de estos tipos de métodos.

- **Métodos basados en modelos:**
  Estos métodos usan modelos parámetros que describan el comportamiento espectral y/o temporal de las fuentes presentes en la señal. Los parámetros pueden aprender desde señales específicas de entrenamiento, o bien, de la propia señal mezclada. Algunos algoritmos que usan este tipo de información previa sobre las fuentes son [49].

- **Métodos de aprendizaje no supervisado:**
  Los métodos de aprendizaje no supervisados usan modelos simples no paramétricos, y con menos información previa sobre las fuentes presentes en la señal. Para suplir esta carencia de información, tratan de aprenderla desde la misma señal que se pretende separar. Estos métodos suelen usar información de principios teóricos, como la suposición de independencia entre las fuentes. La reducción de la redundancia de información produce representaciones de las señales aisladas. Algunos algoritmos que siguen estos métodos son ICA [50, 51 y 52], NMF [53, 54, 55 y 56], ANLS [57 y 58] y SC, como en [59].

- **Métodos psicoacústicos:**
  Las habilidades cognitivas humanas permiten pervivir y reconocer las fuentes independientes de un sonido mezclado. Generalmente, los métodos de este tipo cuentan con dos fases. En la primera la señal de entrada se descompone en componentes tiempo/frecuencias básicas, siendo éstas clasificadas entre las diferentes fuentes en la segunda fase, mediante el seguimiento de determinadas

reglas de clasificación. Con ellas, algunos trabajos han implementado algoritmos de separación de fuentes [60].

2.5.3 *Introducción a la separación de fuentes sonoras aplicando NMF*

El primer concepto que debe conocerse es NMF (Non-negative matrix factorization) [61]: es un grupo de algoritmos en el ámbito matemático en el que se realiza una separación de una matriz de entrada **(X)** en dos matrices que son multiplicadas entre sí (ver ecuación 5), y que pretenden mediante otros algoritmos de cálculo de distancias, asemejarse a la matriz de entrada original mediante aproximaciones, siempre sabiendo que las componentes separadas de la matriz de entrada, serán mayores que cero (no negativas).

$$\mathbf{X} = \mathbf{W} \times \mathbf{H} \tag{5}$$

Los algoritmos NMF pueden ser de tres tipos, a continuación, se pasa a describir cada uno de ellos:

- **Sistema supervisado:** En este tipo de sistemas existe una fase de entrenamiento y una fase de separación. En la fase de entrenamiento se deberán entrenar los patrones espectrales de los sonidos que compongan la señal mezcla que se quiera separar. Por tanto, en la fase de separación las matrices de componentes **(W)** vendrán fijadas de la fase de entrenamiento, y sólo será necesario calcular las matrices de activaciones **(H)** de los sonidos a separar.

- **Sistema semisupervisado:** En este tipo de sistemas al igual que en el supervisado existe una fase de entrenamiento y una fase de separación. Pero en la fase de entrenamiento sólo se entrenarán los patrones espectrales de uno de los sonidos que se quieran separar. Por tanto, para la fase de separación la matriz de componentes vendrá fijada de la fase de entrenamiento, y será necesario calcular las matrices de activaciones **(H)** de los sonidos a separar, y la matriz de componentes **(W)** que no ha sido entrenada previamente.

- **Sistema ciego (o no supervisado):**
  En este tipo de sistema no existe fase de entrenamiento. Por lo tanto, en la fase de separación el sistema no conocerá ningunos patrones espectrales y deberá por medio de restricciones, como: sparse, de suavidad temporal o espectral, ser capaz de separar los sonidos mezclados. Los modelos de separación llevados a cabo en este Trabajo Fin de Master se basan en este tipo de sistemas.

2.5.3.1 Explicación del modelo NMF

Non-negative Matrix Factorization**(NMF)** es un conocido método de descomposición no supervisada que permite representar datos en dos dimensiones como una combinación lineal de un conjunto de elementos básicos representativos. Dada la matriz de datos **X** con dimensión $\mathbf{F} \times \mathbf{T}$ y valores no negativos, **NMF** es el problema de encontrar una factorización de la forma:

$$\boldsymbol{X} \approx \boldsymbol{WH} = \boldsymbol{Y} \tag{6}$$

donde **W** y **H** son matrices de valores no negativos de dimensiones $\mathbf{F} \times \mathbf{K}$ **y** $\mathbf{K} \times \mathbf{T}$**,** respectivamente. Habitualmente, **K** toma valores tal que **FK + KT << FT**, reduciendo, por tanto, la dimensión de las matrices a tratar.

Esta herramienta se ha aplicado en multitud de aplicaciones y campos, tales como el aprendizaje de patrones de rostros, patrones semánticos de texto, transcripción automática de música polifónica, separación de fuentes de sonido o clasificación de whiskies escoceses.

En las aplicaciones sobre señales de audio, la matriz **X** se emplea como una representación tiempo-frecuencia (por ejemplo, el espectro de magnitud o potencia), en la que se encuadran los ***bins*** frecuenciales $\boldsymbol{f}$ de cada trama temporal $\boldsymbol{t}$.

Es importante destacar, llegados a este punto, que la señal que debe entrar en el sistema NMF lo hace mediante un espectrograma. Un espectrograma es el resultado de calcular el espectro de tramas enventanadas de una señal, por lo que resulta una gráfica tridimensional que representa la energía del contenido frecuencia de la señal según va variando está a lo largo del tiempo.

El objetivo del algoritmo **NMF** es encontrar las matrices no negativas **W y H** que cumplan:

$$(\boldsymbol{W}, \boldsymbol{H}) = arg\ min_{W,H>0}\, \boldsymbol{D}(\boldsymbol{X}|\boldsymbol{WH}) \tag{7}$$

donde $\boldsymbol{D}(\boldsymbol{X}|\boldsymbol{WH})$ es una función de coste definida como:

$$D(X|WH) = D(X|Y) = \sum_{f=1}^{F} \sum_{t=1}^{T} d(X(f,t)|Y(f,t)) \tag{8}$$

donde X sería la señal musical original e Y la señal reconstruida final y donde $d(x|y)$ es una función de coste escalar.

Algunas de las funciones de coste más conocidas y usadas son: la distancia Euclídea (EUC), Kullback Leibler (KL) e Itakura Saito (IS).

Distancia Euclídea $$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \quad (9)$$

Kullback Leibler $$d_{KL}(x|y) = x\log\frac{x}{y} - x + y \quad (10)$$

Itakura Saito $$d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \quad (11)$$

Las funciones de coste se pueden obtener también a partir de un sistema de ecuaciones (12), donde dependiendo del valor β (coste β-divergencia) se obtendrá una función de coste u otra.

$$d\beta(x|y) = \begin{cases} \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{1}{\beta(\beta-1)}\left(x_{f,t}^{\beta} + (\beta-1)y_{f,t}^{\beta} - \beta x_{f,t}y_{f,t}^{\beta-1}\right) & \beta \in (0,1)\cup(1,2] \\ \sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t}log\frac{x_{f,t}}{y_{f,t}} - x_{f,t} + y_{f,t} & \beta = 1 \\ \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{x_{f,t}}{y_{f,t}} - log\frac{x_{f,t}}{y_{f,t}} - 1 & \beta = 0 \end{cases} \quad (12)$$

Donde en función del parámetro β las funciones de coste definidas serían:

- **La distancia Euclídea (β=2).**
- **Kullback Leibler (β=1).**
- **Itakura Saito (β=0).**

En la figura 2.12 se muestran las diferentes funciones de coste para x = 1.

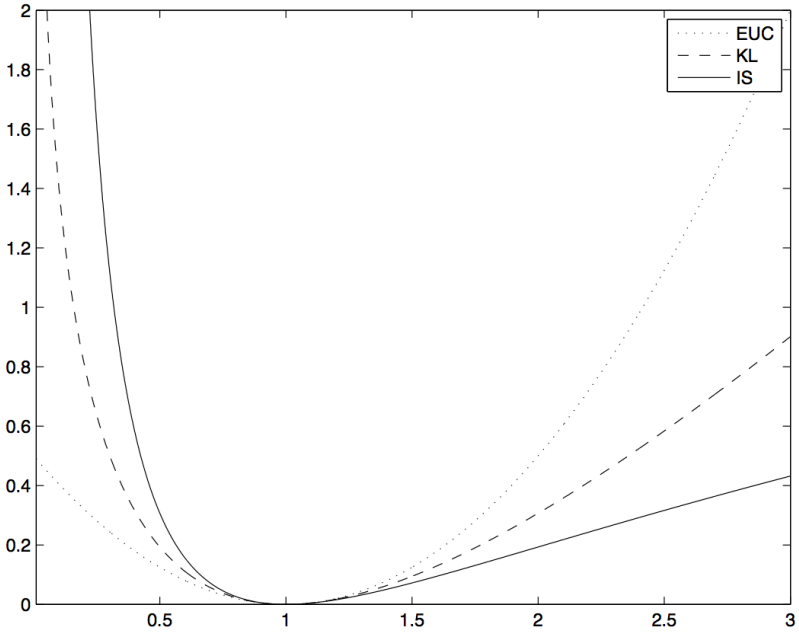

Figura 2.12: Euclídea, KL y IS funciones de coste d(x | y) en función de "y" y para x=1. Las divergencias Euclídea y KL son convexas en (0,∞). La divergencia IS es convexa en (0,2x] y cóncava en [2x,∞).

Las funciones de coste EUC y KL son positivas, toman valor cero si y sólo si $x = y$. Lee y Seung [62] proponen un algoritmo de gradiente descendente para resolver el problema de minimización (ver ecuación 8) usando las funciones de coste descritas. Si en el algoritmo del gradiente se usa un paso adecuado, las ecuaciones de actualización resultantes son multiplicativas, por lo que la función de coste no crecerá. Indudablemente, la simplicidad de las reglas de actualización ha hecho que NMF sea una técnica muy popular, y la mayoría de las aplicaciones, antes mencionadas, usen el algoritmo de Lee y Seung para minimizar la distancia Euclidea o la divergencia de KL.

Esta función de coste fue definida por Itakura y Saito [63] a partir de la estimación de máxima probabilidad (Maximum Likelihood, ML) del espectro de voz con modelado autoregresivo. Se presentó como "una medida de parecido entre dos espectros" y se convirtió en una función de coste muy usada entre los investigadores en el campo de la voz durante los años 70. Una de las características más interesantes de la divergencia IS es la invarianza de escala, es decir, a las componentes de baja energía en X se les asigna la misma relevancia relativa que a las de alta energía. Esta peculiaridad es de gran importancia cuando se tratan espectros de señales de audio, en los que los coeficientes de X se engloban en un gran rango dinámico de valores.

Para cerrar el sistema una vez que se conoce el mecanismo que utiliza la técnica NMF, el cual, básicamente consiste en minimizar las funciones de coste, es necesario introducir el concepto de las ecuaciones de actualización, como aquellas ecuaciones que nos irán permitiendo actualizar los valores de W (matriz de las frecuencias espectrales) o de H (matriz de activaciones temporales), para al fin y al cabo llegar al objetivo final de la separación, que la señal reconstruida sea lo más parecida a la señal de entrada. Para llegar a cumplir el objetivo general es necesario la realización de un determinado número de iteraciones conocido como “maxiter”. Las ecuaciones de actualización de forma general se muestran en las ecuaciones (13) y (14).

$$H <-- H.* \frac{W^T*(WH)^{(\beta-2)}*X}{W^T*(WH)^{(\beta-1)}} \quad (13)$$

$$W <-- W.* \frac{((WH)^{(\beta-2)}*X)*H^T}{(WH)^{(\beta-1)}*H^T} \quad (14)$$

Para poder obtener estas ecuaciones de actualización es necesario partir de una de las funciones de coste (ver ecuación 15). El siguiente paso sería realizar la función del gradiente decreciente (la derivada respecto de una variable, y la derivada respecto de la otra variable) (16) y (17). Por último, obtendríamos las ecuaciones de actualización sabiendo que la parte negativa de la función gradiente iría en el numerador y la parte positiva en el denominador (18) y (19).

Función de coste: $$d\beta = y^{\beta-2}(y - x) \quad (15)$$

Gradiente decreciente:

$$\nabla_H D_\beta (X|WH) = W^T * ((WH)^{(\beta-2)} * (WH - X)) \quad (16)$$

$$\nabla_W D_\beta (X|WH) = \left((WH)^{(\beta-2)} * (WH - X)\right) * H^T \quad (17)$$

Ec. Actualización:

$$H = H \odot \frac{\left[\frac{\partial d_\beta}{\partial H}\right]^-}{\left[\frac{\partial d_\beta}{\partial H}\right]^+} \quad (18)$$

$$W = W \odot \frac{\left[\frac{\partial d_\beta}{\partial W}\right]^-}{\left[\frac{\partial d_\beta}{\partial W}\right]^+} \quad (19)$$

En el modelo NMF básico, la única restricción es la no negatividad de los elementos de todas las matrices intervinientes. El resto de propiedades de la descomposición son efectos adicionales e intrínsecos al algoritmo. Por tanto, el hecho de que el algoritmo sea capaz de obtener información sobre la señal de entrada, desarrolle una separación u obtenga datos interpretables y con significado, se puede considerar como una "buena noticia". Aun así, no es una herramienta mágica, la mayoría de los autores contrarrestan esta falta de control en la descomposición con ciertas restricciones que dan lugar a variantes del NMF básico. Tales restricciones, como, por ejemplo, la dispersión, localización espacial o continuidad temporal, mejoran dichos efectos adicionales en los que los autores se apoyan para lograr datos que se adecúen a su propósito. Algunas de esas restricciones se describen en este capítulo. Una variante típica consiste en agregar un término de penalización, también llamado de regularización, a la función de coste y minimizarla.

2.5.3.2 Principales restricciones utilizadas

Como se ha comentado anteriormente, en la versión estándar de NMF, la única restricción es la no negatividad de los valores de todas matrices del modelo. El resto de las propiedades de la descomposición no son controladas, y todas ellas se producen como "efectos secundarios". A pesar de ser efectos no controlados, la descomposición adquiere ciertas propiedades de la señal original que pueden ser muy útiles para llevar a cabo una separación o la extracción de información relativa a la interpretación. El siguiente paso natural, es tratar de potenciar y controlar estos efectos añadiendo más restricciones, de manera explícita, al problema de factorización.

La manera de introducir restricciones al modelo es usando un término de penalización. De esta forma, además de minimizar el error de reconstrucción Dr, la función de coste incluye un término Dc que cuantifica la característica que se desea restringir. Por tanto, el problema NMF con restricción se puede formular de la siguiente manera:

$$min_{W,H} Dr(X|WH) + \lambda DC(X|WH) \quad (20)$$

donde λ es un factor que se puede ajustar para regular la influencia de la restricción sobre el procedimiento de minimización NMF.

A continuación, se presentan algunas de las restricciones más conocidas:

- **Escasez [45]:** NMF ha demostrado ser una herramienta de análisis útil para tratar diversos tipos de datos. Una de las propiedades más útiles que presenta es que sus descomposiciones son fácilmente interpretables de manera intuitiva por ser dispersas. Sin embargo, en algunas ocasiones, la dispersión de los datos que obtiene NMF, por sí sólo, no es suficiente; en estos casos resulta muy útil poder controlar en grado de dispersión que se produce en el proceso de factorización. Por ejemplo, cuando el espectro de una fuente cubre parcialmente el espectro de otra, la última fuente se puede modelar como la suma del primer sonido más un residuo. El uso de ganancias dispersas puede ayudar a poder identificar la segunda fuente con espectro propio.

- **Monofonía [36]:** Cuando se está trabajando con señales monofónicas, o bien instrumentos monofónicos, se puede usar esta propiedad como restricción a la hora de descomponer la señal. Si se conoce que el instrumento sólo es capaz de generar una nota en cada instante, no tendría sentido permitir que más de un pitch se activara a la vez. Este es un caso extremo de dispersión, que se puede aplicar a toda la señal, o bien, a cada instrumento si la señal se compone de varios instrumentos monofónicos.

- **Localización espacial [62]:** esta restricción es particularmente interesante para muchas aplicaciones de procesado de imagen. La localización de rasgos es de utilidad en el reconocimiento de objetos, incluyendo estabilidad frente a deformaciones, variaciones de luminosidad y oclusiones parciales. Básicamente, la función de coste se modifica para imponer el área de las características que se buscan y adaptando la representación para que la identificación de elementos sea más sencilla, como en el caso del reconocimiento facial.

- **Continuidad temporal [45]:** en el algoritmo NMF estándar, y en la mayoría de sus variantes, las tramas temporales se consideran independientes entre sí, lo cual no es del todo cierto en las señales de audio, por ejemplo. Un rasgo distintivo inherente al sonido de instrumentos musicales es la continuidad temporal. De hecho, como consecuencia de la presencia de una nota, con un pitch constante, las ganancias de tramas adyacentes gnt tienden a ser similares unas a otras. En [45] se introduce un término de penalización en la función de coste NMF para tener en cuenta esta continuidad temporal. En [8], el término está relacionado directamente con la diferencia $\boldsymbol{g_{nt}} - \boldsymbol{g_{n(t-1)}}$.

Los sistemas que se van a implementar en el TFM sobre separación de fuentes sonoras basándose en los modelos NMF van a utilizar las restricciones de escasez, monofonía y continuidad.

2.5.3.3 Conclusión y resumen

En este apartado se realizado un enfoque generalizado sobre la técnica NMF aplicada a la separación de señales musicales para poder explicar de una forma más clara y eficaz el sistema desarrollado en el Trabajo Fin de Master. El cual permitirá realizar una separación de fuentes sonoras de forma automática (empleando el segmentador automático que nos proporcionará los tramos instrumentales de forma inteligente).

Queda por tanto definido el modo de funcionamiento de los algoritmos NMF. Conociendo cual es la técnica **NMF** y las 3 funciones de coste más conocidas, la de distancia **Euclídea (DEUC),** la divergencia de **Kullback- Leibler (DKL)** y la de divergencia de **Itakura-Saito (DIS).**

Dichos algoritmos están basados en el principio de la estimación de máxima similitud mediante variación de parámetros. $\boldsymbol{X}$ (Espectrograma de la señal de entrada), tiene unas dimensiones definidas $\boldsymbol{F} \times \boldsymbol{T}$, por lo que **W y H** tendrán respectivamente dimensiones $\boldsymbol{F} \times \boldsymbol{K}$ y $\boldsymbol{K} \times \boldsymbol{T}$ para que las dimensiones de las matrices sean correctas.

El significado de dichas dimensiones es el siguiente:

- $\boldsymbol{F}$ (número de filas de **W**), será el número total de bins que se ven involucrados en el archivo de audio y estarán contenidos en la matriz **W**.
- $\boldsymbol{T}$ (número de columnas de **H**), el número de frames que se introducirán en la matriz **H**.

- K (número de filas de **H** y de columnas de **W**), es el número de componentes usados para el proceso de reconstrucción. Serán el número de bases con las que se representarán cada uno de los sonidos que queramos entrenar en nuestro sistema. Por lo tanto, existirá un número de bases distintas para cada una de las fuentes sonoras que queramos separar.

Uno de los conceptos más importantes de la técnica NMF es el cálculo de la función de coste. Las funciones de coste serán siempre positivas y la única manera de que sean igual a cero es que X e Y sean iguales, demostrando así que el grado de semejanza es máximo. El resultado de dicha función, es el número total de iteraciones que deberán producirse hasta que la función de coste sea constante, y por tanto se alcance un mínimo local. En la fase de entrenamiento cuando se alcanza el mínimo local, querrá decir que se obtendrán unas bases lo más representativas posible de la señal a entrenar. En la fase de separación cuando la función de coste sea constante, querrá decir que para ese número de iteraciones y con las matrices fijadas de la fase de entrenamiento (W), se tendrá un grado de semejanza suficientemente bueno para que la señal mezcla de entrada y la señal reconstruida por el sistema sean lo más parecidas posible. Consiguiendo así que el programa reconstruya las señales de las distintas fuentes sonoras por separado sin que existan interferencias entre las fuentes y por lo tanto consiguiendo unas medidas objetivas razonablemente buenas.

Una vez que se ha entendido el funcionamiento de las funciones de coste se deberán de emplear las ecuaciones de actualización obtenidas de ellas para poder realizar una iteración (de maxiter iteraciones) para obtener las matrices (H) que no se hayan fijado. Esta explicación está enfocada a un sistema supervisado, ya que si el sistema a implementar es semisupervisado únicamente deberán de entrenar los patrones espectrales de alguno de las fuentes sonoras a separar, por lo que las ecuaciones de actualización se deberán de implementar también para los patrones espectrales que no se hayan fijado en la fase de entrenamiento. Para el caso del sistema ciego, no habrá fase de entrenamiento, por lo tanto, las ecuaciones de actualización se deberán de utilizar para todas las componentes de las fuentes sonoras que intervienen en la señal musical original (patrones espectrales: W y activaciones: H).

Por último, destacar que el sistema llevado a cabo en el TFM se va a comparar con uno de los investigadores que ha abordado el modelo NMF de forma profunda. Este investigador es Durrieu [3] y vamos a realizar una comparación entre los resultados obtenidos en [64] y los obtenidos por el TFM.

3 OBJETIVOS

3.1 Objetivo principal

El objetivo principal de este Trabajo Fin de Master es el diseño y la implementación de un sistema de segmentación y separación de señales musicales a nivel general. Por lo tanto, al finalizar dicho trabajo se debe disponer de un sistema capaz de obtener las distintas fuentes sonoras que componen la señal musical de entrada (armónica, percusiva y vocal) para cualquier señal musical. Para la implementación de dicho sistema se llevarán a cabo distintos algoritmos que permitan, mediante una serie de resultados objetivos, determinar cuál es la mejor opción para conseguir un sistema óptimo. Una vez que se han llevado a cabo las pruebas necesarias con respecto a los distintos algoritmos desarrollados y por lo tanto se ha escogido la mejor opción para desarrollar el sistema automático, se realizara una serie de pruebas para poder analizar las pérdidas del sistema automático con respecto al sistema manual (previamente implementado).

El sistema automático corresponde con aquel sistema que incluye tanto la parte del segmentador automático, como la parte del separador.

El sistema manual corresponde con aquel sistema en el que la parte del segmentador no es necesaria, ya que los ficheros que determinan donde están las zonas instrumentales son creados de forma manual por el investigador. Por lo tanto, solo incluye la parte del separador.

3.2 Objetivos secundarios

Este objetivo principal se divide en los siguientes objetivos secundarios necesarios para poder alcanzar el anterior y que se detallan a continuación.

3.2.1 Segmentador automático

Este es el objetivo inicial de nuestro sistema que nos permitirá realizar el sistema totalmente automático. El objetivo consiste en obtener las zonas de la partitura donde solo hay música de forma automática, analizando únicamente el contenido de la señal y sin que el usuario proporcione ningún tipo de información. En el caso de que el sistema a desarrollar sea el manual esta fase del sistema no sería necesaria, ya que los tiempos en los que solo hay música (para obtener las bases musicales) se proporcionarían de forma manual. Este es el objetivo fundamental del Trabajo fin de Master, ya que le proporciona al sistema el atractivo de construirlo de forma inteligente para que no sea necesario intervenir

en el de forma manual. Por lo tanto, se abordarán diferentes opciones que se estudiaran y se analizaran a lo largo del Trabajo fin de Master para que posteriormente, estudiando sus resultados, se pueda determinar cuál es la mejor opción para el segmentador.

3.2.2 Separador de señales musicales

Este objetivo comprende la segunda fase del sistema, es decir, una vez que ya tenemos los tiempos donde solo se produce música, ya se hayan obtenido de forma manual o de forma automática, se utilizan esos tiempos para extraer las características instrumentales de dichos fragmentos y poder utilizarlas dentro de los tramos donde si hay voz, para así discretizar entre lo que es música y lo que es voz. Además, mediante una serie de restricciones que caracterizan los sonidos armónicos, percusivos y vocales se puede realizar una separación entre ellos, proporcionando a la salida de esta fase tres ficheros correspondientes a las tres fuentes sonoras separadas (vocal, armónico y percusiva). Dentro de este objetivo secundario se abordarán dos fases. La primera permite desarrollar un separador NMF, donde únicamente están presentes las fuentes armónicas y percusivas. Una vez que se ha optimizado el sistema separador NMF para la fuente armónica y percusiva, entra en juego la segunda fase donde se lleva a cabo el sistema separador para las tres fuentes presentes (incluyendo en este caso la fuente vocal). De esta forma se parte de un sistema con solo dos fuentes (armónica y percusiva) y cuando los resultados sean satisfactorios se incluye una nueva fuente (la vocal), de esta forma se afronta el problema en dos bloques.

Es importante destacar que los objetivos secundarios se han ordenado, por lo tanto, dependiendo de la calidad de los resultados de la fase 1, la fase 2 tendrá mejores resultados.

4 MATERIALES Y MÉTODOS

El Trabajo Fin de Master presenta un sistema innovador que consta de dos etapas bien diferenciadas, las cuales son implementadas sobre la plataforma Matlab. La primera etapa consiste en un segmentador automático que permite obtener las zonas musicales de una canción cualquiera y almacena los tiempos instrumentales en un fichero ".txt", esto facilitara mucho la realización de la separación de las fuentes sonoras, ya que para realizar un sistema separador de fuentes basado en la técnica NMF es necesario partir de las zonas instrumentales para poder obtener las bases de las fuentes armónico y percusivas, y de esta forma partir de información que caracteriza las bases instrumentales para obtener una separación lo más efectiva y óptima posible de las fuentes presentes en las zonas vocales. El hecho de poder obtener los tramos musicales de forma automática nos aporta una gran ventaja a la hora de no tener que anotar de forma manual dichos tiempos. Como se ha mencionado en otras partes de la memoria esta tarea resulta ser el gran atractivo del Trabajo Fin de Master, ya que la mayoría de los investigadores se centran en la realización de sistemas que realicen separación de dos o más fuentes sonoras, pero pocos se encargan de detectar esas zonas instrumentales de forma automática para facilitar mucho más el proceso.

La segunda etapa consiste en realizar la separación de una señal musical en sus distintas fuentes sonoras, para este Trabajo Fin de Master, se va a realizar una separación de la señal musical en tres fuentes sonoras, las cuales son: la fuente vocal, la fuente armónica y la fuente percusiva. La realización de esta fase se llevará a cabo utilizando el método Non-Negative Matrix Factorization (NMF). Este método consiste en la factorización de la matriz no negativa del espectro de la señal de entrada llegando a dos matrices, también no negativas, de manera que, una contiene la información correspondiente a las bases de la matriz original y la otra sus activaciones. Para ser más exactos se va a utilizar una modificación de este método que nos permita realizar dicha separación, partiendo siempre de que necesitamos los tramos donde se encuentran las bases instrumentales, ya que en un primer proceso se realizara un entrenamiento de las bases armónico y percusivas, básicamente las llamadas bases instrumentales, para que en una segunda fase se pueda partir de esas bases para realizar una separación global de las fuentes armónico-percusivo y vocal. Sin esas zonas instrumentales se tendría que inicializar las bases armónicas y percusivas previamente con matrices ya entrenadas, para poder realizar una separación sin partir de "cero" y con unos resultados aceptables. El hecho de poder utilizar estos tramos musicales nos permite obtener unos resultados mucho más representativos y exitosos.

Para las dos fases descritas anteriormente se van a llevar a cabo distintos pasos y se abordaran distintas opciones para llegar a la solución más óptima en cada fase, llegando así al sistema óptimo. En comparación con lo desarrollado en el Trabajo Fin de Grado, este sistema general no se especializa en ningún estilo musical, si no que se ha llevado a cabo de forma genérica, por lo tanto, aparece ese valor añadido. De esta forma a lo largo del trabajo Fin de Master se llevará a cabo una optimización del sistema para que funcione de la forma más óptima teniendo en cuenta que la base de datos no será ninguna específica, sino que se parte de una base de datos genérica.

Es importante tener en cuenta que el segmentador que vamos a llevar a cabo parte de la utilización de un preprocesado basado en los algoritmos NMF. Por lo tanto, como el sistema separador se basa también en la misma técnica, se va a llevar a cabo una explicación detallada de cada uno de los sistemas NMF, para que queden bien diferenciados.

En la Figura 4.1 podemos ver un esquema general de la unión entre la fase de segmentación y la fase de separación.

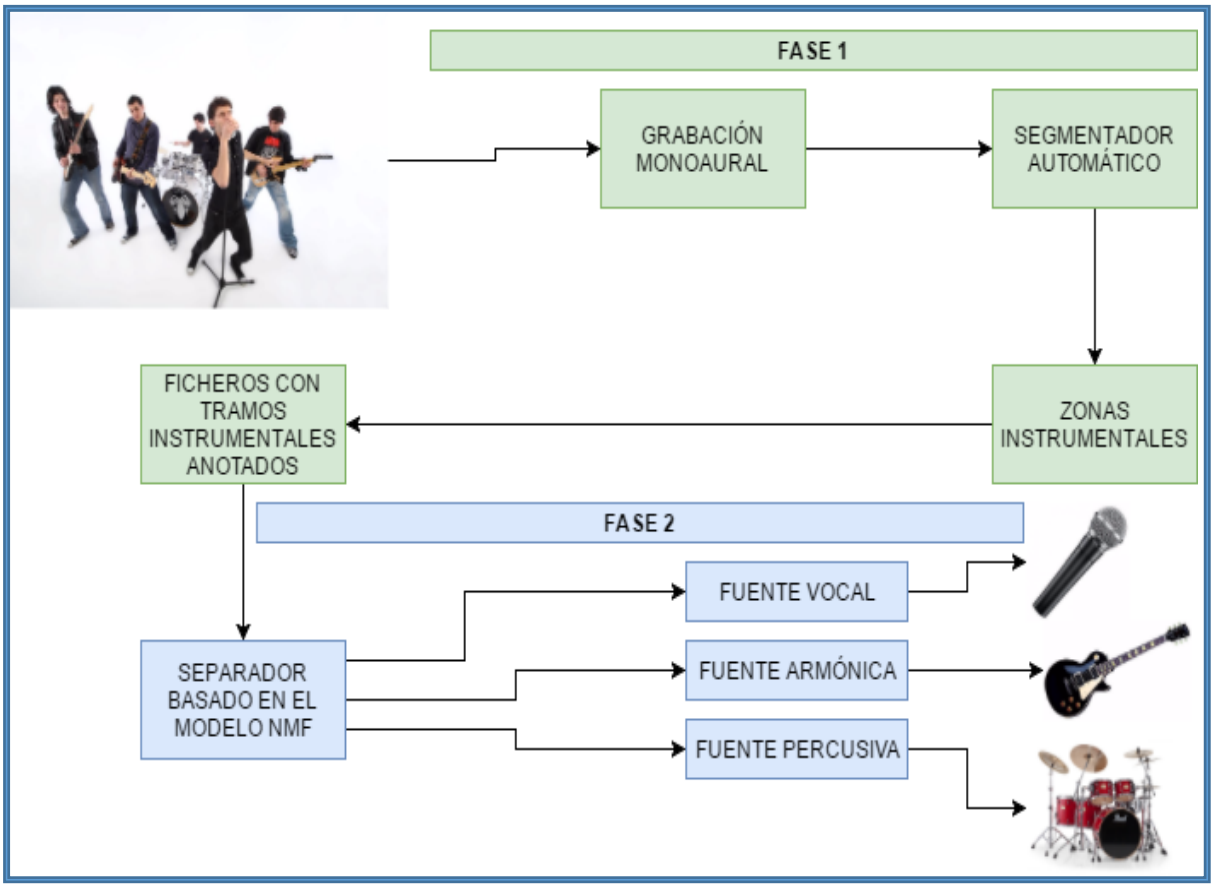

Figura 4.1: Esquema explicativo sobre el sistema a realizar en el TFM con dos fases bien diferenciadas.

A continuación, se desarrolla con detalle el trabajo de implementación llevado a cabo para llegar a la realización de estas dos etapas del sistema.

4.1 IMPLEMENTACIÓN DE LA ETAPA DEL SEGMENTADOR AUTOMÁTICO

La primera etapa del sistema es la clave del Trabajo Fin de Master, como ya se ha mencionado esta etapa es la que nos permite realizar una separación entre la parte instrumental y la parte vocal, es decir, nos permite detectar aquellos tramos temporales donde únicamente existe música y de esta forma poder realizar la fase de separación de las fuentes sonoras con los resultados más óptimos. Esta fase se divide en una serie de bloques para su mejor entendimiento e implementación. En la Figura 4.2 podemos observar los bloques principales en los que se descompone el sistema del segmentador automático a implementar.

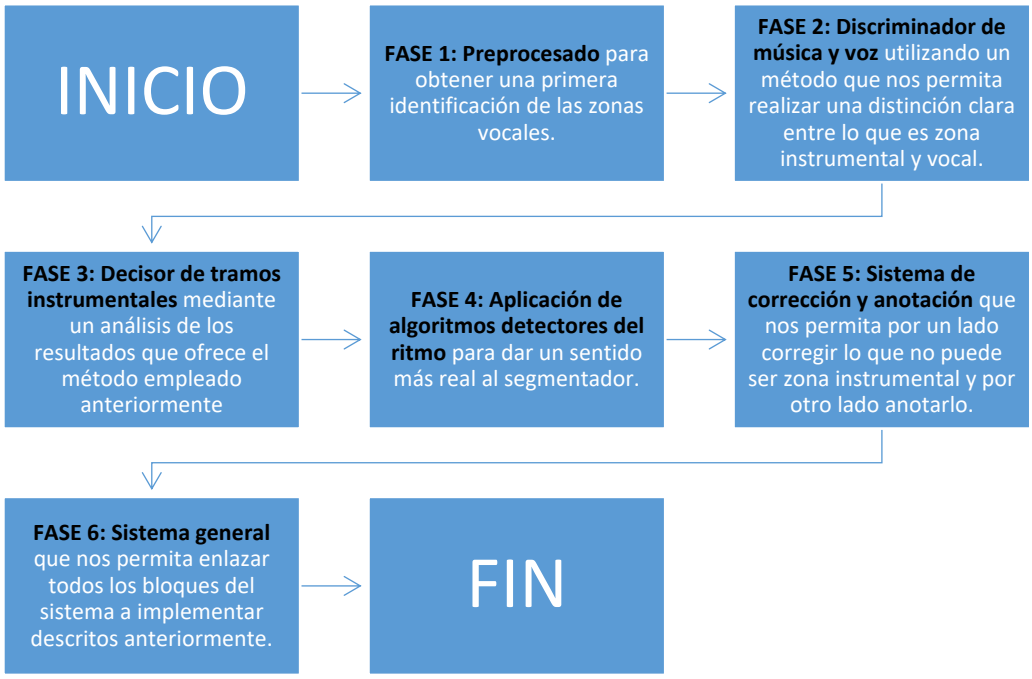

Figura 4.2: Estructura del sistema del segmentador automático dividido en fases.

Como se puede observar en la Figura 4.2 existen 6 fases bien diferenciadas para poder implementar el sistema de segmentación automática, el hecho de que existan un numero de fases amplio se debe a conseguir un mejor entendimiento del sistema de segmentación y poder llevarlo a cabo de la forma más eficaz posible.

En esta sección de la memoria se realizará una explicación, lo más precisa posible, del segmentador automático a implementar. Como se podrá ver a continuación en cada una de las fases se llevará un estudio de las distintas opciones para su implementación y

un análisis que nos permita deliberar cuál es la opción correcta para un sistema óptimo. Los Fases que compone este sistema son:

- **Fase 1:** Preprocesado para obtener una primera identificación de las zonas vocales. En esta fase se va a utilizar la técnica NMF que nos va a permitir identificar de alguna forma y a grandes rasgos dónde se encuentran las zonas vocales. Por lo tanto, es necesario realizar una primera separación de las fuentes sonoras sin la utilización de los tramos temporales donde únicamente existen bases instrumentales, es decir, esta primera separación se realizará mediante una modificación del modelo NMF. En lugar de utilizar las zonas instrumentales para poder obtener las bases armónico y percusivas se utilizará una matriz de bases armónicas entrenadas previamente para poder inicializar de alguna forma las bases y así obtener unos resultados aceptables, pero incomparables con el sistema separador que se pretende realizar con las zonas instrumentales. El objetivo de realizar esta primera separación es poder identificar aquellas zonas donde la voz del cantante está presente y por consiguiente identificar las zonas vocales. Es importante tener claro que lo que se busca en esta fase no es conseguir una separación óptima de las tres fuentes presentes, si no localizar aquellas zonas donde la voz está presente. Al fin y al cabo, lo que nos interesa es llevar a cabo una primera separación de la parte vocal para poder analizar en la siguiente fase y determinar dónde están presentes las zonas instrumentales y las vocales. Para esta fase se van a llevar a cabo dos opciones que nos permitirán analizar de una forma objetiva cual ofrece mejores resultados, las opciones son:

    a) **Preprocesado basado en un sistema NMF sin inicializar las bases armónicas con bases entrenadas.**

    b) **Preprocesado basado en un sistema NMF inicializando las bases armónicas con bases entrenadas correspondientes al piano. Incluyendo además el concepto de $\boldsymbol{\lambda_{Dinámico}}$.**

- **Fase 2:** Discriminador de música y voz utilizando un método que nos permita realizar una distinción clara entre lo que es zona instrumental y lo que es zona vocal. En esta sección, de la memoria del Trabajo Fin de Master, se realizará en primer lugar una implementación de las primeras técnicas utilizadas para conseguir una separación óptima entre las zonas instrumentales y las zonas vocales. Las primeras técnicas o modelos utilizados se basan en la divergencia y en la aplicación de un umbral de energía. Tras la evidencia de que las técnicas mencionadas anteriormente dieron resultados poco prometedores surgió la idea de utilizar un modelo basado directamente

en el modelo NMF utilizado. De esta forma se pensó en utilizar la máscara vocal que forma parte del modelo NMF para determinar un umbral que nos permita quedarnos con la mayor parte de las zonas vocales. De esta forma hablemos desarrollado un segmentador basado totalmente en un sistema NMF. Como se ha dicho anteriormente para esta fase se van a llevar a cabo tres opciones para determinar cuál será la opción optima, las cuales son:

a) **Modelo basado en la divergencia.**
b) **Modelo basado en la intensidad.**
c) **Modelo basado en la máscara vocal del sistema NMF.**

- **Fase 3:** Decisor de tramos instrumentales mediante un análisis de los resultados que ofrece el método empleado anteriormente. Una vez que se ha obtenido una clasificación de las tramas en dos grupos, se debe de seleccionar de alguna forma cual es la zona instrumental y cuál sería la zona vocal. Para esta fase se van a llevar a cabo dos opciones, las cuales son:

  a) **Un análisis mediante ventanas de unos 1.5 segundos de duración, de tal forma que si el 60% de la ventana es mayor que un determinado umbral (fijado por el algoritmo de Otsu) la ventana completa se etiquetara como zona instrumental y en caso contrario se etiquetara como zona vocal.**
  b) **Clasificador HMM (Hidden Markov Model) que permita llevar a cabo un análisis más exhaustivo de los resultados obtenidos en la fase anterior y llevar a cabo un sistema de decisión basado en la probabilidad.**

- **Fase 4:** Aplicación de algoritmos detectores del ritmo para dar un sentido más real al segmentador. Es importante tener en cuenta que en la mayoría de las canciones (independientemente del género musical), cuando el cantante empieza a cantar lo hace de forma acorde al ritmo. Por lo tanto, podemos pensar que si de alguna forma podemos detectar el ritmo de la canción podríamos determinar los posibles momentos en los que el cantante podría comenzar a cantar. De esta forma se podrá llevar a cabo una implementación más exacta del segmentador, dándole un valor añadido con el detector del ritmo. Para esta fase se van a llevar a cabo dos opciones para analizar cuál será la óptima, las cuales son:

  a) **REPET: Repeating Pattern Extraction Technique.**
  b) **ELLIS.**

- **Fase 5:** Sistema de corrección y anotación que nos permita por un lado corregir lo que no puede ser zona instrumental y por otro lado anotar las zonas instrumentales en un fichero de texto. Esta fase del sistema es la que nos permite realizar en primer lugar las correcciones oportunas de los tramos seleccionados como zona instrumental, ya que si, por ejemplo, detecta una zona instrumental de duración inferior a 1 segundo debemos de obviarlo, ya que la posibilidad de que se produzca un segundo de zona instrumental es muy reducida y el sistema tendera a equivocarse. Por otro lado, se proporcionará un margen de seguridad, en el sentido de realizar una disminución de los tramos instrumentales, esto se debe a que nuestro sistema busca detectar las zonas instrumentales en las que únicamente exista la música producida por los instrumentos armónicos y percusivos, pero no nos interesa que aparezcan interferencias correspondientes a la fuente vocal. En segundo lugar, dicha fase nos permitirá señalar de alguna forma estos tramos instrumentales y poder anotarlos en un fichero de texto realizando una pequeña conversión entre tramas (espectrograma) y segundos (tiempo).

- **Fase 6:** Sistema general que nos permita enlazar todos los bloques del sistema a implementar descritos anteriormente. Esta última fase del sistema nos permitirá enlazar cada una de las fases anteriores en un sistema general, que nos permitirá realizar la segmentación de forma totalmente automática.

Para que todo quede lo suficientemente claro en la figura 4.3 podemos observar un esquema detallado de las distintas fases a desarrollar y de las distintas opciones que se llevaran a cabo en dichas fases. Es necesario entender que el hecho de explicar las distintas opciones es para determinar de una forma más clara y sencilla cual es el sistema óptimo. Por lo tanto, con este esquema ofreceremos una visión de las opciones que no se deben de llevar a cabo debido a las razones que se indican en su correspondiente caso. Evitando de esta forma las opciones menos eficientes, podremos llegar a la solución óptima en cada una de las fases, al fin y al cabo, nuestro objetivo. Un segmentador óptimo para cualquier señal musical.

Cada una de las fases descritas anteriormente se explicará de la forma más clara y eficaz posible, con la utilización de una serie de ejemplos para hacerlo de la forma más interactiva posible para el lector.

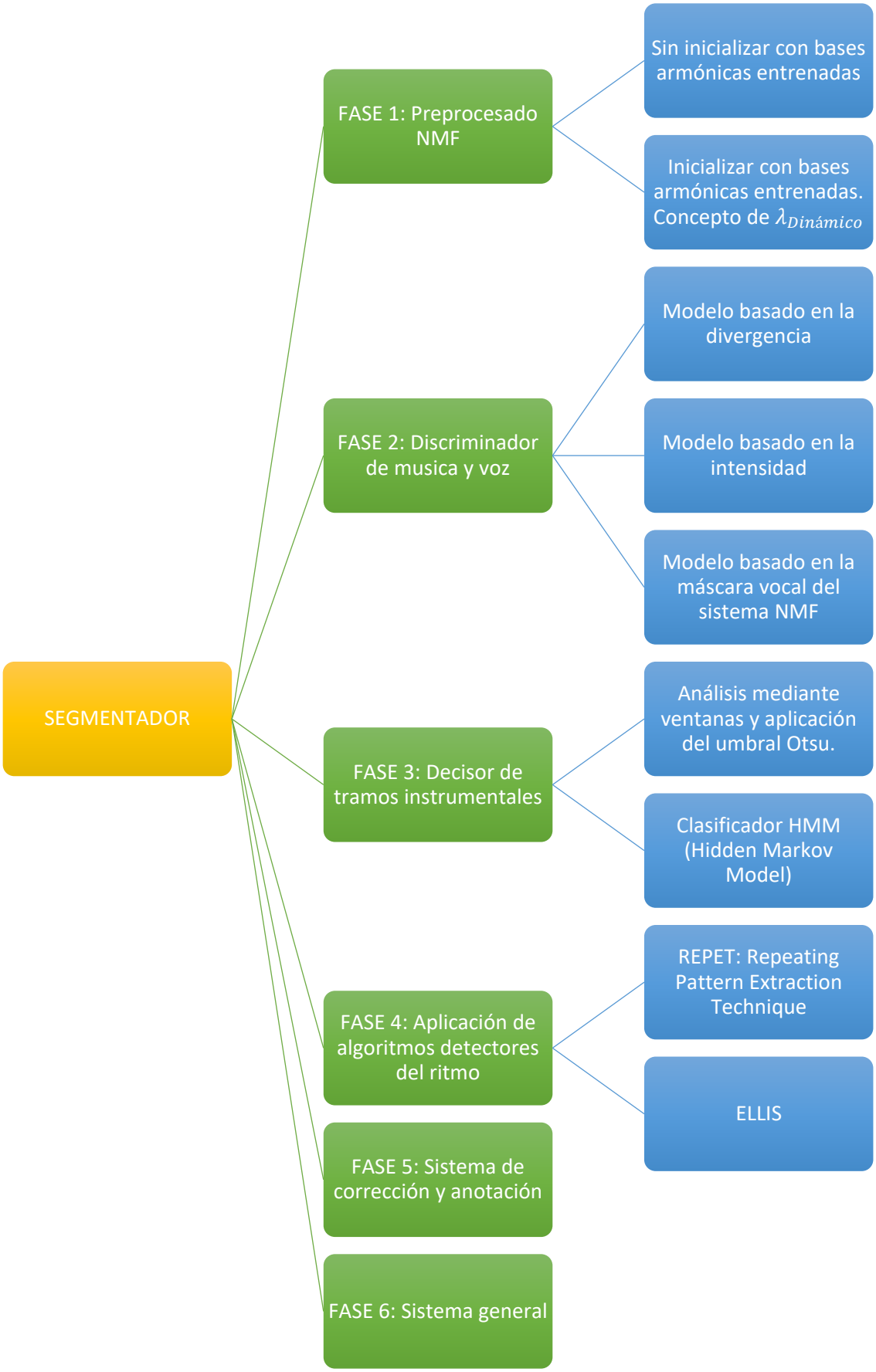

Figura 4.3: Esquema detallado de las distintas etapas y opciones a llevar a cabo para implementar el segmentador.

4.1.1 FASE 1: Preprocesado para obtener una primera identificación de las zonas vocales.

La primera fase que inicia nuestro sistema de segmentación corresponde con un modelo de separación, que al igual que para la fase de separación de las fuentes sonoras (armónico-percusivo-vocal) se va a utilizar un sistema de separación basado en NMF.

El principal motivo de la utilización de esta etapa de separación en el segmentador es conseguir desarrollar un sistema de segmentación/separación basado en una técnica común, conocida como NMF. De esta forma nuestro sistema completo queda totalmente implementado con la misma técnica de una forma más sostenible y sin que interfieran varios tipos de algoritmos.

Es importante considerar que con esta etapa de separación NMF no buscamos separar las distintas fuentes sonoras (armónico/percusivo/vocal) con resultados altamente satisfactorios, sino que por lo consiguiente lo que buscamos es identificar claramente las zonas donde existe la parte vocal. Por lo tanto, el sistema de separación NMF será implementado de tal forma que, aunque el resultado de la fuente vocal no sea totalmente satisfactorio en términos sonoros, tenga bien identificadas las zonas donde si aparece esa fuente vocal mediante su espectrograma.

El motivo por el cual surgió esta idea es básicamente, que si con un sistema de separación basado en NMF somos capaces de distinguir entre las distintas fuentes (armónico/percusivo/vocal), de alguna forma se podría ajustar el algoritmo para que identificase aquellas zonas donde existe la parte vocal.

A diferencia del sistema de separación NMF que se utilizara en la etapa del separador de las distintas fuentes (sección 4.2), este sistema no parte de las bases (armónico y percusivas) previamente entrenadas en las zonas instrumentales. Por lo que es imposible conseguir resultados de separación de las tres fuentes que sean satisfactorios. Pero si podemos conseguir identificar bien las zonas donde está presente la parte vocal.

Al fin y al cabo, el objetivo del segmentador es distinguir entre las zonas vocales y las zonas instrumentales. Una vez que conseguimos obtener las zonas vocales, por consiguiente, conseguimos obtener las zonas instrumentales, que son aquellas donde no existe parte vocal.

Por lo tanto, en esta etapa de separación NMF nuestro objetivo no es separar las distintas fuentes sonoras con unos resultados objetivos y subjetivos óptimos. Nuestro objetivo es ajustar el sistema para que identifique esas zonas donde está presente la parte vocal.

Puede ser confuso el hecho de utilizar varios sistemas de separación basados en la misma técnica y utilizados para fines distintos, pero al fin y al cabo NMF es una técnica solida con una gran cantidad de aplicaciones. Para que todo quede bien explicado y para un mejor entendimiento de los distintos sistemas NMF que se van a utilizar vamos a dedicar un punto de la sección a explicarlo.

El hecho de utilizar el concepto de separación en esta etapa es porque al fin y al cabo lo que queremos es separar la parte vocal del resto y con ello identificar las zonas vocales.

4.1.1.1 Representación de los modelos NMF utilizados para el desarrollo del Trabajo Fin de Master

Para el desarrollo de este Trabajo Fin de Master se han usado tres modelos distintos del algoritmo NMF. Este algoritmo se basa en la factorización de una matriz de entrada, es decir, la matriz de entrada es considerada el producto de un conjunto de matrices. En los modelos desarrollados, se ha tomado como referencia un espectrograma X de entrada, con dimensiones frecuencia-tiempo, para la factorización. Este espectrograma es descompuesto en dos matrices distintas; por un lado, se obtiene una matriz W que se corresponde con la excitación de una fuente sonora en frecuencia (las bases); por otro lado, se obtiene una matriz H, que corresponde con la activación de esa excitación en tiempo (las activaciones).

Como se ha mencionado anteriormente en el desarrollo del Trabajo Fin de Master se han usado tres modelos distintos del algoritmo NMF, como se puede observar en la figura 4.4 los tres algoritmos basados en modelos NMF se utilizan para:

- **El primer modelo NMF** utilizado realiza una separación vocal/armónico/percusiva mediante la utilización de restricciones, ya que se desconocen las tramas temporales correspondientes a las zonas instrumentales. Este algoritmo es el que nos permite realizar una segmentación automática de forma óptima. En conclusión, lo que busca este algoritmo es identificar las zonas vocales sin que los resultados de separación en

esta etapa tengan que ser lo suficientemente satisfactorios, únicamente nos interesa identificar las zonas vocales.

- **El segundo modelo NMF** utilizado realiza una separación armónica/percusiva (sin parte de voz) en las zonas instrumentales detectadas por el segmentador automático. Este algoritmo es el que permite aprender las bases instrumentales para su posterior utilización en el tercer modelo. Por lo tanto, el primer modelo permite obtener las zonas instrumentales (aquellas donde no hay voz) para después utilizarlas en el segundo modelo y conseguir las bases armónico y percusivas que son las que se utilizaran en el tercer modelo NMF.

- **El tercer modelo NMF** utilizado realiza una separación vocal/armónico/percusiva utilizando las bases instrumentales aprendidas en el segundo modelo NMF. Este algoritmo es el que permite realizar una separación optima de las fuentes sonoras que compone la señal musical, partiendo de las bases instrumentales entrenadas y obteniendo unos resultados satisfactorios.

Un punto importante a tener en cuenta es que en la sección 4.2 en la cual se explican los modelos 2 y 3 que son los que nos permite obtener la separación de las fuentes partiendo de las zonas instrumentales obtenidas de la segmentación, también se hace un análisis del modelo de separación NMF utilizado.

Es decir, antes de comenzar a explicar los modelos de separación NMF que nos permitirán realizar la separación optima de las tres fuentes, se desarrolla una implementación de un modelo de separación NMF para separar dos fuentes (armónica y percusiva) con el fin de demostrar que el sistema de separación funciona bien por si solo y así asegurar que combinando la segmentación con la separación conseguimos los resultados esperados.

En la Figura 4.4 podemos observar una explicación detallada de los distintos modelos de separación utilizados para que no haya ningún tipo de confusión entre uno y otro y para que se pueda entender todo de una forma sencilla y amena.

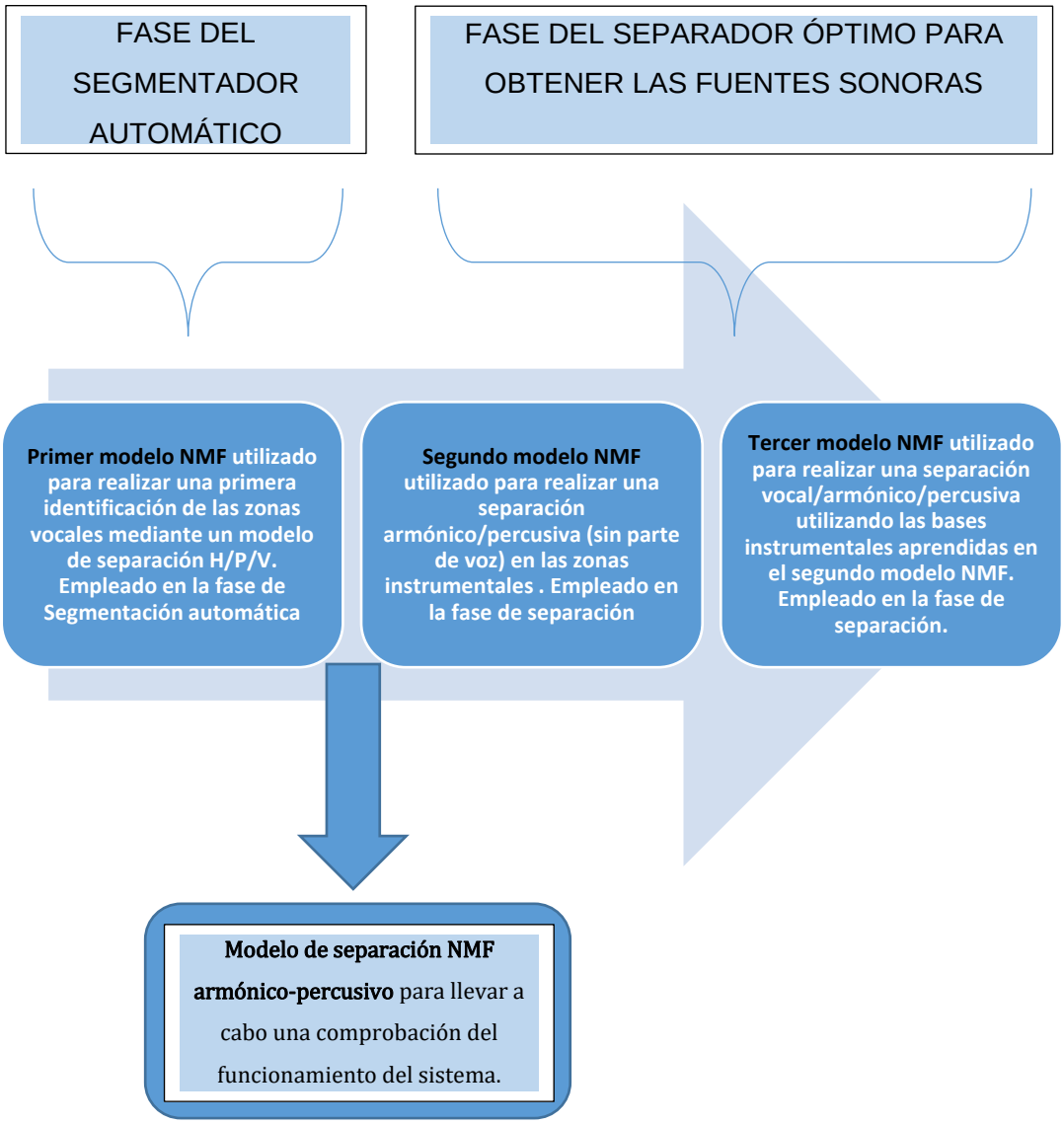

Figura 4.4: Esquema de los tres algoritmos implementados para el desarrollo del TFM basados en modelos NMF

Como se puede ver en el esquema anterior se añade una fase adicional al sistema que nos permite comprobar el funcionamiento del algoritmo de separación NMF independientemente de los resultados del segmentador, ya que únicamente se utilizan dos fuentes (armónico y percusiva) y ambas se consideran instrumentales.

4.1.1.2 *OPCIÓN A: Desarrollo del primer modelo NMF sin inicializar las bases armónicas con bases entrenadas.*

Como se ha mencionado anteriormente lo que se busca con este modelo es hacer una primera detección de las zonas vocales. Por lo que se implementara un sistema de separación NMF para obtener una buena separación de la parte vocal en términos temporales, a pesar de que los resultados auditivos de la separación no sean tan satisfactorios.

Existen diversos métodos que permiten la separación de fuentes sonoras en una pista ya mezclada. Para este TFM, se ha implementado una modificación del algoritmo NMF (Non-Negative Matrix Factorization). Se han asumido una serie de restricciones que ofrecen ciertas ventajas respecto al algoritmo original, en concreto, consideraciones de suavidad (smothness), dispersión (sparseness) y monofonía (single-activation).

Se asume que los sonidos armónicos poseen cualidades de dispersión espectral y suavidad temporal, mientras que los sonidos percusivos poseen cualidades de dispersión temporal y suavidad espectral. Por otro lado, los sonidos correspondientes a la voz presentan un comportamiento correspondiente a la monofonía, ya que un cantante no puede pronunciar más de una nota o fonema en el mismo instante de tiempo.

Entendemos por sonidos percusivos aquellos producidos al golpear instrumentos, mientras que los sonidos armónicos se corresponden con aquellos que surgen como variaciones de frecuencias como es el caso de la guitarra. En cambio, los sonidos procedentes de la voz se caracterizan por la riqueza de la voz del cantante, es decir el comportamiento de la frecuencia no puede dar saltos de octava, sino que sigue un comportamiento melódico.

Estas características permiten modelar sonidos armónicos usando la dispersión en frecuencia, picos espectrales, y la suavidad en tiempo, poca variación en la amplitud. Para los sonidos percusivos encontramos el caso contrario, en frecuencia la energía decrece lentamente mientras en tiempo la mayor parte de la energía se encuentra concentrada en pequeñas regiones. Por otro lado, para los sonidos correspondientes a la parte vocal y a diferencia de los sonidos armónicos, para cada tiempo únicamente debe existir una base presente, en otras palabras, no puede haber acordes.

Ahora se va a llevar a cabo una explicación detallada del algoritmo, teniendo en cuenta que en la sección 4.2 a la hora de explicar los modelos NMF de separación que nos permitirán realizar la separación de las fuentes sonoras, la explicación tendrá una gran similitud. Aun así, es necesario llevar a cabo una explicación de cada modelo para que todo quede lo más claro posible. Aunque para no ser muy monótono en la explicación del NMF, este primer modelo utilizado en la segmentación va a estar explicado de una forma más escueta. Sin embargo, todo el proceso matemático al igual que todo el seguimiento va a estar bien reflejado.

Por lo tanto y sin más demora introducir y recordar que es necesario normalizar la señal de entrada para poder llevar a cabo una separación armónica/percusivo/vocal. Esta normalización es necesario para que las ecuaciones que definen las restricciones que caracterizan a los distintos tipos de sonidos que participan y las ecuaciones de actualización de las matrices de bases y de actualización se puedan implementar de una forma más sencilla.

$$X_{n\beta} = \frac{X}{\sigma . X_{\beta}} = \frac{X}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t}^{\beta}}{FT}\right)^{\frac{1}{\beta}}} \tag{21}$$

Se define una función para factorizar un espectrograma Xn como mezcla de tres espectrogramas separados X_p, X_h y X_v.

$$\begin{aligned} X_{n\beta} \approx \hat{X}_n &= \hat{X}_v + \hat{X}_p + \hat{X}_h \\ &= \left(W_{F0_{F,Re}} diag(eFO_{Re}) H_{F0_{Re,T}} * W_{F_{F,Rf}} diag\left(eF_{Rf}\right) H_{F_{Rf,T}}\right) \\ &+ W_{P_{F,Rp}} diag\left(ep_{Rp}\right) H_{P_{Rp,T}} + W_{H_{F,Rh}} diag(eh_{Rh}) H_{H_{Rh,T}} \end{aligned} \tag{22}$$

Donde:

- $\hat{X}_n$: espectrograma estimado y normalizado de la función mezcla.
- $\hat{X}_p$: espectrograma estimado y normalizado de la fuente percusiva.
- $\hat{X}_h$: espectrograma estimado y normalizado de la fuente armónica.
- $\hat{X}_v$: espectrograma estimado y normalizado de la fuente vocal.
- $W_{P_{F,Rp}}$: matriz de bases percusivas.
- $H_{P_{Rp,T}}$: matriz de activaciones percusivas.
- $W_{H_{F,Rh}}$: matriz de bases armónicas.
- $H_{H_{Rh,T}}$: matriz de activaciones armónicas.

- $W_{F0_{F,Re}}$: matriz de bases de excitaciones vocales.
- $H_{F0_{Re,T}}$: matriz de activaciones de excitaciones vocales.
- $W_{F_{F,Rf}}$: matriz de bases de filtros vocales.
- $H_{F_{Rf,T}}$: matriz de activaciones de filtros vocales.
- $diag(ep_{Rp})$: diagonal de la norma de la parte percusiva.
- $diag(eh_{Rh})$: diagonal de la norma de la parte armónica.
- $diag(eFO_{Re})$: diagonal de la norma de las excitaciones vocales.
- $diag(eF_{Rf})$: diagonal de la norma de los filtros vocales.
- Rp: número de componentes percusivas utilizadas en la factorización.
- Rh: número de componentes armónicas utilizadas en la factorización.
- Re: número de componentes de la matriz de excitaciones vocales.
- Rf: número de componentes de la matriz de filtros vocales.

Es importante destacar que el objetivo principal de los algoritmos NMF es encontrar una estimación de la señal original $\hat{X}_n$ que se asemeje lo suficiente a la señal original $X_{n\beta}$. En más profundidad no nos interesa tanto encontrar una buena estimación de las tres fuentes ($\hat{X}_v + \hat{X}_p + \hat{X}_h$), si no determinar las zonas donde está presente la parte vocal, en otras palabras, encontrar una estimación de la parte vocal $\hat{X}_v$ lo más limpia posible.

Esta separación vocal, se basa en un modelo excitación/filtro, en el cual se considera que para una señal de voz sonora producida por una excitación es dependiente de una frecuencia fundamental. Estas excitaciones o fonemas, son extraídos mediante el uso de un rango discreto de filtros correspondientes a un número limitado de posibles vocales pronunciadas en la voz principal. Bajo ciertos supuestos, podríamos considerar que cada filtro estimado corresponde a un determinado fonema. Se utiliza este modelo excitación/filtro, para determinar la trayectoria de la voz principal de la canción y, así, realizar la separación entre esta y el acompañamiento instrumental.

La matriz de bases de excitaciones vocales $W_{F0_{F,Re}}$ se inicializa con el modelo de excitación global (donde cada columna contiene la excitación, en frecuencia, de un pitch) y se deja fija durante todo el proceso, en cambio, la matriz de bases de fonema $W_{F_{F,Rf}}$ se deja fija inicializándola previamente con bases de fonema entrenadas en una base de datos de voz.

Aparece un nuevo concepto de restricción conocido como monofonía (single activation). El objetivo de esta restricción es que se activen varias componentes de las bases por cada frame en la matriz de activaciones. Por lo tanto, de esta forma aseguramos que la voz del cantante sigue una melodía. Es importante tener en cuenta que, a diferencia de los instrumentos armónicos, como por ejemplo una guitarra, la voz humana no puede pronunciar o cantar varias notas en el mismo tiempo, por lo tanto, es lógico que con esta restricción ayudemos al algoritmo a obtener mejores resultados de separación.

El objetivo del sistema es minimizar una función de coste global para estimar $W_{P_{F,Rp}}$, $H_{P_{Rp,T}}$, $W_{H_{F,Rh}}$, $H_{H_{Rh,T}}$, $W_{F0_{F,Re}}$, $H_{F0_{Re,T}}$, $W_{F_{F,Rf}}$ y $H_{F_{Rf,T}}$. Esta función de coste global depende del coste β-divergencia y de los costes procedentes de las restricciones que se asocian con la parte armónica, percusiva y vocal, es lo que se conoce como coste armónico, coste percusivo y coste vocal.

Por lo tanto, el siguiente paso es definir los distintos costes que nos van a permitir estimar las matrices de las activaciones y de las bases para las distintas fuentes sonoras presentes.

- **Coste β-divergencia:** El espectrograma normalizado de entrada $X_{n\beta}$ se construye minimizando el coste β-divergencia $d\beta(x|y)$ para los tres espectrogramas $\hat{X}_p$, $\hat{X}_h$ y $\hat{X}_v$. Considerando que, $x = X_{n\beta}$ e $y = \hat{X}_p + \hat{X}_h + \hat{X}_v$.

  En la ecuación 23 se puede observar las distintas funciones que existen en función del valor de β. Para la realización de este Trabajo Fin de Master, se ha seleccionado un valor de β=1,3. Este valor ha sido seleccionado tras la realización de diferentes pruebas, ya que se ha demostrado que actúa de la forma más eficiente.

$$d\beta(x|y) = \begin{cases} \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{1}{\beta(\beta-1)}\left(x_{f,t}^{\beta} + (\beta-1)y_{f,t}^{\beta} - \beta x_{f,t}\, y_{f,t}^{\beta-1}\right) & \beta \in (0,1)\cup(1,2] \\ \sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t} log\frac{x_{f,t}}{y_{f|t}} - x_{f,t} + y_{f,t} & \beta = 1 \\ \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{x_{f,t}}{y_{f,t}} - log\frac{x_{f,t}}{y_{f,t}} - 1 & \beta = 0 \end{cases} \quad (23)$$

- **Coste percusivo:** se toman las consideraciones de suavidad en frecuencia y dispersión en tiempo. Los sonidos percusivos suelen representarse como señales de banda ancha, con energía confinada en cortos intervalos de tiempo. Se usa una definición de suavidad espectral, SSM, asociada con la matriz Wp. Un alto coste, se asocia con ganancias distintas de cero, suponiendo que los sonidos percusivos pueden representarse como transitorios con energías concentradas en cortos intervalos de tiempo. Este hecho se asocia con Hp.

$$SSM = \frac{T}{Rp}\sum_{rp=1}^{Rp}\frac{1}{\sigma W_{P_{F,Rp}}{}^2}\sum_{f=2}^{F}\left(W_{P_{f-1,rp}} - W_{P_{f,rp}}\right)^2 \tag{24}$$

$$TSP = \frac{F}{R_p}\sum_{r_p=1}^{R_p}\sum_{t=1}^{T}\left|\frac{H_{P_{rp,t}}}{\sigma H_{P_{rp}}}\right| \tag{25}$$

$$\text{Con } \sigma H_{P_{rp}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} H_{P_{rp,t}}^2} \text{ y } \sigma W_{P_{rp}} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} W_{P_{rp,t}}^2}$$

- **Coste armónico:** para los sonidos armónicos ocurre de forma similar que en el caso anterior. Sin embargo, en este caso se toman consideraciones de dispersión en frecuencia y suavidad en tiempo.

$$SP = \frac{T}{R_h}\sum_{r_h=1}^{R_h}\sum_{f=1}^{F}\left|\frac{W_{H_{f,rh}}}{\sigma W_{H_{rh}}}\right| \tag{26}$$

$$TSM = \frac{F}{R_h}\sum_{r_h=1}^{R_h}\frac{1}{\sigma H_{H_{rh}}{}^2}\sum_{t=2}^{T}(H_{H_{rh,t-1}} - H_{H_{rh,t}})^2 \tag{27}$$

- **Coste vocal:** para los sonidos vocales es importante tener en cuenta que únicamente la voz puede pronunciar o cantar una única nota por cada instante, por lo tanto, se aplicara la restricción de monofonía a las matrices de activaciones de los filtros y las excitaciones vocales.

$$SAHF0 = \frac{1}{Re(Re-1)}\sum_{i=1}^{Re}\sum_{j=1}^{T} HF0 * HF0^T - Trace(HF0 * HF0^T) \tag{28}$$

$$SAHF = \frac{1}{Rf(Rf-1)} \sum_{i=1}^{Rf} \sum_{j=1}^{T} HF * HF^T - Trace(HF * HF^T) \quad (29)$$

Donde Trace se refiere al cálculo de la diagonal de las matrices entre paréntesis.

Por lo tanto y como se muestra en la ecuación siguiente la función de coste global queda determinada de la siguiente forma.

$$D = d\beta\left(X_{n\beta}|\left(\widehat{X_n}\right)\right) + K_{SSM}SSM + K_{TSP}TSP + K_{SSP}SSP + K_{TSM}TSM + K_{SAHFO}SAHFO + K_{SAHF}SAHF \quad (30)$$

Donde:

- K_{SSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las bases percusivas.
- K_{TSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las activaciones percusivas.
- K_{SSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las bases armónicas.
- K_{TSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las activaciones armónicas.
- K_{SAHFO}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de las excitaciones vocales.
- K_{SAHF}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de los filtros vocales.

Esta es la clave que nos permitirá dar más peso a unas restricciones que a otras y lo que nos llevará a conseguir una separación lo más eficiente posible. Estos pesos se ajustan llevando a cabo distintas pruebas y análisis, de tal forma que observando los resultados obtenidos con unos pesos u otros podamos determinar cuáles son los pesos que se considerarían óptimos.

El siguiente paso es usar las llamadas “reglas de actualización multiplicativas”, de forma que $\boldsymbol{D}$ no aumente, mientras se garantiza la no negatividad mediante un algoritmo de gradiente descendente.

- **Reglas de actualización multiplicativas para la base $H_{F0_{Re,T}}$.**

$$H_{F0} = H_{F0} \odot \frac{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{-} + K_{SAHFO}\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{-}}{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{+} + K_{SAHFO}\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{+}} \tag{31}$$

$$\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{-} = (diag(eF0)W_{F0})'\left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{32}$$

$$\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{+} = (diag(ep)W_{F0})'\left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{33}$$

$$\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{-} = \frac{2}{Re(Re-1)}H_{F0} \tag{34}$$

$$\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{+} = \frac{2}{Re(Re-1)}1_{Re,Re}H_{F0} \tag{35}$$

- **Reglas de actualización multiplicativas para la base $H_{F_{Rf,T}}$.**

$$H_F = H_F \odot \frac{\left[\frac{\partial d\beta}{\partial H_F}\right]^{-} + K_{SAHF}\left[\frac{\partial SAHF}{\partial H_F}\right]^{-}}{\left[\frac{\partial d\beta}{\partial H_F}\right]^{+} + K_{SAHF}\left[\frac{\partial SAHF}{\partial H_F}\right]^{+}} \tag{36}$$

$$\left[\frac{\partial d\beta}{\partial H_F}\right]^{-} = (diag(eF)W_F)'\left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{37}$$

$$\left[\frac{\partial d\beta}{\partial H_F}\right]^{+} = (diag(eF)W_F)'\left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{38}$$

$$\left[\frac{\partial SAHF}{\partial H_F}\right]^{-} = \frac{2}{Rf(Rf-1)}H_F \tag{39}$$

$$\left[\frac{\partial SAHF}{\partial H_F}\right]^{+} = \frac{2}{Rf(Rf-1)}1_{Rf,Rf}H_F \tag{40}$$

- **Reglas de actualización multiplicativas para la base $W_{P_{F,Rp}}$.**

$$W_p = W_p \odot \frac{\left[\frac{\partial d\beta}{\partial W_p}\right]^- + K_{SMM}\left[\frac{\partial SSM}{\partial W_p}\right]^-}{\left[\frac{\partial d\beta}{\partial W_p}\right]^+ + K_{SMM}\left[\frac{\partial SSM}{\partial W_p}\right]^+} \tag{41}$$

$$\left[\frac{\partial d\beta}{\partial W_p}\right]^- = \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] diag(ep)' \tag{42}$$

$$\left[\frac{\partial d\beta}{\partial W_p}\right]^+ = \left[\left(\hat{X}_n\right)^{\beta-1}\right] \left(diag(ep) H_p\right)' \tag{43}$$

$$\left[\frac{\partial SSM}{\partial W_p}\right]^+_{f,rp} = \frac{1}{2(F-1)\left(\frac{rp}{F}\right)} 4W_p \tag{44}$$

$$\left[\frac{\partial SSM}{\partial W_p}\right]^-_{f,rp} = \frac{2}{2(F-1)\left(\frac{rp}{F}\right)} \left[\left(W_{P_{f-1,rp}} + W_{P_{f+1,rp}}\right) + \sum_{f=2}^{F}\left(W_{P_{f,rp}} - W_{P_{f-1,rp}}\right)^2\right] \tag{45}$$

- **Reglas de actualización multiplicativas para la base $H_{P_{Rp,T}}$.**

$$H_p = H_p \odot \frac{\left[\frac{\partial d\beta}{\partial H_p}\right]^- + K_{TSM}\left[\frac{\partial TSM}{\partial H_p}\right]^-}{\left[\frac{\partial d\beta}{\partial H_p}\right]^+ + K_{TSM}\left[\frac{\partial TSM}{\partial H_p}\right]^+} \tag{46}$$

$$\left[\frac{\partial d\beta}{\partial H_p}\right]^- = \left(diag(ep) W_p\right)' \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{47}$$

$$\left[\frac{\partial d\beta}{\partial H_p}\right]^+ = \left(diag(ep) W_p\right)' \left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{48}$$

$$\left[\frac{\partial TSP}{\partial H_p}\right]^+_{rp,t} = \frac{1}{\sqrt{T} * rp} \tag{49}$$

$$\left[\frac{\partial TSP}{\partial H_p}\right]^-_{rp,t} = 0 \tag{50}$$

- **Reglas de actualización multiplicativas para la base $W_{H_{F,Rh}}$.**

$$W_H = W_H \odot \frac{\left[\frac{\partial d\beta}{\partial W_H}\right]^- + K_{SSP}\left[\frac{\partial SSP}{\partial W_H}\right]^-}{\left[\frac{\partial d\beta}{\partial W_H}\right]^+ + K_{SSP}\left[\frac{\partial SSP}{\partial W_H}\right]^+} \tag{51}$$

$$\left[\frac{\partial d\beta}{\partial W_h}\right]^- = \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] diag(eh)' \tag{52}$$

$$\left[\frac{\partial d\beta}{\partial W_h}\right]^+ = \left[\left(\hat{X}_n\right)^{\beta-1}\right] (diag(eh)H_h)' \tag{53}$$

$$\left[\frac{\partial SSP}{\partial W_h}\right]^+_{f,rh} = \frac{1}{\sqrt{F} * rh} \tag{54}$$

$$\left[\frac{\partial SSP}{\partial W_h}\right]^-_{f,rh} = 0 \tag{55}$$

- **Reglas de actualización multiplicativas para la base $H_{H_{Rh,T}}$.**

$$H_H = H_H \odot \frac{\left[\frac{\partial d\beta}{\partial H_H}\right]^- + K_{TSM}\left[\frac{\partial TSM}{\partial H_H}\right]^-}{\left[\frac{\partial d\beta}{\partial H_H}\right]^+ + K_{TSM}\left[\frac{\partial TSM}{\partial H_H}\right]^+} \tag{56}$$

$$\left[\frac{\partial d\beta}{\partial H_h}\right]^- = (diag(ep)W_h)'\left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{57}$$

$$\left[\frac{\partial d\beta}{\partial H_h}\right]^+ = (diag(ep)W_h)'\left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{58}$$

$$\left[\frac{\partial TSM}{\partial H_H}\right]^+_{rh,t} = \frac{1}{2(T-1)\left(\frac{rh}{T}\right)} 4H_h \tag{59}$$

$$\left[\frac{\partial TSM}{\partial H_h}\right]^-_{rh,t} = \frac{2}{2(T-1)\left(\frac{rh}{T}\right)}\left[\left(H_{H_{rh,t-1}} + H_{H_{rh,t+1}}\right) + \sum_{t=2}^{T}\left(H_{H_{rh,t}} - H_{H_{rh,t-1}}\right)^2\right] \tag{60}$$

Una vez que se ha llevado a cabo una explicación detallada de las distintas ecuaciones de actualización con sus respectivas restricciones es necesario dejar claro el proceso interno que se lleva a cabo dentro del bucle iterativo en el que se llevan a cabo la actualización y reconstrucción de las matrices que entran en juego.

Por lo tanto, es preciso comentar que todas las matrices correspondientes a las bases (W) y a las activaciones (H) están normalizadas. La norma de estas matrices se guarda en un vector definido como “e” y su multiplicación con la matriz se realizada diagonalizando el vector.

Dentro de cada iteración del modelo NMF, se actualizan las matrices H y W mediante las correspondientes ecuaciones de actualización, y después se vuelven a normalizar pasando los valores de norma extraídos al vector “e”.

Todo esto se hace para hacer más fáciles las ecuaciones de actualización ya que las regularizaciones se hacen con respecto a matrices normalizadas. Para terminar, las regularizaciones o restricciones se hacen con respecto a matrices normalizadas para asegurar que su proporción en la regularización con respecto a la beta-divergencia sea siempre constante o se pueda controlar de antemano.

- **Actualización de la norma para las matrices de bases.**

$$ep = ep \odot \sqrt{\sum {W_p}^2} \tag{61}$$

$$eh = eh \odot \sqrt{\sum {W_e}^2} \tag{62}$$

$$eF0 = eF0 \odot \sqrt{\sum {W_{F0}}^2} \tag{63}$$

- **Actualización de la norma para las matrices de activaciones.**

$$ep = ep \odot \sqrt{\sum {H_p}^2} \tag{64}$$

$$eh = eh \odot \sqrt{\sum {H_h}^2} \tag{65}$$

$$eF0 = eF0 \odot \sqrt{\sum {H_{F0}}^2} \tag{66}$$

$$eF = eF \odot \sqrt{\sum {H_F}^2} \tag{67}$$

Una vez que se ha explicado el proceso por el cual se estiman todas las matrices que entran en juego en el sistema es necesario hacer una serie de aclaraciones que se consideran importantes en el proceso, para un mejor entendimiento.

- La matriz de bases de excitaciones vocales $W_{F0_{F,Re}}$ se inicializa con el modelo de excitación global (donde cada columna contiene la excitación, en frecuencia, de un pitch) y se deja fija durante todo el proceso, en cambio, la matriz de bases de fonema $W_{F_{F,Rf}}$ se deja fija inicializándolas previamente con bases de fonema entrenadas en una base de datos de voz.

- Las bases $H_{F0_{Re,T}}$ y $H_{F_{Rf,T}}$ se irán actualizando mediante las reglas de actualización multiplicativas, al igual que las bases percusivas y armónicas. Todas estas matrices se inicializan de forma aleatoria y se dejan evolucionar como se indica anteriormente.

- Las normas de cada una de las matrices se inicializan a unos y se dejan evolucionar como se muestra en las ecuaciones anteriores.

- Es importante tener en cuenta que todas estas matrices que entran en juego se deben de normalizar, incluyendo tanto la matriz de entrada original (señal musical a analizar), como la matriz de salida estimada.

- Una vez que se han inicializado todas estas matrices entramos en un bucle en el cual se llevan a cabo las distintas ecuaciones de actualización. El tamaño del bucle es de 100 iteraciones, ya que tras sucesivas pruebas se ha determinado que es el punto donde se alcanza el mínimo local, un concepto que se puede apreciar en el estado de arte a la hora de explicar el proceso de separación NMF original.

- Por otro lado, es importante remarcar que nos encontramos ante un modelo de separación NMF que lo que busca es obtener una separación vocal lo más limpia posible, de tal forma que nos quedemos con las zonas donde está presente la fuente vocal. Independientemente de la calidad auditiva de la fuente vocal extraída lo verdaderamente importante es que los tiempos de esa fuente vocal estén bien identificados.

El siguiente esquema nos muestra el proceso que sigue esta etapa del segmentador hasta llegar a su objetivo final. El cual es conseguir un espectrograma de la voz donde se aprecie de forma significativa las zonas en la que está presente.

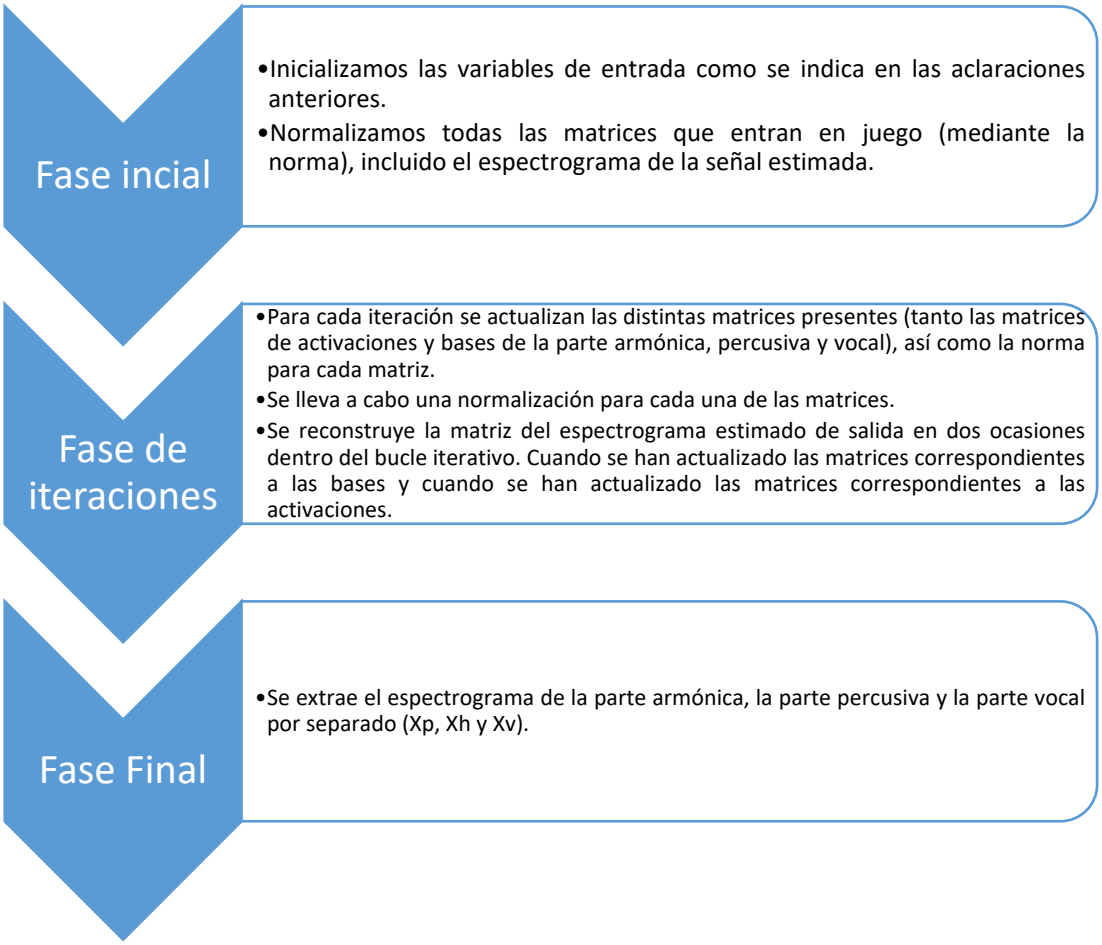

Figura 4.5: Explicación de los pasos que se llevan a cabo en la implementación de la primera fase del segmentador.

Aunque la Figura 4.5 muestra el esquema del proceso de separación llevado a cabo en la primera fase del segmentador, es necesario incluir un esquema que identifique cada uno de los pasos llevados a cabo en la “Fase de Iteraciones” definida anteriormente.

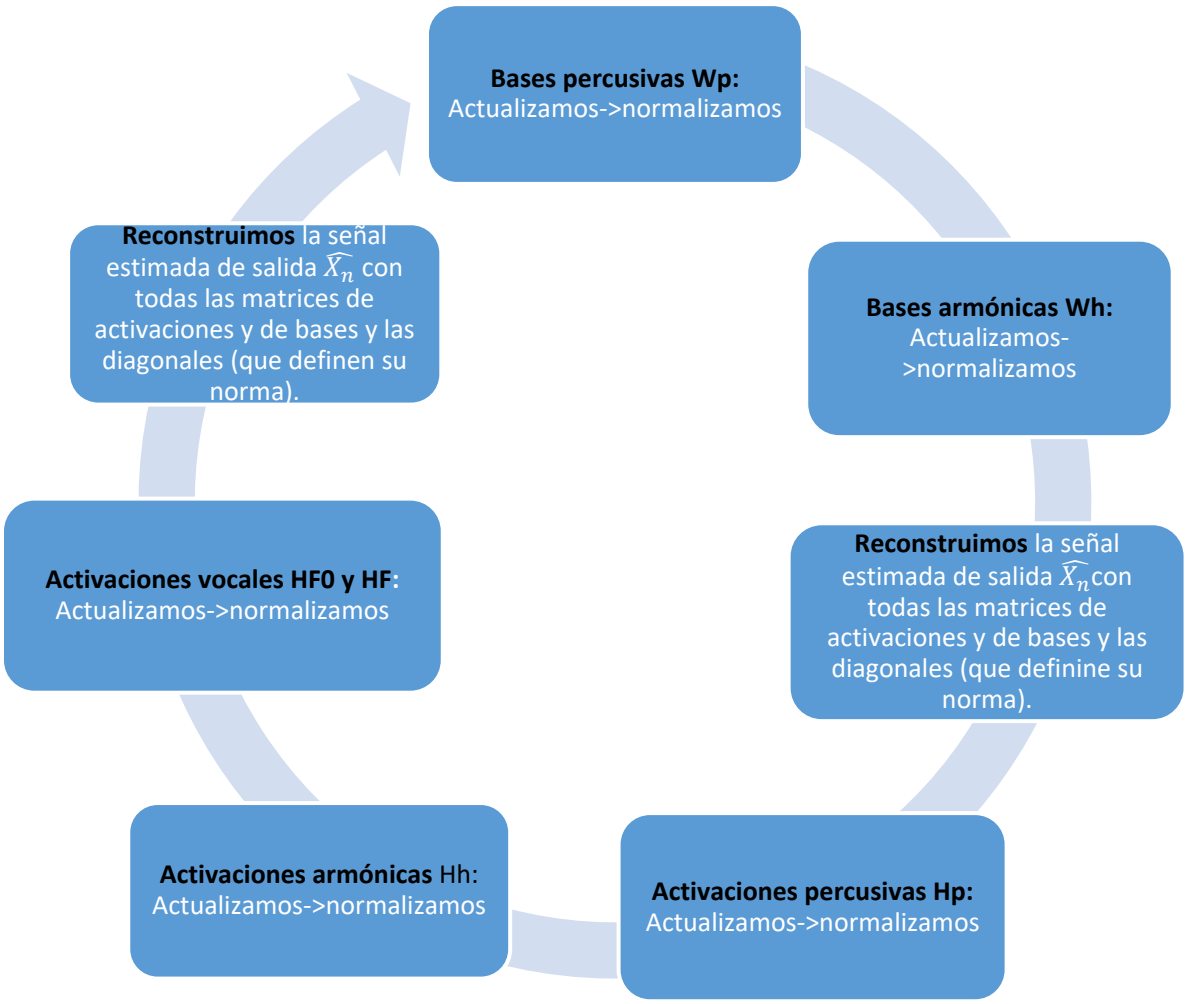

Figura 4.6: Esquema aclarativo de los diferentes pasos llevadas a cabo en la "Fase de Iteraciones" del modelo NMF utilizado en el segmentador.

Es importante dejar claro que siempre que se actualizan las matrices correspondientes a las bases o a las activaciones también se actualizan las normas correspondientes a cada matriz.

Por último, quedaría obtener el espectrograma de cada una de las fuentes sonoras, es decir, la parte correspondiente a cada fuente y que forma parte de la señal estimada de salida en su totalidad.

$$X_p = W_p diag(ep) H_p \tag{68}$$

$$X_h = W_h diag(eh) H_h \tag{69}$$

$$X_v = W_{F0} diag(eFO) H_{F0} * W_F diag(eF) H_F \tag{70}$$

Una vez que se ha entendido el sistema se realiza un análisis de los resultados obtenidos teniendo en consideración el espectrograma de la señal vocal que se ha estimado. Aunque estos resultados no sean objetivos, pueden ser de gran ayuda para realizar un análisis con respecto a las zonas vocales que está determinando nuestro sistema.

En la siguiente imagen podemos observar el espectrograma de la parte vocal que se ha estimado. A lo largo de la memoria se utilizará una señal musical correspondiente al género de música 'pop' para llevar a cabo un análisis visual de las distintas etapas. Esta canción corresponde al fichero de audio denominado como *'Hollywood_Nights'*.

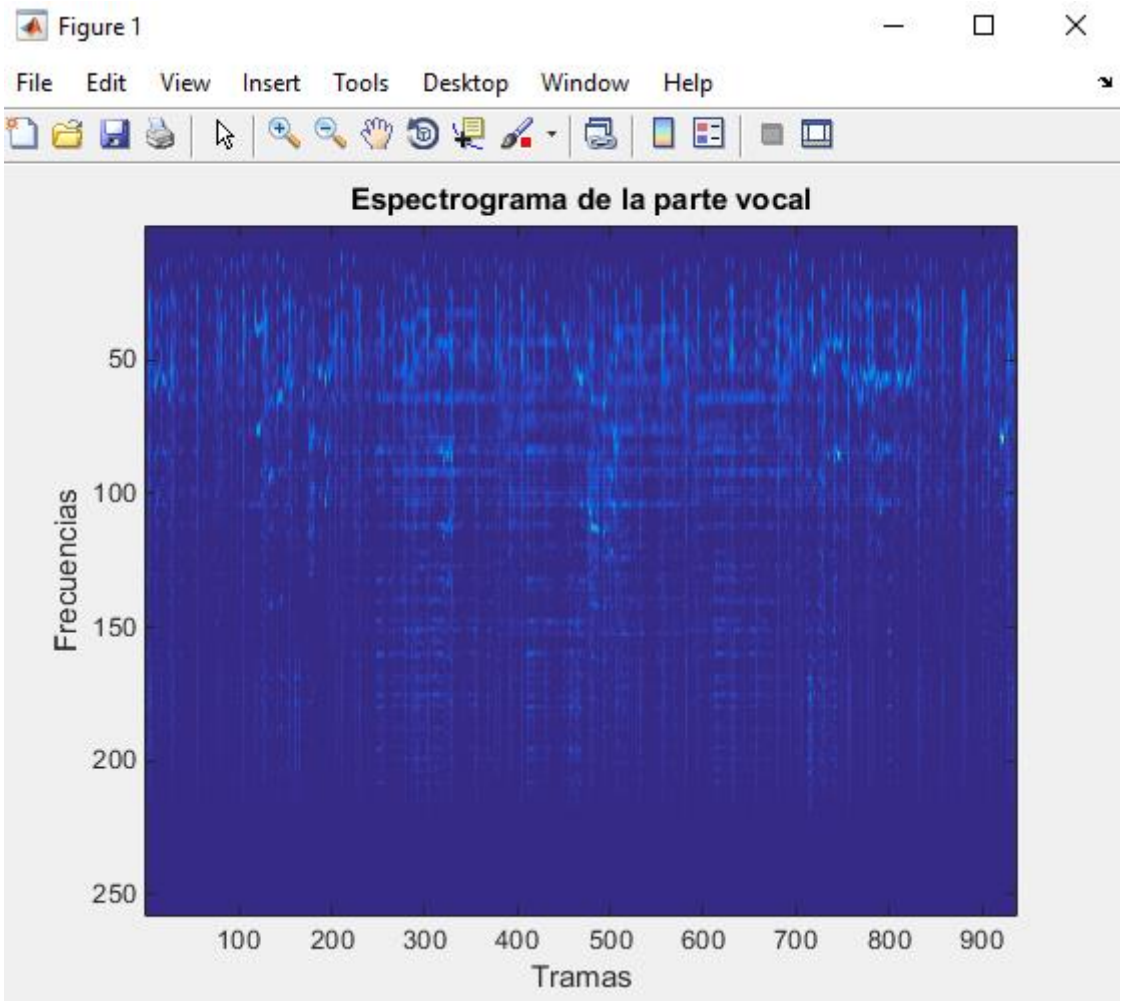

Figura 4.7: Espectrograma de la parte vocal con la opción A

Como se puede observar en la Figura 4.7 se puede ver que el espectrograma de la parte vocal no presenta una separación lo suficientemente buena como para determinar las zonas vocales. Esto se debe principalmente a que por un lado el sistema no dispone de la suficiente información adicional (por ejemplo, bases armónicas entrenadas) como para conseguir una separación lo suficientemente buena.

Por otro lado, es importante tener en cuenta que estamos trabajando con un gran número de matrices, tanto de bases como de activaciones, y eso supone que a su vez estemos trabajando con un gran número de restricciones. Por lo tanto, es lógico pensar que dentro del bucle donde se realizan las iteraciones que llevan a cabo la actualización y la reconstrucción de las distintas matrices correspondientes a las bases y a las activaciones, pueda ocurrir lo siguiente:

- Por un lado, el coste β-divergencia va a ser distinto para las matrices de las bases que para las de las activaciones. Esto se debe a que en primer lugar se lleva a cabo la actualización de las matrices correspondientes a las bases, después se reconstruye la señal y seguidamente se lleva a cabo la actualización de las matrices correspondientes a las activaciones como se puede ver en la figura 18. Por lo tanto y como el coste β-divergencia depende de la señal reconstruida va a ser distinto en las matrices de bases que en las de activaciones.

- Por otro lado, dependiendo del peso que le demos a las restricciones podremos obtener unos resultados u otros. Ya que al fin y al cabo con ese peso lo que hacemos es añadir importancia a la restricción. El problema es que buscar un acuerdo entre los pesos de las restricciones para conseguir una separación idónea es complicado cuando entran en juego muchas restricciones.

- Por otro lado, esto nos lleva a que las ecuaciones de actualización dependen de la función de coste de sus respectivas restricciones por lo tanto se puede producir el hecho de trabajar con valores muy dispares y de esta forma llevar a alguna de las matrices a valores cercanos al cero y de esta forma anular por completo una de las fuentes sonoras presentes en la señal musical. Al fin y al cabo, es imposible aprender tantas bases y ganancias con tantas restricciones y el modelo se ajusta por su cuenta poniendo a cero ciertas matrices.

Es por ello que surgiera la necesidad de hacer una mejora en cuanto al modelo NMF que nos surgió como idea inicial, el cual será explicado en el siguiente punto,

considerado como la opción B y en el cual podremos observar cómo hacer frente a ese tipo de problemas.

4.1.1.3 OPCIÓN B: Desarrollo del primer modelo NMF inicializando las bases armónicas con bases entrenadas correspondientes al piano. Añadiendo el concepto de $\lambda_{Dinámico}$.

Una vez que se han analizado los problemas que presenta la opción A, se va a llevar cabo una modificación del modelo NMF utilizado para conseguir solventar esos problemas. El modelo que se va a definir para esta opción es similar al anterior, pero con una serie de modificaciones que son las que se van a analizar en esta opción.

El hecho de inicializar la matriz de bases armónicas con una matriz previamente entrenada con bases correspondientes al piano, es para evitar el aprendizaje de la matriz de bases armónicas mediante su restricción (sparseness).

En primer lugar, además de partir de información adicional para la parte vocal como ocurre en la opción anterior vamos a partir de información adicional correspondiente a la parte instrumental. Para ello vamos a inicializar las bases armónicas con una matriz de bases previamente entrenadas correspondientes al piano, ya que al fin y al cabo se puede considerar uno de los grandes instrumentos tomados como ejemplo a la hora de hablar de sonidos armónicos. Por lo tanto, es importante destacar que en esta opción las inicializaciones que se van a llevar a cabo con información útil, que sirven como información adicional acorde a conseguir una mejor separación son:

- La matriz de bases de excitaciones vocales $W_{F0_{F,Re}}$ se inicializa con el modelo de excitación global (donde cada columna contiene la excitación, en frecuencia, de un pitch) y se deja fija durante todo el proceso, en cambio, la matriz de bases de fonema $W_{F_{F,Rf}}$ se deja fija inicializándola previamente con bases de fonema entrenadas en una base de datos de voz.
- La matriz de bases correspondiente a los sonidos armónicos $W_{H_{F,Rh}}$ se inicializa con una matriz de bases armónicas previamente entrenadas correspondientes a las bases espectrales del piano.
- Las demás matrices se inicializan de la misma forma que en la opción anterior.

Esta modificación nos permitirá partir de una mayor cantidad de información útil, de esta forma habría menos cosas que aprender y se conseguiría una estimación de los distintos espectrogramas de una forma óptima.

Por otro lado, y para solventar el problema descrito anteriormente y que ocurre dentro de la 'Fase de Iteración' se pensó en utilizar una constante definida como $\lambda_{Dinámico}$ que nos permita ajusta el valor de la restricción en función del valor en el que se encuentre en ese momento el coste β-divergencia. Para ello es importante definir un valor de $\lambda_{Dinámico}$ para cada una de las matrices que participan en el proceso de separación, ya que al fin y al cabo lo que buscamos es que los valores de las restricciones que forman parte de las ecuaciones de actualización sean acordes a los valores de la parte del coste β-divergencia y de esta forma se apliquen todas las restricciones además de conseguir una buena reconstrucción final del espectrograma de salida.

La clave de conseguir una buena separación de las fuentes al fin y al cabo es conseguir que se apliquen esas restricciones dentro de su medida y evitar que el algoritmo se ciegue en el papel de reconstruir la señal final para conseguir un espectrograma de salida lo más similar al de entrada. Es por esto el motivo de utilizar esta constante que nos permite por así decirlo intentar ajustar los valores de las restricciones a los valores del coste β-divergencia, dicho con otras palabras, estaríamos forzando de alguna forma a que el sistema piense en separar aplicando las restricciones y no únicamente en reconstruir la señal final. Las nuevas reglas de actualización tomarían la siguiente forma:

- **Reglas de actualización multiplicativas para la base $\boldsymbol{H_{F0_{Re,T}}}$.**

$$H_{F0} = H_{F0} \odot \frac{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{-} + \lambda_{Dinámico_{H_{F0}}} K_{SAHFO} \left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{-}}{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{+} + \lambda_{Dinámico_{H_{F0}}} K_{SAHFO} \left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{+}} \tag{71}$$

$$\lambda_{Dinámico_{H_{F0}}} = P_{H_{F0}} \frac{d\beta\left(X_{n\beta} | \left(\widehat{X_n}\right)\right)}{K_{SAHFO} SAHFO} \tag{72}$$

Donde:

- K_{SAHFO}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de las excitaciones vocales.
- SAHFO: es el coste correspondiente a la monofonía para la matriz de las activaciones de las excitaciones vocales.
- $P_{H_{F0}}$: es el peso que se le va a dar a esta restricción con respecto a las otras.

- **Reglas de actualización multiplicativas para la base $H_{F_{Rf,T}}$.**

$$H_F = H_F \odot \frac{\left[\frac{\partial d\beta}{\partial H_F}\right]^- + \lambda_{Dinámico_{H_F}} K_{SAHF} \left[\frac{\partial SAHF}{\partial H_F}\right]^-}{\left[\frac{\partial d\beta}{\partial H_F}\right]^+ + \lambda_{Dinámico_{H_F}} K_{SAHF} \left[\frac{\partial SAHF}{\partial H_F}\right]^+} \quad (73)$$

$$\lambda_{Dinámico_{H_F}} = P_{H_F} \frac{d\beta\left(X_{n\beta}|\left(\widehat{X_n}\right)\right)}{K_{SAHF} SAHF} \quad (74)$$

Donde:

- K_{SAHF}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de los filtros vocales.
- SAHF: es el coste correspondiente a la monofonía para la matriz de las activaciones de los filtros vocales.
- P_{H_F}: es el peso que se le va a dar a esta restricción con respecto a las otras.

- **Reglas de actualización multiplicativas para la base $W_{P_{F,Rp}}$.**

$$W_p = W_p \odot \frac{\left[\frac{\partial d\beta}{\partial W_p}\right]^- + \lambda_{Dinámico_{W_p}} K_{SMM} \left[\frac{\partial SSM}{\partial W_p}\right]^-}{\left[\frac{\partial d\beta}{\partial W_p}\right]^+ + \lambda_{Dinámico_{W_p}} K_{SMM} \left[\frac{\partial SSM}{\partial W_p}\right]^+} \quad (75)$$

$$\lambda_{Dinámico_{W_p}} = P_{W_p} \frac{d\beta\left(X_{n\beta}|\left(\widehat{X_n}\right)\right)}{K_{SSM} SSM} \quad (76)$$

Donde:

- K_{SSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las bases percusivas.
- SSM: es el coste correspondiente a la restricción de suavidad para la matriz de las bases percusivas.
- P_{W_p}: es el peso que se le va a dar a esta restricción con respecto a las otras.

- **Reglas de actualización multiplicativas para la base** $H_{P_{Rp,T}}$.

$$H_p = H_p \odot \frac{\left[\frac{\partial d\beta}{\partial H_p}\right]^- + \lambda_{Dinámico_{H_p}} K_{TSP} \left[\frac{\partial TSP}{\partial H_p}\right]^-}{\left[\frac{\partial d\beta}{\partial H_p}\right]^+ + \lambda_{Dinámico_{H_p}} K_{TSP} \left[\frac{\partial TSP}{\partial H_p}\right]^+} \quad (77)$$

$$\lambda_{Dinámico_{H_p}} = P_{H_p} \frac{d\beta\left(X_{n\beta}|\left(\widehat{X_n}\right)\right)}{K_{TSP} TSP} \quad (78)$$

Donde:

- K_{TSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las activaciones percusivas.
- TSP: es el coste correspondiente a la restricción de dispersión para la matriz de las activaciones percusivas.
- P_{H_p}: es el peso que se le va a dar a esta restricción con respecto a las otras.

- **Reglas de actualización multiplicativas para la base** $W_{H_{F,Rh}}$**.**

$$W_H = W_H \odot \frac{\left[\frac{\partial d\beta}{\partial W_H}\right]^- + \lambda_{Dinámico_{W_H}} K_{SSP} \left[\frac{\partial SSP}{\partial W_H}\right]^-}{\left[\frac{\partial d\beta}{\partial W_H}\right]^+ + \lambda_{Dinámico_{W_H}} K_{SSP} \left[\frac{\partial SSP}{\partial W_H}\right]^+} \quad (79)$$

$$\lambda_{Dinámico_{W_H}} = P_{W_H} \frac{d\beta\left(X_{n\beta}|\left(\widehat{X_n}\right)\right)}{K_{SSP} SSP} \quad (80)$$

Donde:

- K_{SSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las bases armónicas.
- SSP: es el coste correspondiente a la restricción de dispersión para la matriz de las bases armónicas.
- P_{W_H}: es el peso que se le va a dar a esta restricción con respecto a las otras.

- **Reglas de actualización multiplicativas para la base $H_{H_{Rh,T}}$.**

$$H_H = H_H \odot \frac{\left[\frac{\partial d\beta}{\partial H_H}\right]^- + \lambda_{Dinámico_{H_H}} K_{TSM} \left[\frac{\partial TSM}{\partial H_H}\right]^-}{\left[\frac{\partial d\beta}{\partial H_H}\right]^+ + \lambda_{Dinámico_{H_H}} K_{TSM} \left[\frac{\partial TSM}{\partial H_H}\right]^+} \quad (81)$$

$$\lambda_{Dinámico_{H_H}} = P_{H_H} \frac{d\beta\left(X_{n\beta} | (\widehat{X_n})\right)}{K_{TSM} TSM} \quad (82)$$

Donde:

- K_{TSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las activaciones armónicas.
- SSP: es el coste correspondiente a la restricción de suavidad para la matriz de las activaciones armónicas.
- P_{W_H}: es el peso que se le va a dar a esta restricción con respecto a las otras.

El valor de los distintos pesos (P_x) asociado a cada uno de los $\lambda_{Dinámico}$ es el que nos permite darle mayor importancia a una restricción con respecto a la otra.

Una vez que se han explicado las distintas modificaciones que se han llevado a cabo para mejorar el modelo que lleva a cabo la primera fase del segmentador es necesario, al igual que se hizo con la opción A, mostrar el espectrograma estimado de la parte vocal y así poder compararlo con el de la opción anterior.

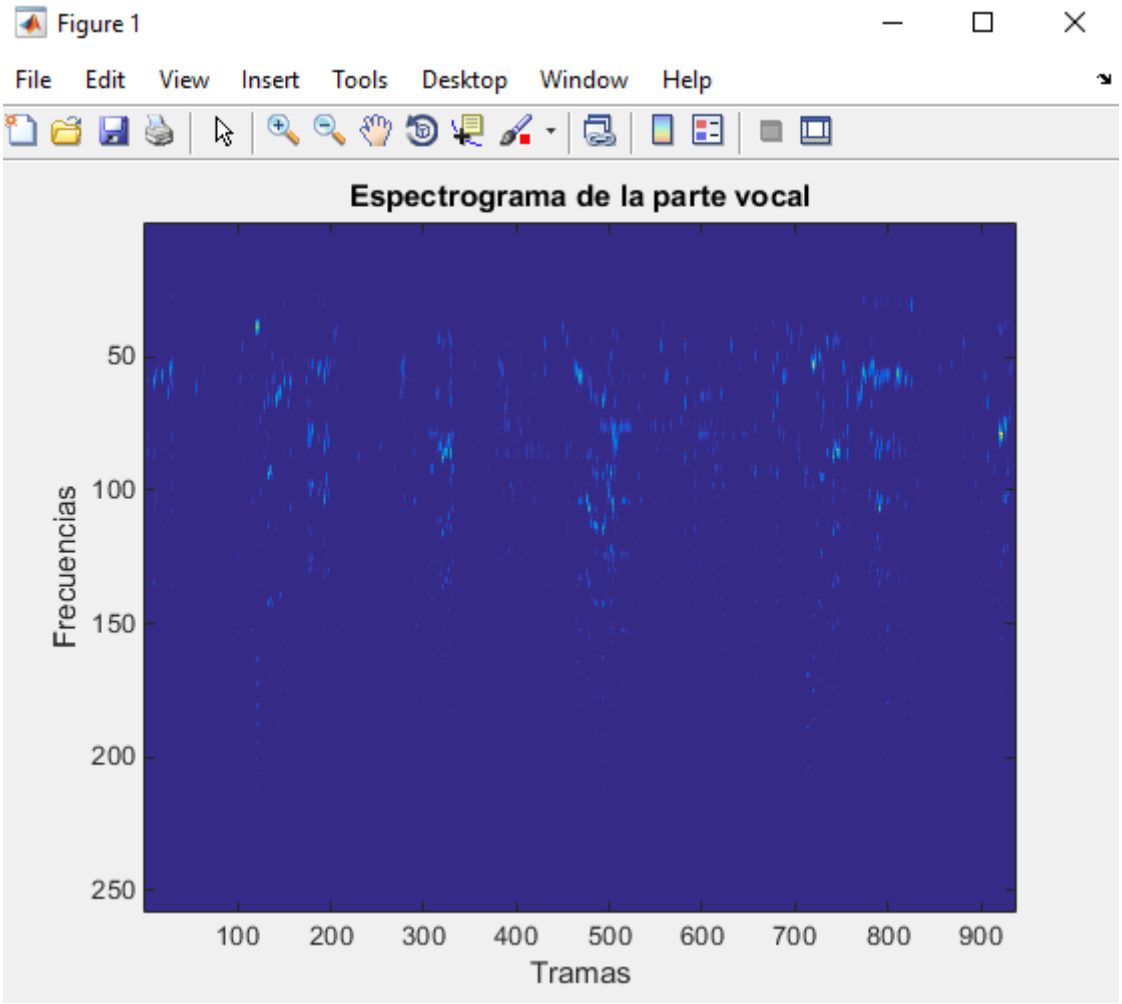

Figura 4.8: Espectrograma de la parte vocal opción B

Como se puede observar en el espectrograma vocal estimado mostrado en la Figura 4.8, los resultados a simple vista son mucho mejores que los obtenidos con la opción anterior. Con esta opción podemos hacer un análisis visual e identificar las zonas vocales de una forma sencilla como se muestra en la figura anterior. Por lo tanto, han quedado solucionados los problemas que se presentaron al inicio.

4.1.2 FASE 2: Discriminador de zonas instrumentales y vocales

Una vez que en la fase anterior se ha conseguido llevar a cabo una primera identificación de las zonas vocales y por consiguiente de las zonas instrumentales, necesitamos realizar una interpretación del espectrograma vocal obtenido en la fase anterior para determinar y clasificar las distintas zonas como instrumentales o como vocales. Este paso al igual que el anterior es crucial para conseguir llevar a cabo una correcta segmentación de cualquier señal musical, ya que en la primera fase estamos obteniendo una información visual que nos permitirá en la segunda fase identificar lo que se consideran como zonas vocales o instrumentales.

Durante la realización del Trabajo Fin de Master se utilizaron una serie de técnicas para implementar el segmentador automático que no dieron los resultados que se buscaban. Es necesario la explicación de las técnicas implementadas previas a la técnica óptica para evitar una posterior utilización de ellas a la hora de realizar segmentadores automáticos futuros. Estas técnicas, que resultaron fallidas, surgieron por consecuencia de analizar cada una de las señales musicales de forma individual, cometiendo el fallo de centrarse en una única canción o un grupo de canciones reducido. En esta sección se detallarán en primer lugar estas técnicas para desmontarlas y dar una explicación del porqué no siempre funciona.

Por otro lado, una vez que se ha llevado a cabo un análisis de lo que no debemos de utilizar como se hizo en la fase anterior, se va a llevar a cabo la explicación del modelo finalmente utilizado, cuyos resultados son mejores y se pueden considerar óptimos.

4.1.2.1 Modelo basado en la divergencia

La razón principal que motivo la utilización de la divergencia como base para la segunda fase del segmentador automático es la siguiente:

- Es difícil encontrar una canción en la que las zonas iniciales y finales tenga la fuente vocal, ya que al fin y al cabo se puede apreciar que siempre empiezan con una base musical introductoria y de igual forma terminan con una base musical para finalizar. Por lo tanto, si aplicamos la divergencia entre la señal musical original y la señal instrumental (suma de la fuente armónica y la fuente percusiva) obtenida de la fase anterior se podría seleccionar de alguna forma las zonas instrumentales aplicando un determinado umbral. Esto se debe a que en las zonas donde únicamente hay fuentes instrumentales la divergencia proporcionara resultados muy pequeños, mientras que

en aquellas zonas donde además existen fuentes vocales los resultados de la divergencia serán mucho mayores.

Para realizar una visión general del sistema se ha realizado un esquema, mostrado en la Figura 4.9, donde se puede observar las distintas fases del sistema implementado el cual se basa en el cálculo de la función de coste β-divergencia.

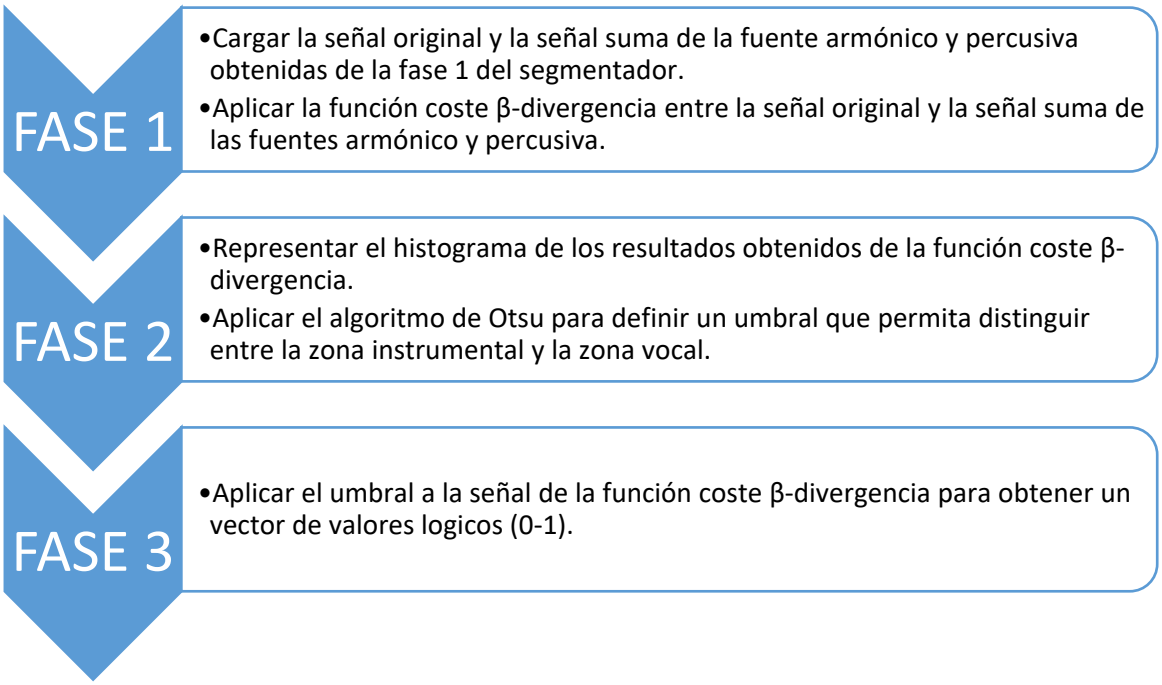

Figura 4.9: Representación de las fases que componen el sistema segmentador automático basado en la función coste β-divergencia

En la primera fase se debe cargar la señal musical original y la señal suma de la fuente armónica y percusiva obtenidas la primera fase del segmentador. Para ello y a la hora de cargar cualquier señal en Matlab se ha utilizado la función que nos permite cargar un directorio determinado, para que en el caso de que existan más de una canción, el proceso se pueda realizar de forma iterativa para todas las canciones existentes en la carpeta seleccionada. Las señales con las que se va a trabajar son:

- $\boldsymbol{Xm}$: Señal musical original
- $\boldsymbol{Xp}$ + $\boldsymbol{Xh}$: Señal suma de la fuente armónica y la fuente percusiva obtenida del primer modelo de separación NMF.

El siguiente paso sería aplicar la función coste β-divergencia descrita anteriormente, mediante la explicación del modelo NMF, entre las dos señales con las que trabajar $d\beta\left(X_m|(X_p + X_h)\right)$.

Se representa a continuación la señal de divergencia para la canción seleccionada a métodos de ejemplo para el análisis de los siguientes algoritmos y métodos.

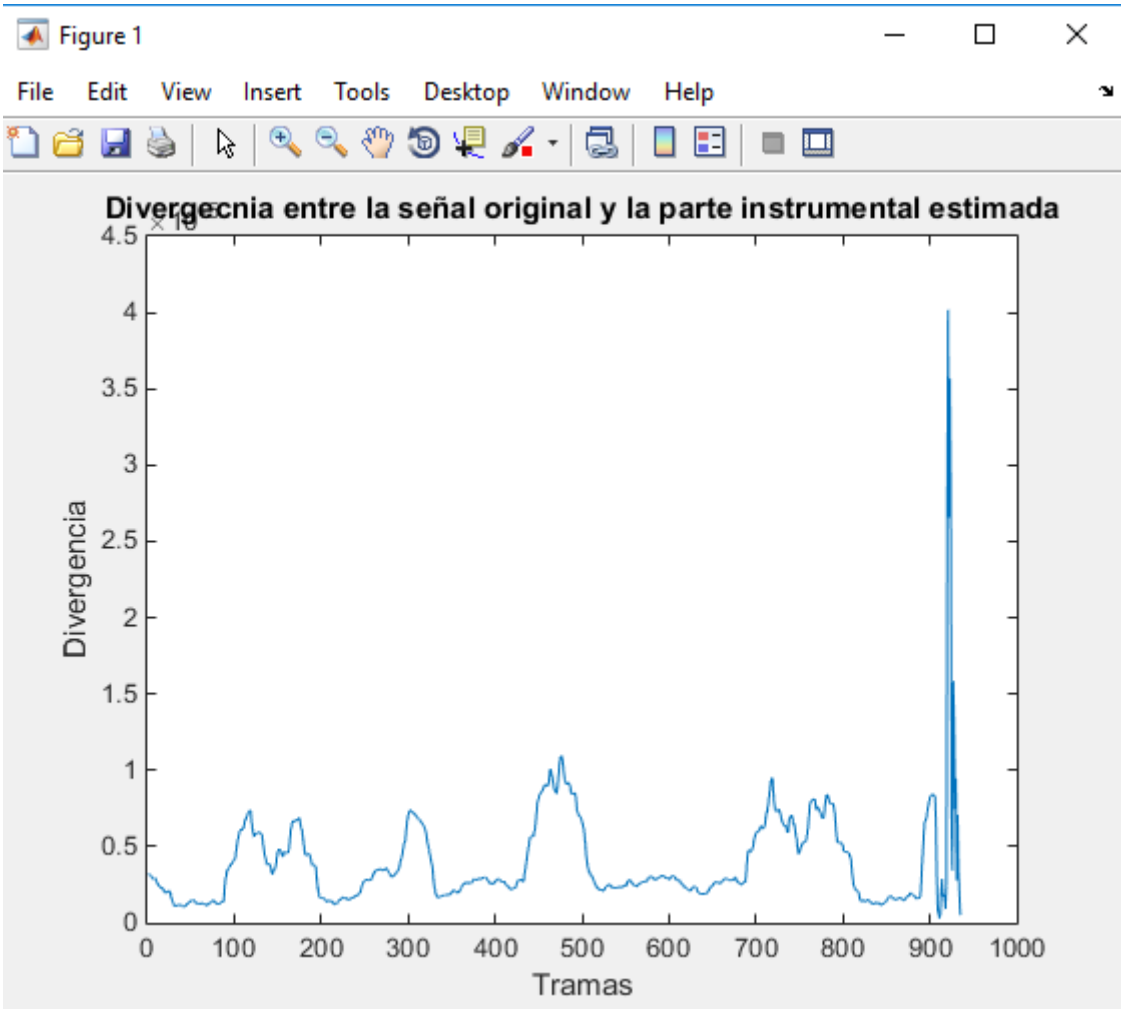

Figura 4.10: Representación de la función Divergencia entre la señal original y la parte instrumental estimada.

Como se puede observar en la Figura 4.10, la función de divergencia a simple vista permite detectar las zonas instrumentales de las zonas vocales, ya que los resultados de divergencia son bajos para las zonas instrumentales (se parecen mucho con la original) y son altos para las zonas vocales (se parece poco con la señal estimada, ya que en estas zonas existe la parte vocal).

En la segunda fase del sistema se debe de realizar una representación del histograma de la función de divergencia para poder realizar una separación entre la zona instrumental y la zona vocal, para ello se representará y almacenará un histograma de 100 barras como se puede ver en la siguiente imagen.

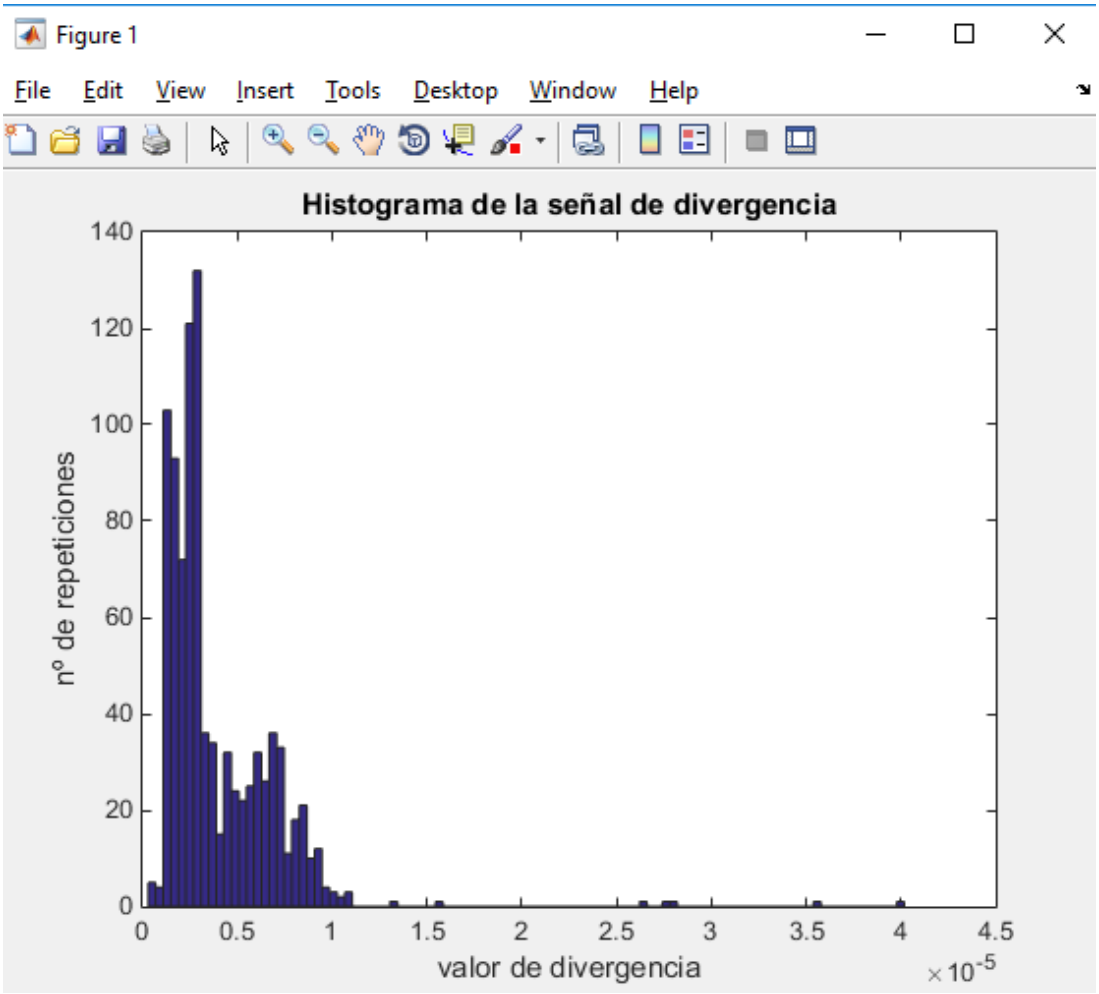

Figura 4.11: Histograma de la señal de divergencia

Se puede observar la existencia de dos grupos diferenciados, uno correspondiente a las zonas instrumentales (los de menor valor) y otro correspondiente a las zonas vocales (los de mayor valor).

Una vez que se ha almacenado el histograma correspondiente a la función divergencia debemos seleccionar el umbral que nos permita diferenciar entre lo que es música y lo que es voz, para ello se aplica el algoritmo de Otsu [65]. El método Otsu es utilizado principalmente para segmentar gráficos rasterizados, es decir separar los objetos de una imagen que nos interesen del resto de la imagen. Aunque si se conoce la técnica se puede aplicar a cualquier tipo de señal con el objetivo de determinar un umbral que nos permita realizar una separación entre dos grupos claramente diferenciados en un histograma.

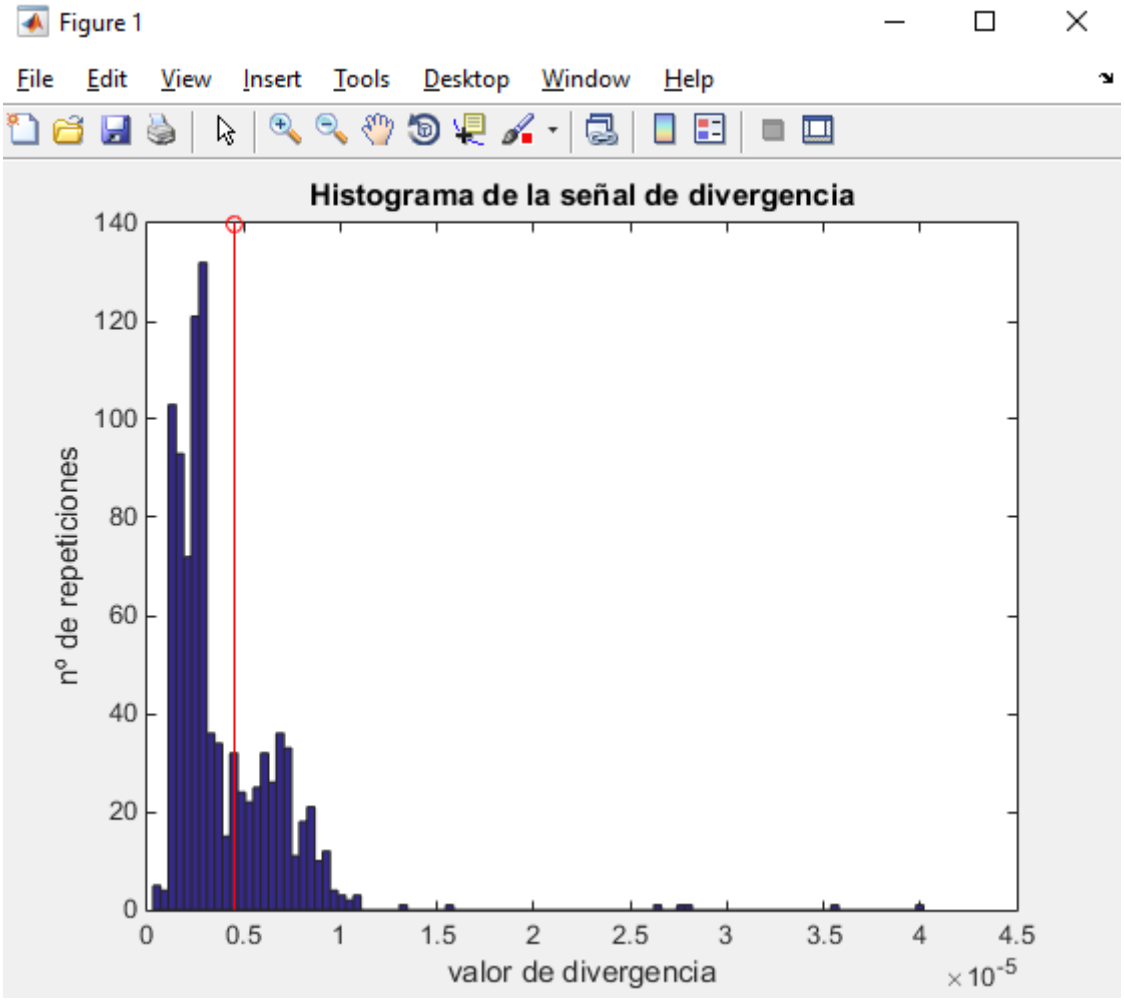

Figura 4.12: Representación del histograma con el umbral definido por Otsu para diferenciar las zonas instrumentales y las vocales.

Como se puede apreciar de la imagen anterior, el umbral de Otsu nos permite diferenciar claramente entre las zonas vocales y las instrumentales, ya que selecciona exactamente la mitad del valor del histograma. Por lo tanto, se podría decir, en una primera aproximación que la segmentación se realizaría de forma correcta.

La tercera fase corresponde con aplicar un umbral (el obtenido por el algoritmo de Otsu) a la función obtenida anteriormente de divergencia con el objetivo de crear un vector formado con valores lógicos, que permita diferenciar entre aquellos tramos correspondientes a la música y aquellos tramos correspondientes a la voz. Se representa a continuación el vector de valores lógicos que muestra las zonas instrumentales seleccionadas por el algoritmo

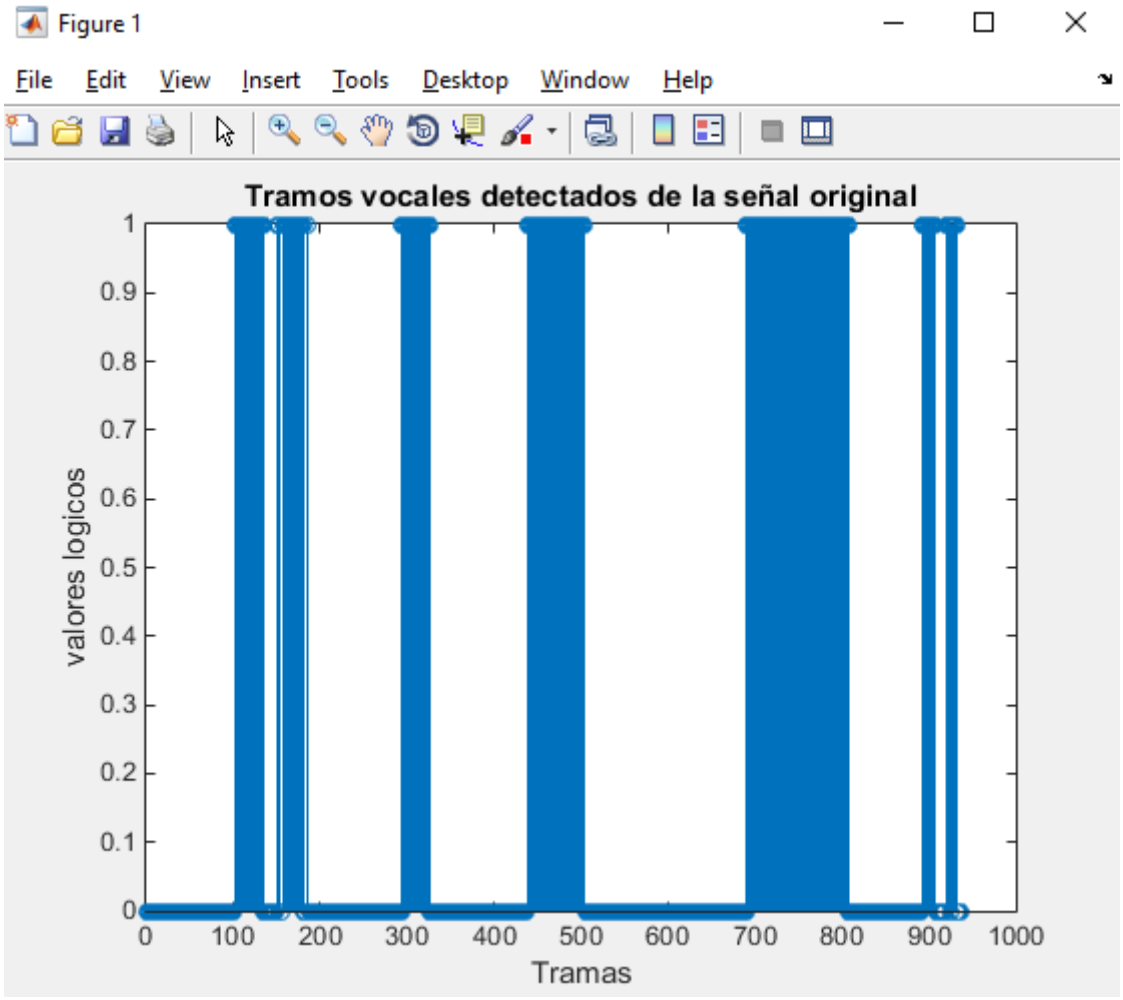

Figura 4.13: Zonas vocales detectadas por el segmentador basado en divergencia.

Los resultados que se obtienen al visualizar varios ejemplos y al compararlos con las zonas reales obtenidas del fichero original, es que el segmentador no ofrece buenos resultados al utilizar la divergencia, ya que para ello los resultados obtenidos de la primera etapa del segmentador tendrían que ser casi perfectos. Por lo tanto, en algunos casos funciona bien por los resultados de separación de la primera etapa son muy buenos, pero no llega a ser la técnica que nos permitirá llevar a cabo un segmentador automático.

4.1.2.2 *Modelo basado en la intensidad*

Las razones que motivaron la utilización del umbral de intensidad como base para la implementación del segmentador automático son las siguientes:

- Tras realizar un análisis con un grupo de canciones pertenecientes a distintos géneros musicales (formado por canciones de ópera y flamenco) se pudo comprobar que en un reducido grupo de ellas existía una característica común para la mayoría de ellas. En la mayoría de ese tipo de señales musicales, cuando comienza a aparecer la parte vocal el nivel de intensidad aumenta considerablemente en comparación al nivel de intensidad cuando solo existen zonas instrumentales. Por lo tanto, se puede poner un umbral de energía de tal forma que todo lo que se situé por debajo de ese umbral se etiquete como parte instrumental y el resto como parte vocal, realizando de esta forma una segmentación clara y sencilla.

Los pasos que se deberían seguir para la implementación de este segmentador basado en un umbral de intensidad serían los mismos que se han realizado en la implementación del segmentador basado en la función de divergencia, únicamente realizando un cambio. En este caso se calcula un umbral de intensidad mientras que en el caso anterior se calculaba un umbral de divergencia.

Para realizar una visión general del sistema se ha realizado un esquema, mostrado en la Figura 4.14, donde se puede observar las distintas fases del sistema implementado el cual se basa en el cálculo del umbral de intensidad.

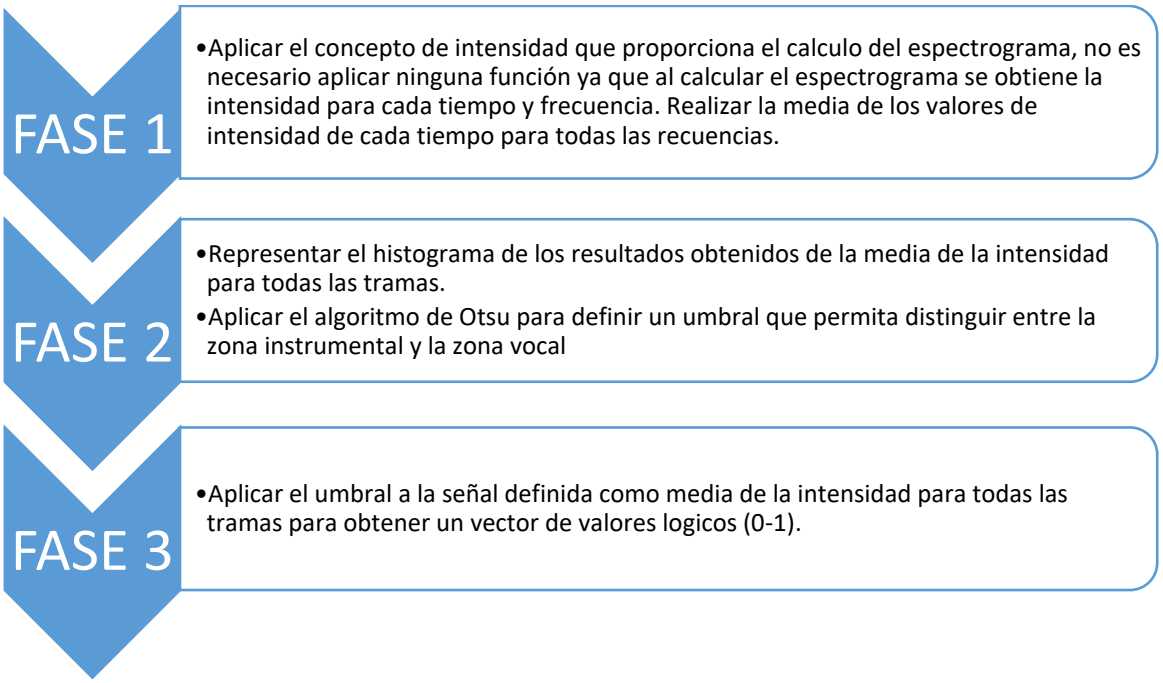

Figura 4.14: Representación de las fases que componen el sistema segmentador automático basado en un umbral de intensidad.

La única diferencia que presenta el sistema actualmente mencionado con el sistema segmentador basado en la función de la divergencia es la siguiente:

- En este sistema no se realiza un cálculo de la función de divergencia entre la señal musical original y la señal suma de las fuentes armónico y percusiva obtenida de la primera fase del segmentador, sino que se realiza el cálculo de la media de las intensidades para cada una de las tramas, obteniendo de esta forma un vector que almacena el valor medio de la intensidad (media calculada para todas las frecuencias) para cada una de las tramas. En la siguiente figura se puede observar la representación del espectrograma de la señal original de la canción con la que se va a trabajar a lo largo de la memoria del TFM Por lo tanto al fin y al cabo este segmentador se puede implementar de una forma muy sencilla ya que se parte de la señal original, por lo tanto, no requerirá de la fase anterior del segmentador.

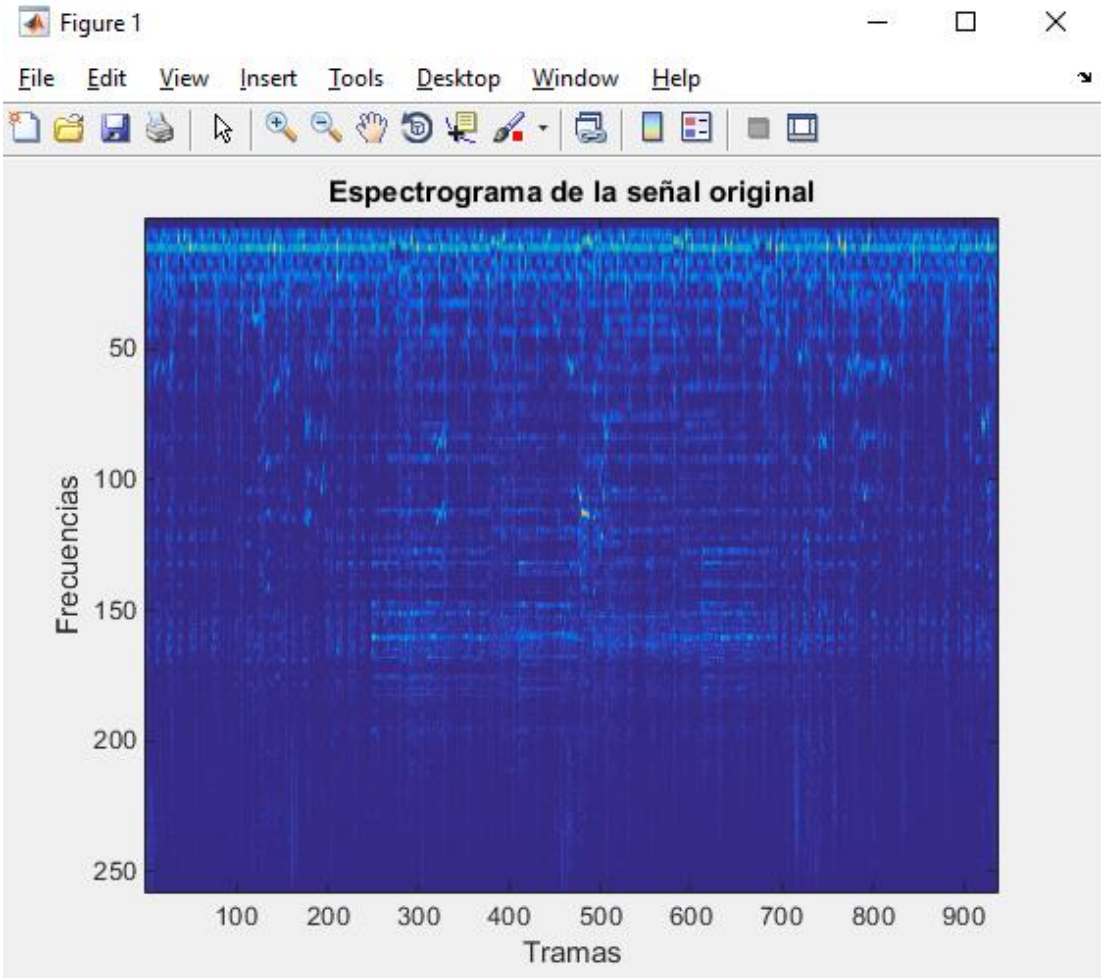

Figura 4.15: Espectrograma de la señal original.

Como se puede ver en el espectrograma de la señal original, mostrado en la Figura 4.15, es imposible determinar las zonas vocales o las zonas instrumentales partiendo del espectrograma de la señal de entrada, por lo tanto, esta técnica carece de sentido utilizarla en un segmentador. Como se ha mencionado anteriormente, esta técnica podría ser útil para canciones pertenecientes al género de la opera o para algunos palos del flamenco, ya que cuando aparece la fuente vocal el nivel de la intensidad aumenta considerable mente, sin embargo, esta teoría no es cierta para cualquier género musical, por lo que nos lleva a descartarla.

Se ha escogido esta canción porque representa un comportamiento muy común a la mayor cantidad de canciones existentes hoy en día, en las que la intensidad de la voz no tiene por qué ser mayor que la de los instrumentos, ni viceversa. De esta forma y como se puede observar en la imagen anterior, queda desmentida completamente la teoría y aunque sería una técnica útil en el género de la ópera y el del flamenco, no nos interesa para nuestro segmentador que se ha llevado a cabo con un carácter general, para cualquier género musical.

Como conclusión final lo más importante a tener en cuenta es:

- El segmentador basado en la divergencia es útil cuando no existe mucha complejidad en términos musicales, es decir cuando no existen muchos instrumentos musicales y además mucha variación entre ellos. Sin embargo, esto no es así ya que teniendo en cuenta el amplio abanico que engloba al grupo de la música no podemos basarnos en esta técnica.
- Por otro lado, el segmentador basado en intensidad tampoco es útil, debido a que en la mayoría de las señales musicales no aparece una fuente con una intensidad dominante, es decir, no tiene porque la fuente vocal tener más intensidad que las fuentes instrumentales y viceversa.
- Por lo tanto, han quedado una vez que han quedado desmentidas las dos teorías anteriores se va a llevar a cabo el algoritmo óptimo de segmentación, en el cual se va a tomar como partida el espectrograma de la parte vocal determinado en la primera fase del segmentador.
- Añadir además que en las dos opciones vistas anteriormente no se buscaba mostrar un comportamiento correcto de esa fase del segmentador, sino todo lo contrario, justificar por qué no se deben de utilizar esas técnicas.
- Por último y para concluir, la clave para llevar a cabo un segmentador de carácter general es no cerrar la puerta a ningún tipo de género y no centrarse en señales musicales específicas. Es decir, aunque trabajes con una a modo de ejemplo trata en primer lugar que no tenga un comportamiento especifico y en segundo lugar trata de trabajar con muchas otras y de distintos tipos.

4.1.2.3 Modelo basado en la máscara vocal del sistema NMF

Una vez que se han analizado las técnicas a través de las cuales no se va a implementar el sistema y se ha justificada las distintas razones que nos han llevado a esa conclusión, se va a llevar a cabo un análisis de la técnica que finalmente se va a implementar en esta segunda fase del segmentador.

En primer lugar, es necesario tener claro que nuestro objetivo en esta fase es interpretar los resultados de la primera fase del segmentador, de tal forma que consigamos determinar las distintas zonas presentes en nuestra señal de partida.

La señal con la que vamos a partir en esta fase es el espectrograma de la señal vocal. En este espectrograma se encuentra toda la información necesaria para determinar qué zonas son instrumentales y que zonas serian vocales.

La primera de las razones porque se ha utilizado la técnica que se describe a continuación es para intentar llevar a cabo un sistema de segmentación/separación en el cual la mayor parte del ámbito técnico quede implementado mediante los modelos NMF.

Teniendo en cuenta que los resultados obtenidos en la primera fase del segmentador cumplen con las expectativas deseadas, es lógico pensar que partiendo del espectrograma vocal obtenido se puede hacer un análisis del mismo para determinar las distintas zonas presentes en la señal.

Y sin más demora comenzamos con la explicación de la técnica optimizada para esta fase del segmentador.

En la fase anterior se han obtenido los espectrogramas correspondientes a cada una de las fuentes presentes en la señal musical (armónico/percusivo/vocal). Como ya se ha comentado un espectrograma tiene dimensiones de Tramas/Frecuencias y la información se recoge en niveles de energía que presenta cada una de las tramas a una frecuencia determinada.

Nosotros para llevar a cabo la implementación de esta etapa vamos a utilizar el concepto de máscara. Estas más caras son las que nos van a permitir luego en la etapa del separador pasar de los espectrogramas obtenidos a las señales definidas en el tiempo, multiplicando esa mascara por un espectrograma complejo relacionado con la señal mezcla $x(t)$, es decir, la señal original.

Existe una máscara definida para cada una delas fuentes como se pueden ver en las siguientes ecuaciones.

$$M_p = \frac{X_p}{X_p + X_h + X_v} \tag{83}$$

$$M_h = \frac{X_h}{X_p + X_h + X_v} \tag{84}$$

$$M_v = \frac{X_v}{X_p + X_h + X_v} \tag{85}$$

Y como se ha dicho antes multiplicando esa máscara por un espectrograma complejo relacionado con la señal mezcla $x(t)$, es decir, la señal original, se puede obtener cada una de las fuentes sonoras separadas en el tiempo aplicando la IDFT (Inverse Discrete Fourier Transform).

$$x_p(t) = IDFT(M_p X_c) \tag{86}$$

$$x_h(t) = IDFT(M_h X_c) \tag{87}$$

$$x_v(t) = IDFT(M_v X_c) \tag{88}$$

Esto se aplicará en su totalidad en la etapa del separador, sin embargo, para esta fase del segmentador lo que nos interesa es determinar donde se localizan las zonas vocales a partir de la máscara del espectrograma vocal.

En la siguiente imagen mostramos el aspecto que tiene la máscara de la parte vocal.

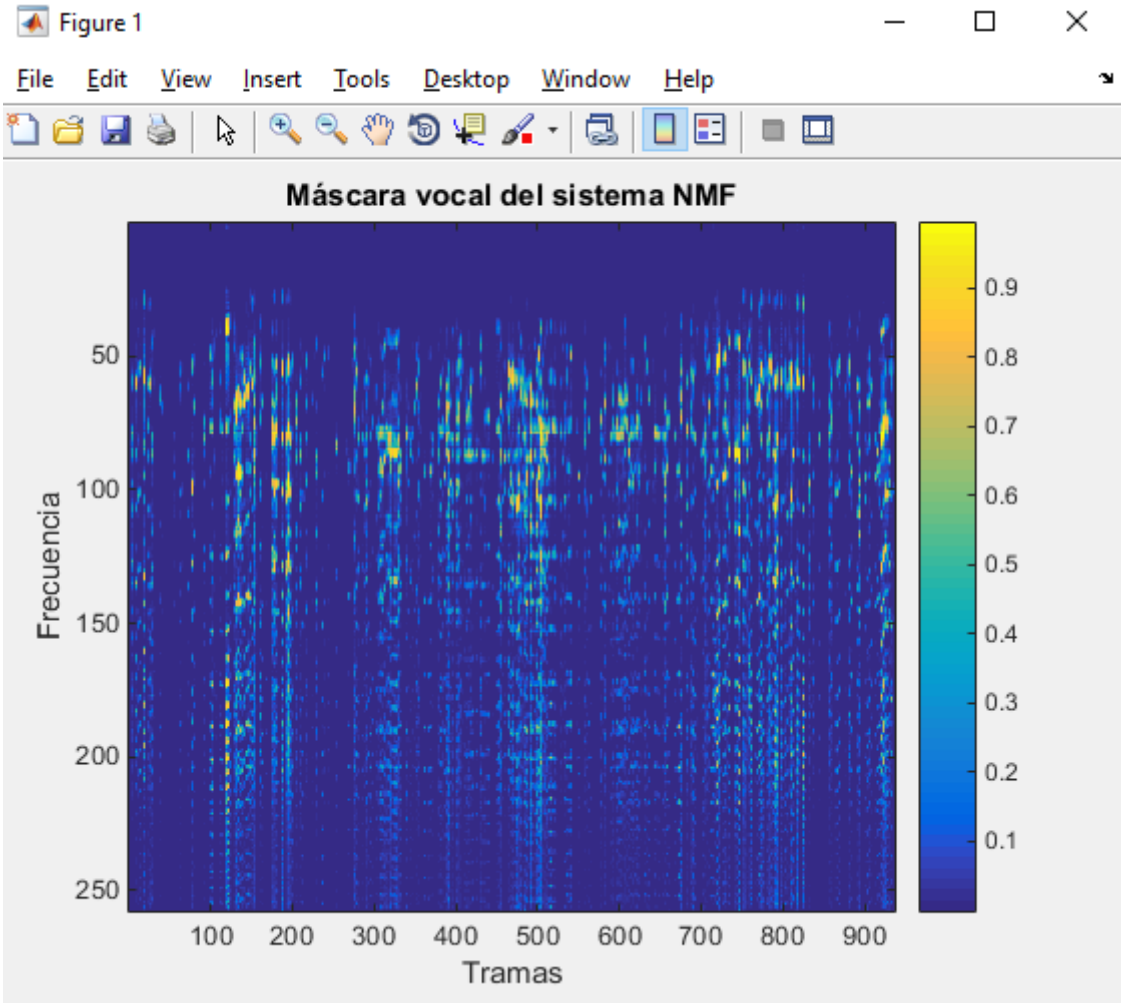

Figura 4.16: Máscara de la parte vocal extraída del sistema NMF de la fase anterior.

Como se puede ver en la imagen anterior es necesario restringir de alguna forma la información obtenida para que únicamente se quede con la información que realmente pertenece a la parte vocal. Para ello se utiliza un determinado peso, que nos permitirá limitar la información vocal que se puede ver en la imagen anterior y considerar que únicamente estamos seleccionando lo que se considera realmente parte vocal como se muestra en la siguiente imagen.

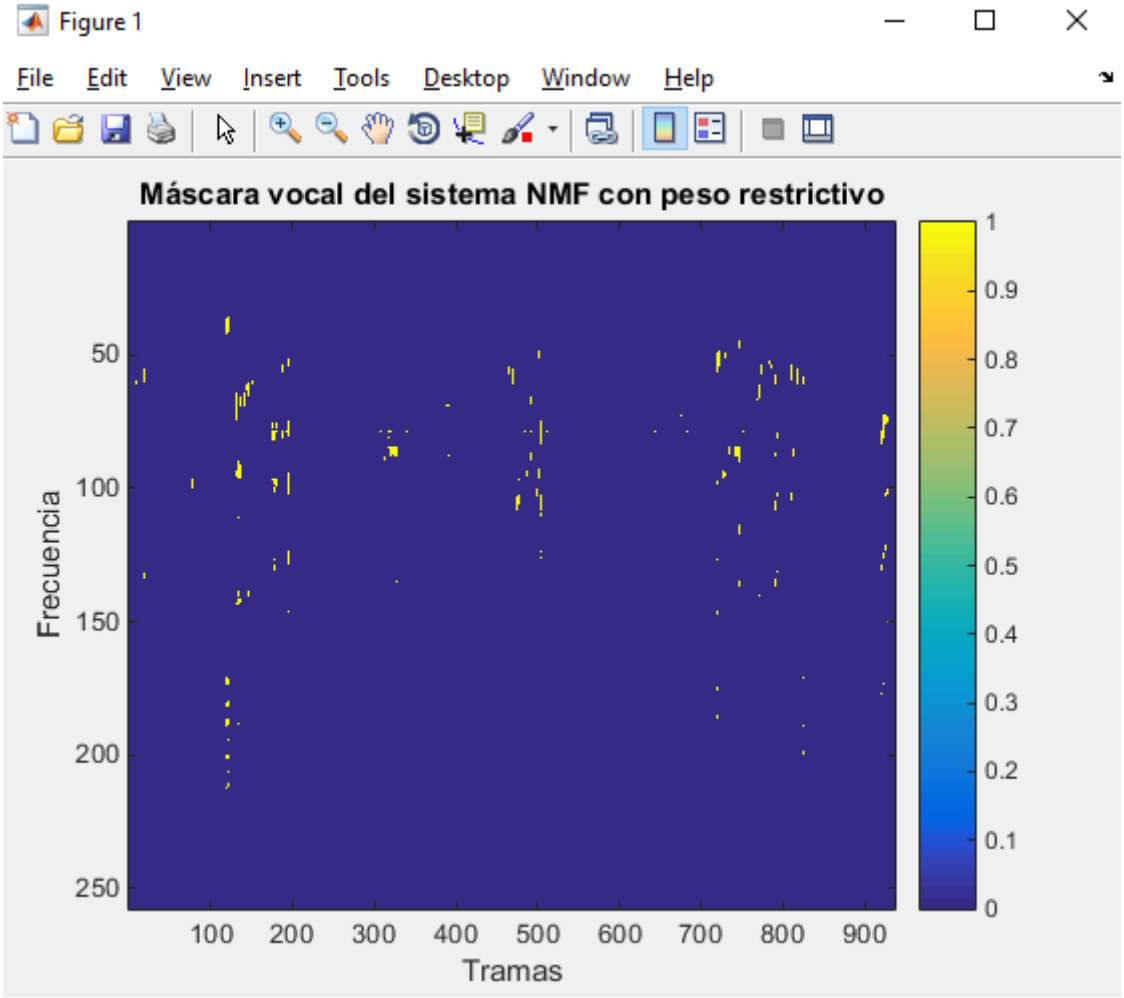

Figura 4.17: Máscara vocal extraída del sistema NMF de la fase anterior con un peso restrictivo añadido.

Una vez que se ha restringido de alguna forma la información, para únicamente visualizar en gran medida la parte vocal, es necesario implementar un umbral por intensidad. El procedimiento es similar al que se utiliza cuando se ha explicado el segmentador por intensidad visto anteriormente, pero en este caso la señal de entrada sería la máscara de la parte vocal, con un determinado peso para ajustarla y hacerla más restrictiva a seleccionar zonas instrumentales que no lo sean (falsos positivos).

Por lo tanto, lo siguiente es realizar la media en intensidad para cada una de las tramas que forman la señal de partida (Máscara vocal con el peso restrictivo). La media se va realizando para cada una de las tramas de tal forma que se suma la energía total de todas las frecuencias presentes en la trama y se divide entre el nº de frecuencias.

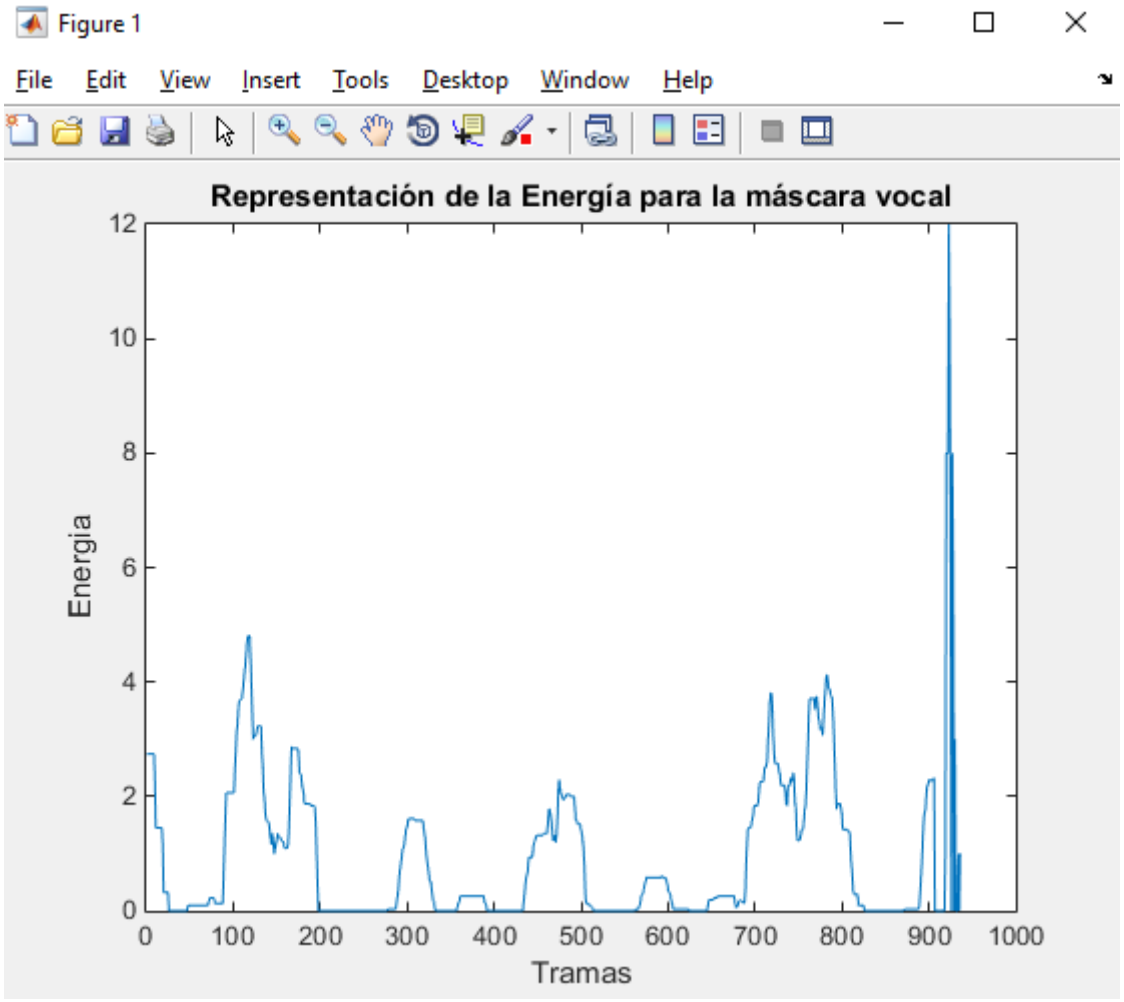

Figura 4.18: Representación de la Energía para la máscara vocal con el peso restrictivo.

Como se puede apreciar en la Figura 4.18 la energía nos marca claramente donde se localizan las zonas vocales, ya que disponen de una energía mucho mayor que las zonas consideradas como zonas instrumentales, por lo tanto, el objetivo de realizar una segmentación lo más óptima posible se está llevando a cabo.

El siguiente paso del sistema consiste en la realización de una representación del histograma de la media de los valores de energía obtenidos para poder realizar una separación entre la zona instrumental y la zona vocal (la zona vocal corresponderá con la de mayor energía, mientras que la zona instrumental apenas tendrá energía), para ello se representará y almacenará un histograma de 100 barras como se puede ver en la siguiente imagen.

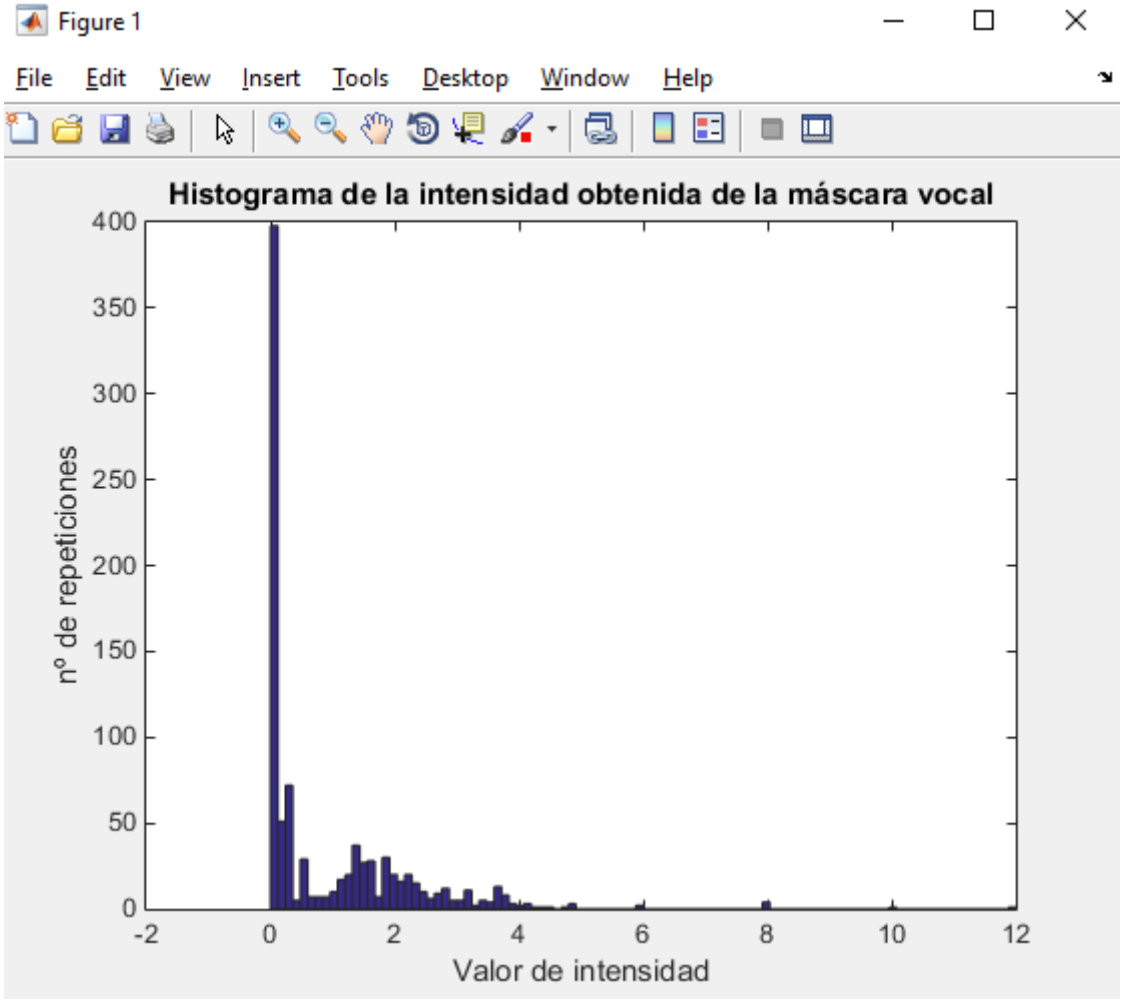

Figura 4.19: Histograma del valor de intensidad de la máscara vocal.

Una vez que se han almacenado el histograma correspondiente a la media de las energías obtenidas de la máscara debemos seleccionar el umbral que nos permita diferenciar entre lo que es música y lo que es voz, para ello se aplica el algoritmo de Otsu mencionado anteriormente.

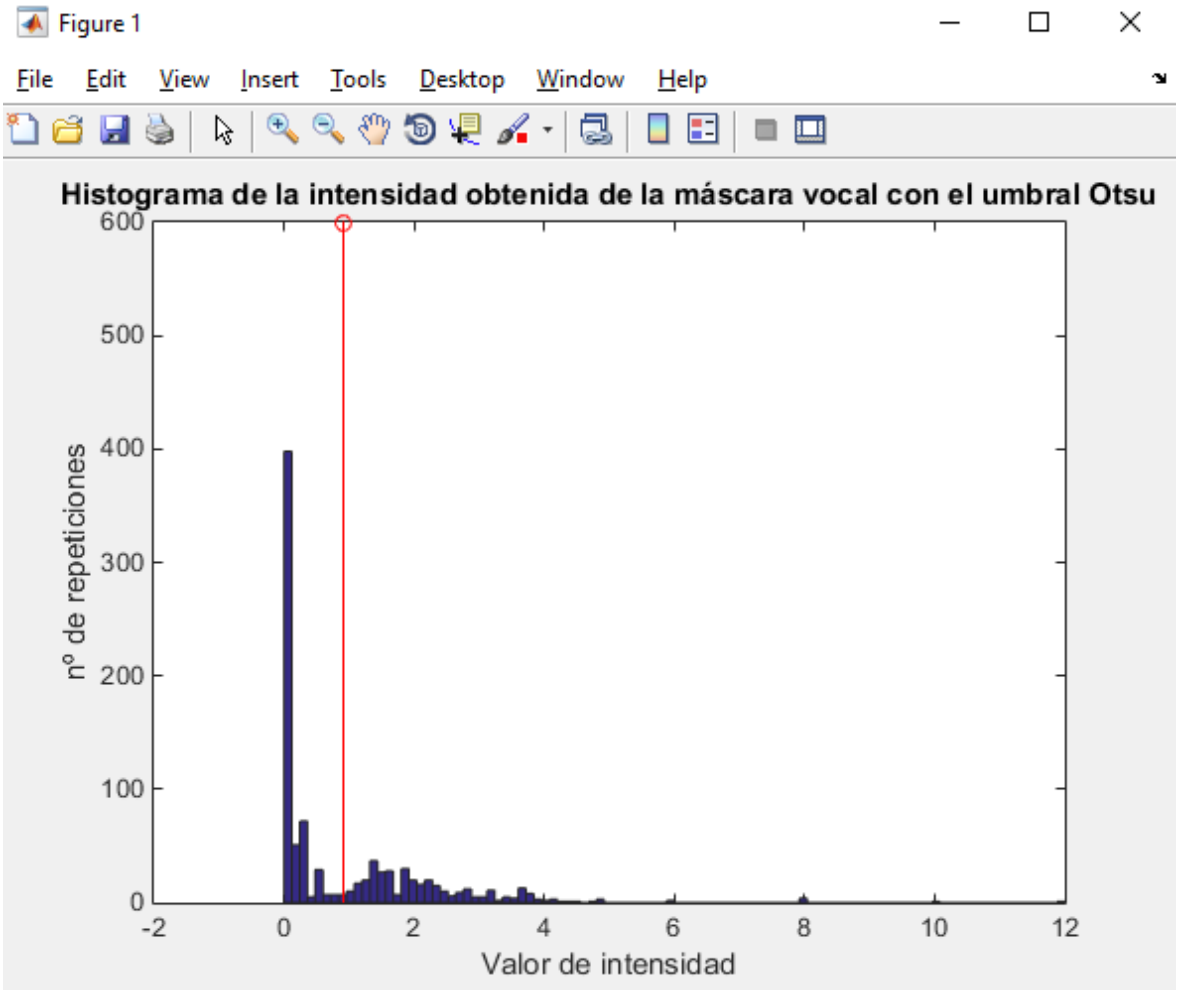

Figura 4.20: Histograma de la intensidad con el umbral Otsu

Como se puede ver en la Figura 4.20 existen dos grupos dentro del histograma, uno corresponde con las zonas vocales que son las que tienen mayor nivel de intensidad, y el otro corresponde con las zonas instrumentales que son las que tienen menor nivel de intensidad.

Finalmente se debe aplicar el umbral obtenido del algoritmo de Otsu a la señal de los valores medios de energía por trama obtenida de la máscara vocal restrictiva. De esta forma se crea un vector formado con valores lógicos, que permite diferenciar entre aquellos tramos correspondientes a la musical y los correspondientes a la voz. Se representa a continuación el vector de valores lógicos que muestra las zonas instrumentales seleccionadas por el algoritmo.

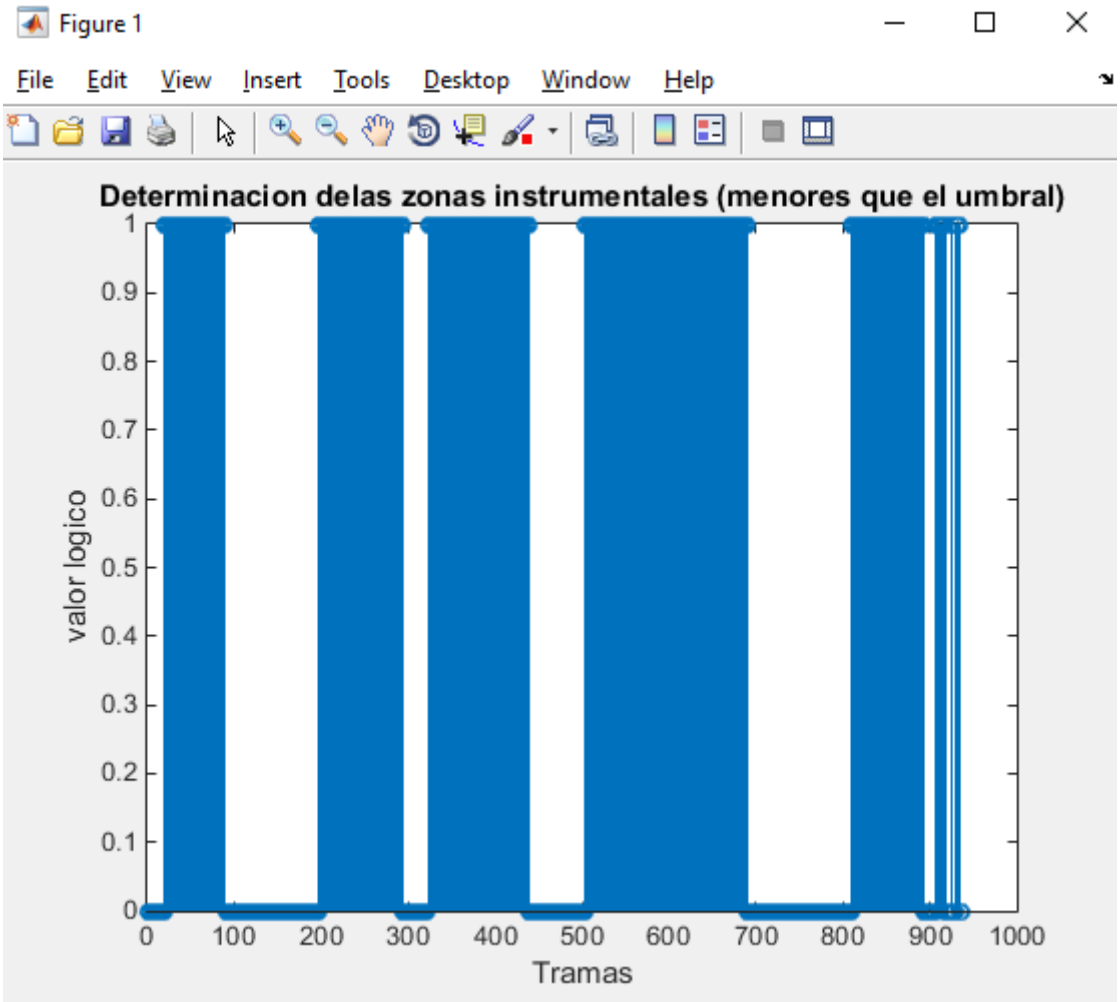

Figura 4.21: Zonas instrumentales detectadas por la fase dos del algoritmo de segmentación.

Por lo tanto, se ha conseguido llevar a cabo una discriminación de lo que se considera parte musical con respecto a lo que se considera parte vocal. En las fases posteriores se tratará de mejorar esta primera aproximación para conseguir unos mejores resultados de segmentación. Por ejemplo, se conseguirá determinar si el comportamiento del final de la canción es zona instrumental o vocal que en esta segunda fase queda un poco difuso.

4.1.3 FASE 3: Decisor de tramos instrumentales

Una vez que se ha implementado la segunda fase del segmentador automático, que corresponde al modelo de discriminación de parte musical y parte vocal aplicando el concepto de máscara vocal al espectrograma de la señal vocal obtenido de la fase 1 del segmentador automático, el siguiente paso sería la implementación de una función que permita decidir cuáles son los tramos musicales y cuáles son los que corresponden a la voz.

Si se observa la Figura 4.21 se puede determinar de forma sencilla cuales son las zonas instrumentales y cuáles son las zonas vocales, por lo tanto, si el ser humano es capaz de interpretar estos resultados de forma visual solo queda implementar una función que permita hacerlo computacionalmente.

Al igual que como se ha hecho en las fases anteriores, en primer lugar, se explica una primera opción que se va a desmentir y en segundo lugar se explica la opción que se ha adoptado para este Trabajo Fin de Master debido a que ofrece mejores resultados.

4.1.3.1 Análisis mediante ventanas y aplicación del umbral Otsu.

Para realizar esta función se ha implementado un sistema basado en ventanas, de tal forma que mediante un bucle se recorra la señal obtenida de la fase anterior, de manera que, si la mayoría de las tramas que componen la ventana superan un determinado umbral, la ventana se etiquetara como zona instrumental y en caso contrario como zona vocal. Las características que componen dicho sistema son las siguientes:

• Se realiza un recorrido de la señal mediante un enventanado, de tal forma que cada una de las ventanas tenga una duración de 1.5 segundos, que si se expresa en tramas tendrán un tamaño de 300 tramas.

• El umbral vendrá fijado por el algoritmo de Otsu (explicado anteriormente en la memoria del Trabajo Fin de Master). Dicho umbral será compensado por un factor de incremento que permita realizar un ajuste optimizado del mismo.

• El número de tramas que deben superar el umbral corresponde al 60% del tamaño de la ventana (200 tramas).

• Esta función nos permite realizar un análisis eficaz de la señal, ya que de esta forma se toma en consideración tramos suficientemente grandes como música o como voz

y se evita la aparición de tramas aisladas que conduzcan a error o a la aparición de futuros artefactos en el proceso de separación NMF.

La función que se ha implementado en MATLAB es la siguiente:

```
function [decision] = decisor(prom, T, inc, ajus)
```

Donde las variables de entrada son:

- **input:** corresponde con la señal de salida de la segunda fase del segmentador.
- **T:** se refiere al tamaño de la ventana que permitirá recorrer la salida de la fase anterior y analizarla de forma eficaz. El valor seleccionado es T=300 (1.5 segundos).
- **inc:** corresponde al incremento que se desea realizar en el umbral tras la aplicación del algoritmo de Otsu. El valor seleccionado es 1.18 (optimizado mediante análisis).
- **ajus:** se refiere al número de tramas que deben de superar el umbral escogido para asignar a la ventana como zona instrumental. El valor seleccionado es T=200.

En la Figura 4.22 se puede observar la sucesión de los distintos pasos para la implementación del decisor de zonas instrumentales.

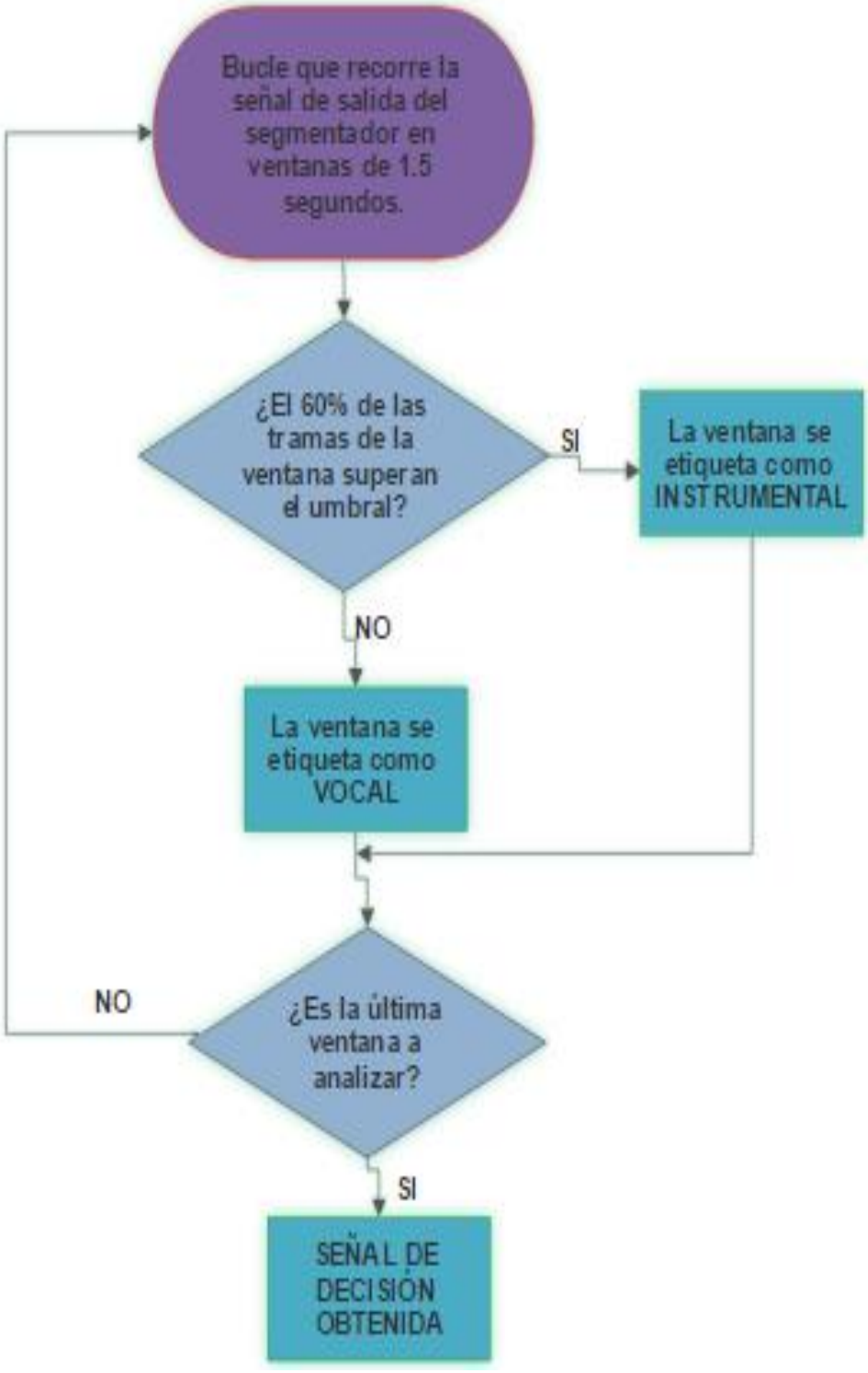

Figura 4.22: Diagrama de flujo de la implementación del decisor de zonas instrumentales

En la Figura 4.23 se puede observar las zonas determinadas como instrumentales, con la aplicación de la fase 3 utilizando la opción del análisis por ventanas.

Como se puede observar esta opción se puede considerar buena, sin embargo, la parte final de la canción sigue quedando un poco difusa, ya que nos presenta en un pequeño tramo de menos de 0.5 segundos de duración que existe zona instrumental, y este hecho no se puede considerar posible, ya que en ninguna señal instrumental nos encontraremos con zonas tan corta duración donde únicamente existe la zona instrumenta.

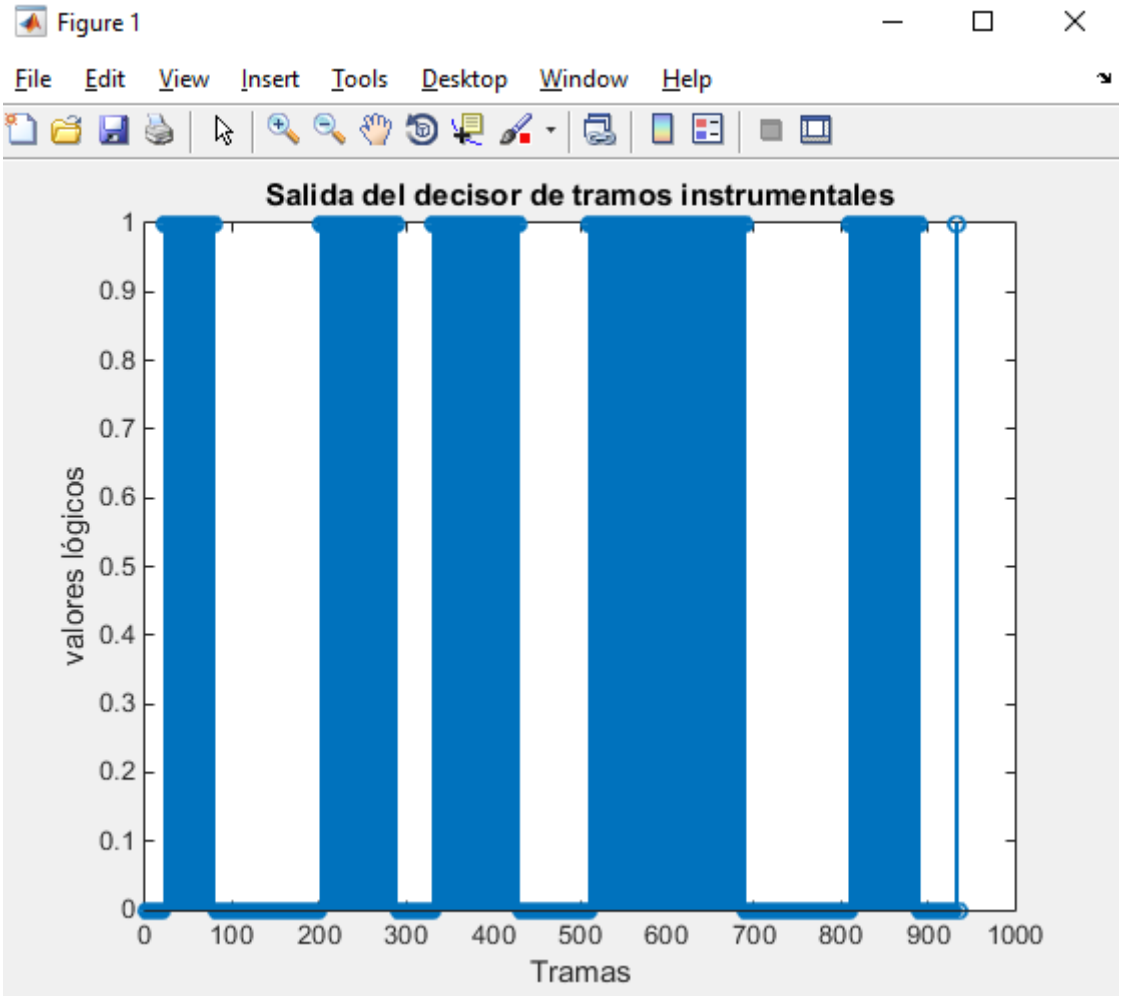

Figura 4.23: Representación de las zonas instrumentales mediante la opción de enventanado (fase 3).

Por lo tanto, debido a lo explicado anteriormente, es necesario buscar una nueva técnica para desarrollar esta fase del segmentador.

4.1.3.2 Clasificador HMM (Hidden Markov Model)

En este caso no utilizamos una técnica basada en análisis por ventanas si no que utilizamos el modelo oculto de Márkov.

Un modelo oculto de Márkov o HMM (Hidden Markov Model) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos. El objetivo es determinar los parámetros desconocidos (u ocultos) de dicha cadena a partir de los parámetros observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis.

En un modelo oculto de Márkov, el estado no es visible directamente, sino que sólo lo son las variables influidas por el estado. Cada estado tiene una distribución de probabilidad sobre los posibles símbolos de salida. Consecuentemente, la secuencia de símbolos generada por un HMM proporciona cierta información acerca de la secuencia de estados.

En la siguiente imagen podemos observar un ejemplo de transición de estados en un modelo oculto de Márkov.

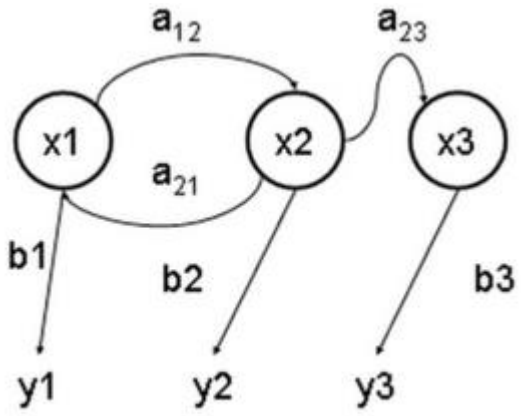

Figura 4.24: Ejemplo del modelo oculto de Márkov.

Partiendo de la imagen anterior podemos relacionar los datos de partida con los que vamos a trabajar para llevar a cabo esta fase del segmentador, con las variables mostradas en la misma.

- **x:** corresponden con los estados ocultos, que en nuestro caso tenemos dos. Es decir que corresponda a una zona instrumental o a una zona vocal.
- **y:** Corresponde a las salidas observables, que en nuestro caso corresponde con las zonas instrumentales y las zonas vocales obtenidas de la fase anterior, es decir, una

vez que se han obtenido la discriminación de las zonas podemos utilizarlas en esta fase.

- **a:** Corresponde con la probabilidad de transición de un estado a otro, que como veremos más a delante en un esquema serán muy restrictivas para asegurar una correcta determinación de las zonas.
- **b:** Corresponde con la probabilidad de salida, que en nuestro caso será la misma para los dos estados, ya que no partimos de información adicional para emplearla, únicamente de las zonas obtenidas en la fase anterior.

Es importante tener en cuenta que la señal de salida de la tercera fase es un conjunto de valores lógicos, indicando una primera aproximación de cuáles son las zonas vocales y cuales corresponden a las zonas instrumentales. Pero esta fase se hace necesaria para evitar ciertos tramos de las zonas detectadas como zonas instrumentales que sean de muy corta duración y que por lo tanto no representen ningún tipo de información. Por lo tanto, la entrada a esta fase va a ser una primera aproximación de las zonas instrumentales y vocales detectadas, y la salida de esta fase va a ser una clasificación de las zonas instrumentales y vocales de una forma más efectiva. En la Figura 4.25 podemos observar un esquema de las variables que forman parte del proceso HMM.

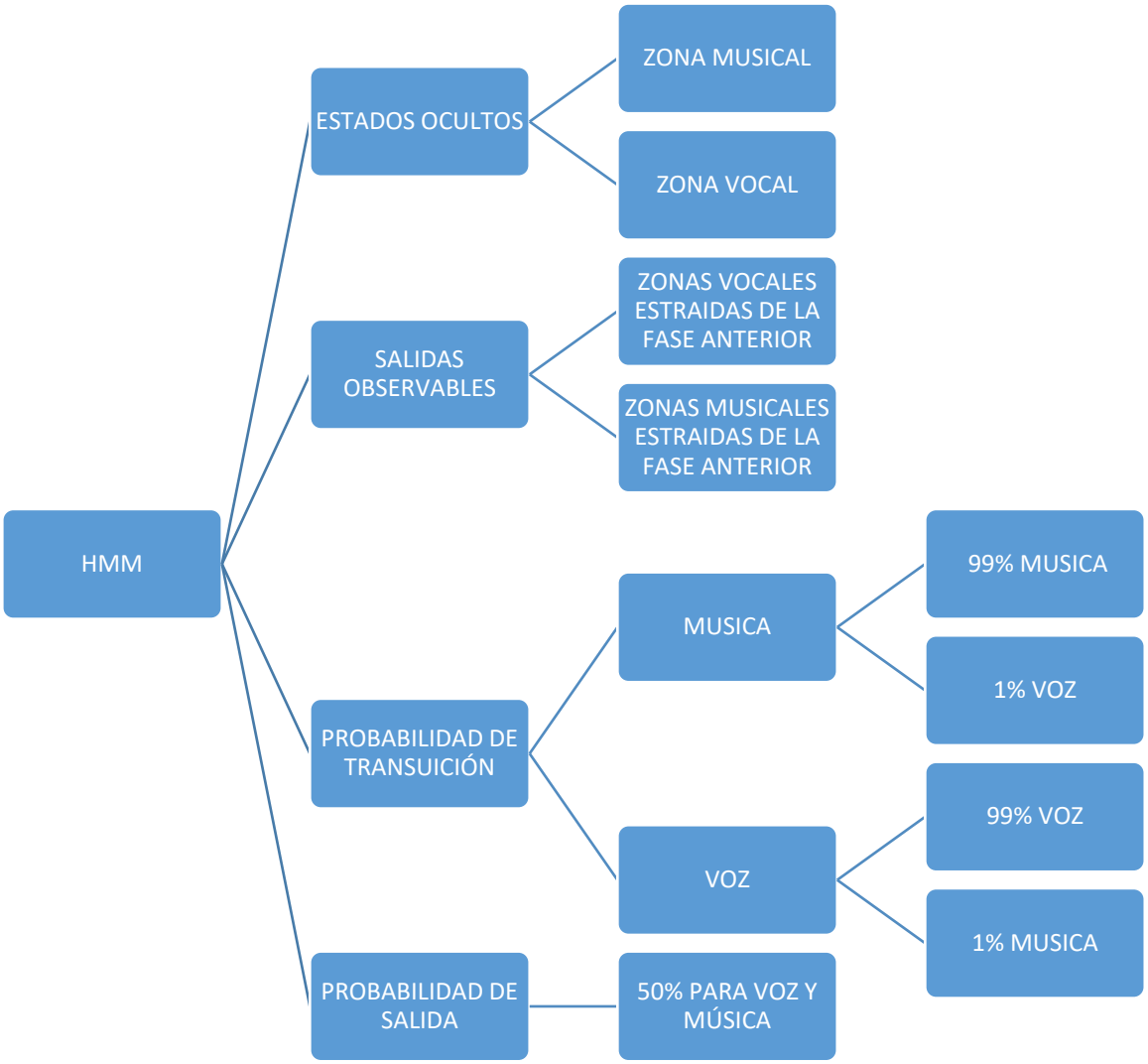

Figura 4.25: Esquema aclaratorio de la aplicación del HMM en la tercera fase del segmentador

Matlab dispone de una función que automáticamente realiza un modelo oculto de Markón aplicando el algoritmo de Viterbi. La función se explica a continuación:

```
function path = viterbi_path(prior, transmat, obslik)
```

Donde:

- **prior:** probabilidad de salida.
- **transmat:** probabilidad de transición.
- **obslik:** salidas observables.
- **path:** estados ocultos.

Aplicando esta función obtenemos las zonas donde existe parte musical y las zonas donde existen parte vocal de una forma más exacta. En la siguiente imagen podemos observar las zonas donde solo existe parte música en azul.

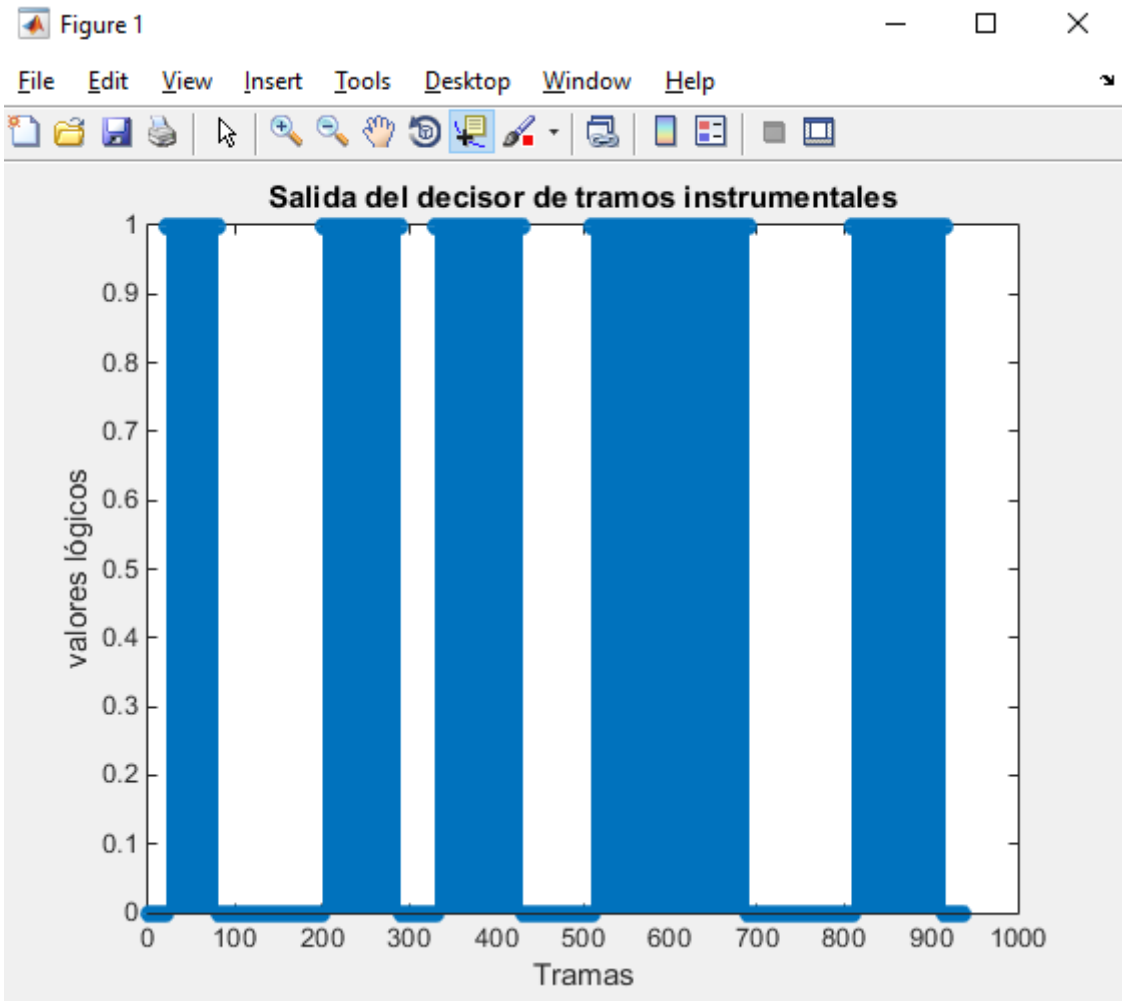

Figura 4.26: Representación de las zonas instrumentales mediante la opción HMM (fase 3).

En este caso, utilizando esta opción se puede visualizar que tanto las zonas instrumentales, como las zonas vocales presentan una mejor localización.

4.1.4 FASE 4: Aplicación de algoritmos detectores del ritmo

Una vez que se ha llevado a cabo la implementación del modelo oculto de Markóv descrito anteriormente para obtener una clasificación de las zonas vocales e instrumentales de forma más efectiva, se pensó en la utilización de los algoritmos encargados de detectar el ritmo de la señal musical para aplicar esos tiempos.

En un principio y haciendo una pequeña evaluación subjetiva, se determinó que las zonas vocales comienzan o finalizan en uno de los golpes del ritmo de la canción, por lo tanto, esto nos permitiría ajustar los resultados obtenidos anteriormente para mejorarlos a nivel objetivo y conseguir que nuestro segmentador funcionase mejor.

Conociendo esto, se pretende obtener lo que se conoce como Beat Spectrum gracias al cual se deduce el "Tiempo de Beat" que nos indica los tiempos en los que se produce un evento que marca el ritmo de la señal. Para llevar a cabo esta tarea se dispone a día de hoy de ciertos algoritmos ya implementados, en concreto en este documento nos centramos en el estudio de los algoritmos REPET [66] y ELLIS [67].

4.1.4.1 REPET: Repeating Pattern Extraction Technique

Repeating Pattern Extraction Technique es un algoritmo que viene descrito en el documento: A Simple Music/Voice Separation Method based on the extraction of the Underlying Repeating Structure [66]. En este se indica el procedimiento a seguir para la obtención de una determinada marca de tiempo, que supuestamente corresponde con el ritmo de la canción.

Las distintas fases que componen el algoritmo se describen a continuación de una forma visual e intuitiva.

- En primer lugar, se calcula un "correlograma" de la señal. Esto se hace calculando la autocorrelación de las filas del espectrograma en potencia de la señal.

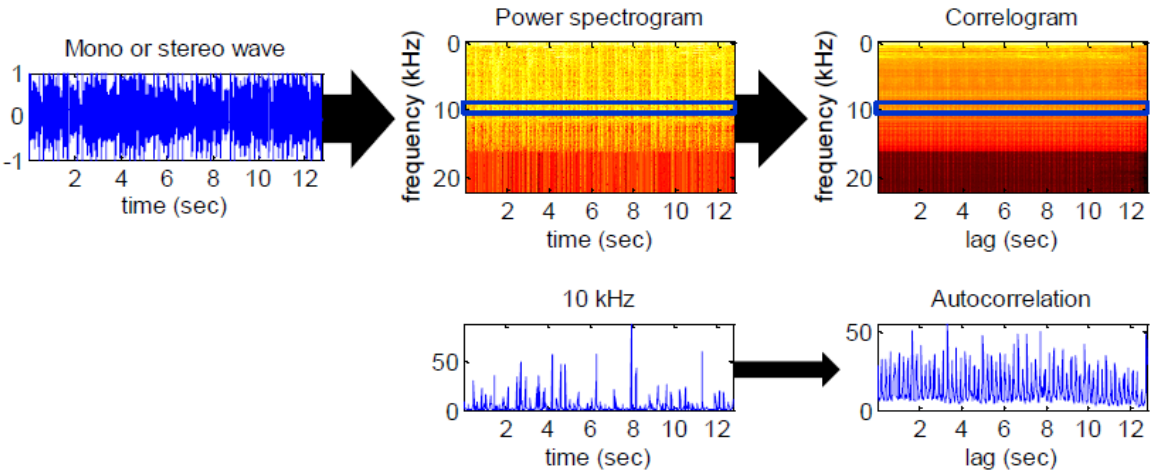

Figura 4.27: Imagen extraída del documento: A Simple Music/Voice Separation Method base don the extraction of the Underlying Repeating Structure [66].

- A continuación, se toma la media de las filas del "correlograma" obtenido, de este modo se obtiene lo que se conoce como Beat Spectrum.

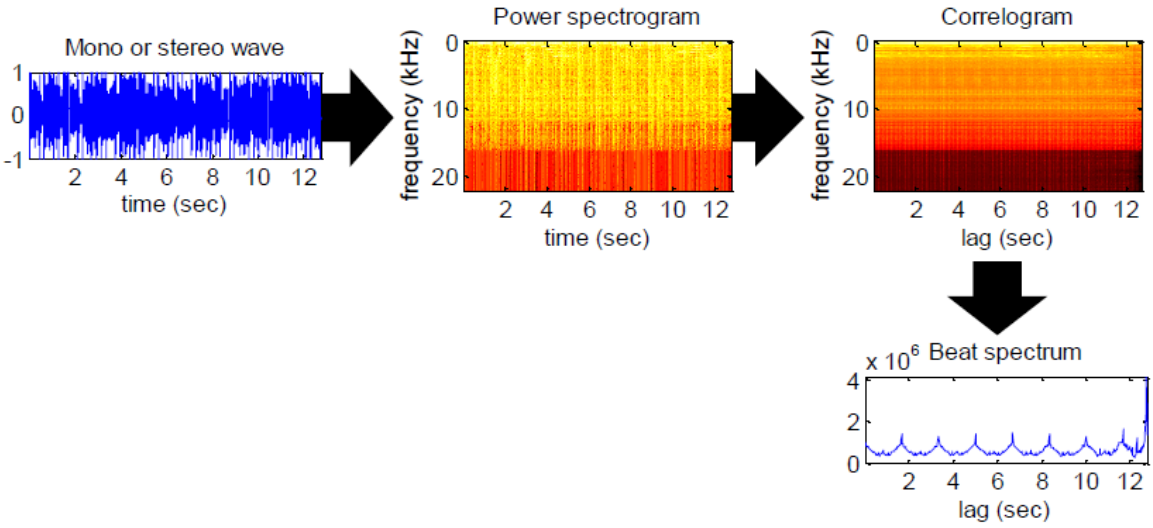

Figura 4.28: Imagen extraída del documento: A Simple Music/Voice Separation Method base don the extraction of the Underlying Repeating Structure [66].

- Este Beat Spectrum nos indica el periodo de repetición (P) de la estructura musical analizada.

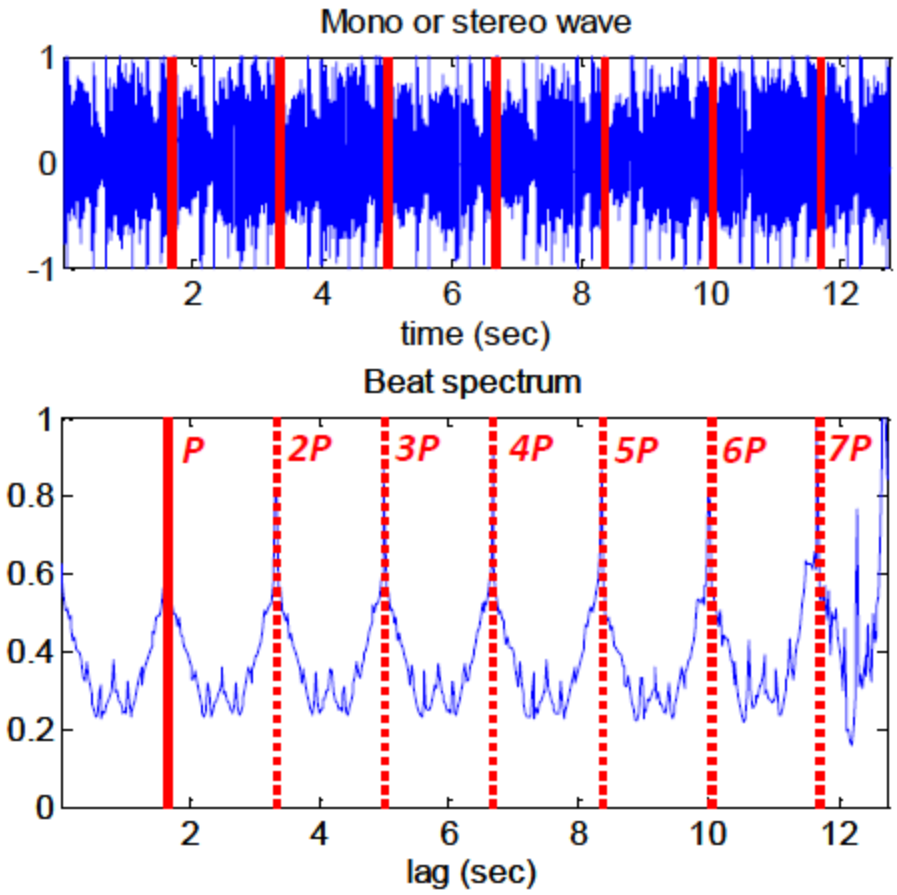

Figura 4.29: Imagen extraída del documento: A Simple Music/Voice Separation Method base don the extraction of the Underlying Repeating Structure [66].

4.1.4.2 ELLIS

El beat tracking o seguimiento de beat consiste en detectar automáticamente el momento en el que, en una señal de audio se produce un énfasis. Esta operación es la que realiza un músico con el pie, cuando sigue el ritmo de alguna canción.

Existen varios modelos para la detección del tempo, pero la gran mayoría siguen una estructura básica como la que se muestra en la siguiente figura, y al igual que utiliza el algoritmo del ELLIS.

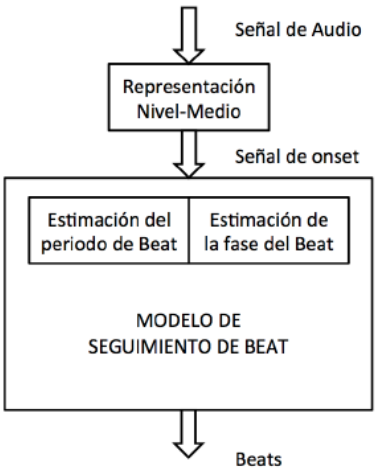

Figura 4.30: Estructura básica de un modelo de seguimiento de beat.

En el seguimiento de beat, se utiliza normalmente una función de onset como representación de nivel medio, esta función marca las zonas de transitorios en la pista de audio. Esta función se encarga de coger los máximos locales, y la señal resultante entrará al sistema de reconocimiento de beats. Este último es el encargado de calcular el tiempo entre beat y beat (periodo), además también se encargará de detectar en qué momento se produce cada beat (fase).

La estimación del tempo de Ellis [68,69], es una parte dentro de su enfoque a la programación dinámica para el seguimiento de beat. La función principal dentro de este algoritmo es una función de detección de onset. La función de detección se normaliza y posteriormente se analiza su periodicidad mediante el cálculo de la auto correlación. La señal resultante tendrá un valor de correlación alto para los golpes de beat.

Las fases que forman parte del detector de beat de ELLIS se puede ver en la Figura 4.31.

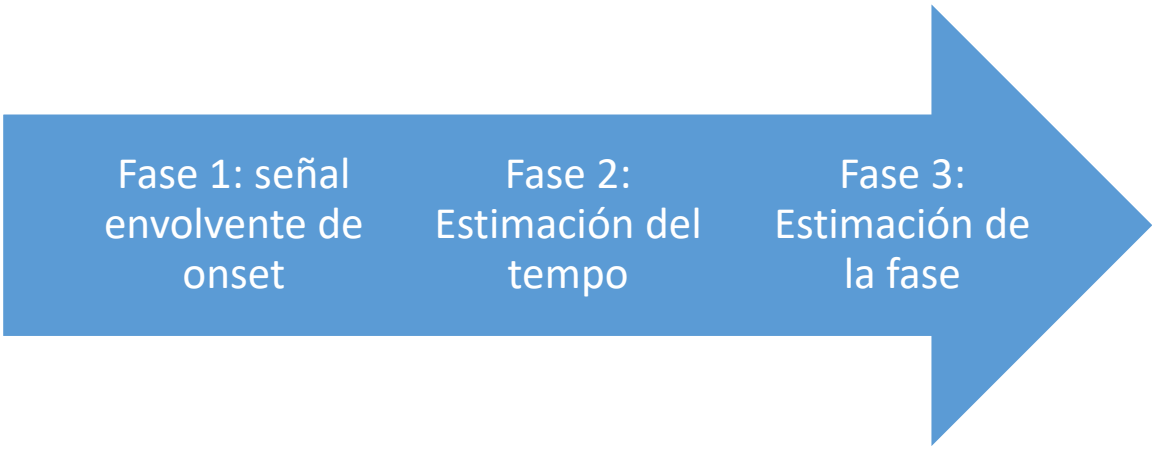

Figura 4.31: Etapas del detector de beat de Ellis.

- **Señal envolvente de onset:** De forma similar a la que se utiliza en el cálculo de otros modelos de onset, se calcula la envolvente de onset desde un modelo perceptual. El primer paso es muestrear de nuevo la señal de entrada a 8 KHz, para ello partimos de señales de audio muestreadas a 48 KHz que es la frecuencia a la que funcionan los dispositivos móviles. Para ello antes de realizar el downsampling vamos a filtrar la señal de entrada por un filtro paso bajo con frecuencia de corte inferior a $\frac{\pi}{6}$. Este paso es necesario para evitar el solapamiento del espectro de la alta frecuencia con el espectro de la baja:

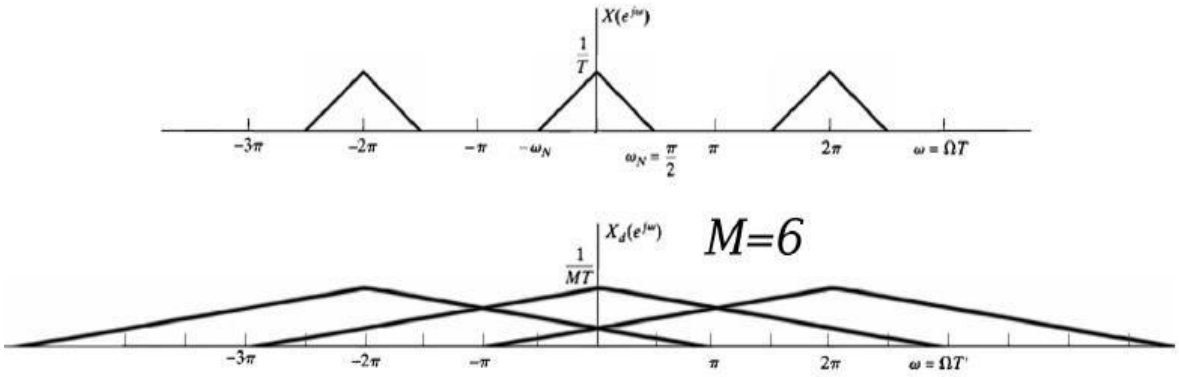

Figura 4.32: Diezmado de la señal de entrada sin filtrar

Una vez tenemos la señal muestreada a 8 KHz, se calcula la transformada corta de Fourier (STFT) usando 32 ms de ventana y 4 ms de salto entre tramas. 32ms a 8KHz son 256 muestras, lo que significa que cada FFT será de unos 256 puntos. Posteriormente la STFT es multiplicada por los coeficientes de Mel. Estos coeficientes forman un banco de filtros de 40 bandas que representan la percepción auditiva.

En la Figura 4.33 se observa una representación de la STFT de los primeros segundos de una canción, como se puede observar se aprecia claramente las partes donde hay un cambio frecuencial importante.

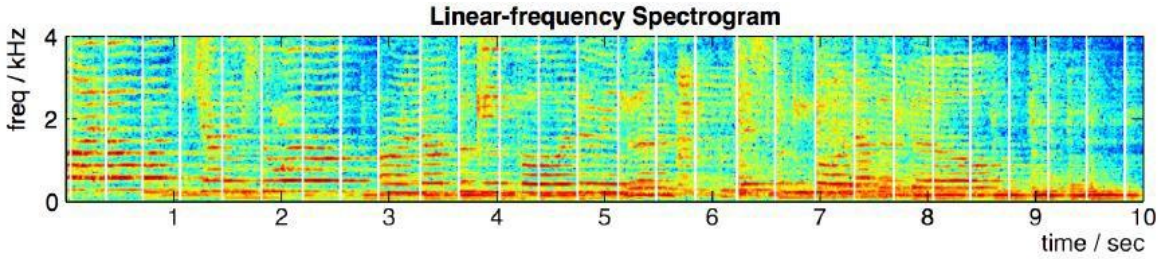

Figura 4.33: STFT de la señal de entrada

Posteriormente, se le aplican los coeficientes de Mel dando como resultado el mostrado en la Figura 4.34:

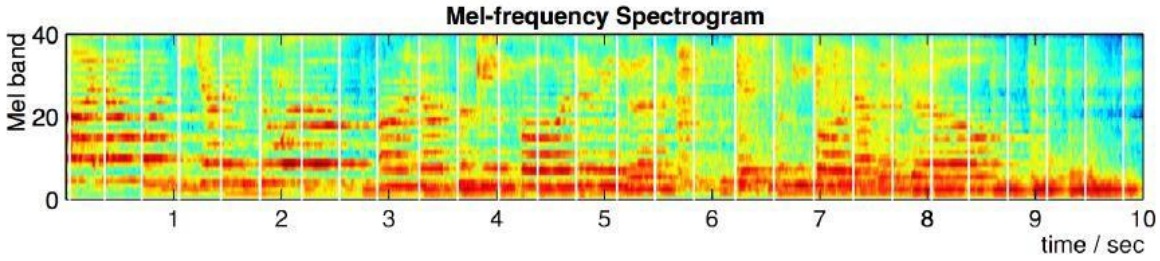

Figura 4.34: STFT aplicando los coeficientes de Mel

Al aplicar los coeficientes de Mel se baja la resolución de 129 a 40 coeficientes, era algo evidente ya que el banco de filtros es de 40 bandas. El espectrograma de Mel se convierte a dB, en las gráficas ya ha sido representado en dB (rojo 20dB y azul -40dB), y se calcula la diferencia de potencia entre bandas $(banda_2 - banda_1, banda_3 - banda_2, \dots, banda_{N-1} - banda_{N-2})$ a lo largo del tiempo

- **Estimación del tempo:** En este apartado se explicará cómo estimar, τp el periodo global que sigue la señal de entrada. Como ya tenemos nuestra señal de onset $O(t)$, aplicando la autocorrelación podemos ver la periodicidad de la señal, ya que, al tomar versiones de la misma señal retardada, cuando estas dos señales coinciden en un tiempo de beat, la correlación es muy alta. Esto se puede observar gráficamente en la siguiente figura, donde se muestra la señal de onset filtrada y a la autocorrelación de la misma.

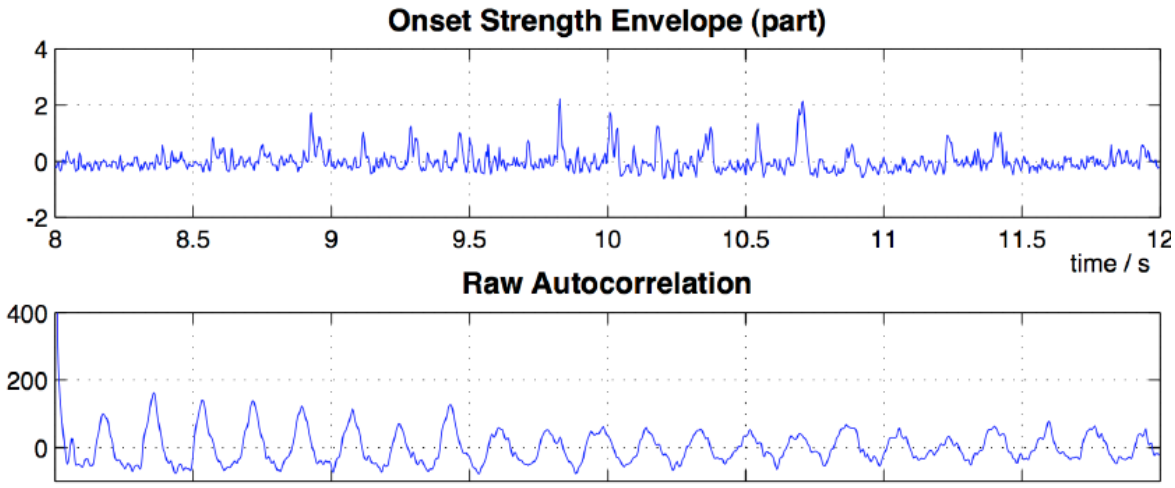

Figura 4.35: Señal de onset y autocorrelación sin enventanar

- **Estimación de la fase:** El objetivo final de un seguidor de beat, es conseguir el periodo entre golpe y golpe, y colocar el primer golpe en el lugar donde debe ir. Para ello en este modelo se utiliza una función objetivo, la clave de esta función es obtener una puntuación sobre cada posible beat para posteriormente conocer cuál es el mejor candidato, dentro de la señal onset, a ser el beat en un instante determinado.

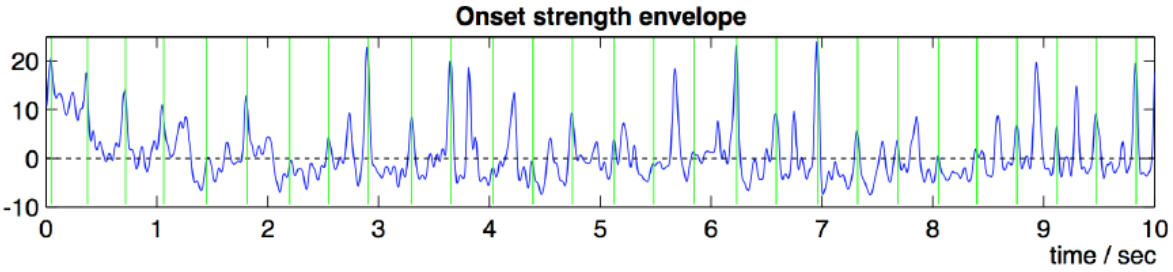

Figura 4.36: Señal de onset (color azul) con los beats detectados (color verde)

4.1.4.3 Conclusión a la hora de utilizar los algoritmos detectores de beat (ritmo):

En un primer análisis y como se ha comentado anteriormente se pudo comprobar que, en la mayoría de las canciones, cuando el cantante canta y por lo tanto cuando la fuente vocal está presente, el inicio coincide con uno de los golpes del ritmo de la canción. Es por ello que se pensó en añadirle esta información adicional al sistema utilizando un algoritmo de detección de beat (ritmo) para mejorar los resultados.

Una vez estudiados los algoritmos de detección de beat, descritos anteriormente se han llevado a las siguientes conclusiones:

- Así, el algoritmo REPET devuelve una única marca de tiempo (P) y asume que la señal presenta un patrón de repetitividad constante a lo largo del tiempo. Esto quiere decir que para REPET los tiempos de beat que marcan el ritmo se producen en instantes de tiempo múltiplos del valor P obtenido. Esta suposición no es del todo cierta dado que, aunque el ritmo suele repetirse con un periodo casi constante, el ser humano es incapaz de generar un ritmo con un patrón de repetitividad tan exacto.

- El algoritmo facilitado por ELLIS no devuelve una única marca de tiempo como ocurría en el caso de REPET sino que nos indica todos los tiempos en los que se producen eventos que marquen el ritmo de la señal, es decir, devuelve los tiempos de beat de toda la señal. Como indicamos antes, el periodo de repetición del ritmo no es constante, sino que presenta pequeñas fluctuaciones a lo largo del tiempo, esto es tenido en cuenta por ELLIS por lo que los tiempos de beat devueltos por este algoritmo no son múltiplos de un valor determinado, sino que varían (obteniendo valores mayores o menores) según indica el ritmo.

- Aunque conseguimos un algoritmo que detecta los tiempos de beat de una forma eficiente, a la hora de analizar en qué tiempos se producían esas detecciones del ritmo, nos dimos cuenta que verdaderamente las zonas vocales no tienen por qué iniciarse en alguno de los tiempos de beat detectados por el algoritmo de ELLIS. Tras un análisis exhaustivo con una base de datos formada por distintos tipos de señales musicales, pudimos observar que los tiempos devueltos por ELLIS no cumplían la relación de que alguno de ellos coincidiera con los momentos en los que la fuente vocal comienza a aparecer.

- Por lo tanto, la primera suposición de que siempre que aparece la fuente vocal coincide con uno de los golpes del ritmo quedo desmentida. Analizando un poco en profundidad esta conclusión, nos damos cuenta que haciendo un análisis subjetivo (escuchando

diferentes canciones) podemos afirmar que cuando el cantante canta aparece un golpe de batería u otro instrumento percusivo (que son los que determinan el ritmo). Pero es necesario tener en cuenta que este análisis es subjetivo y depende de la persona que lo realiza, ya que al fin y al cabo tu oído en ese momento especifico puede determinar que esa relación se cumple siempre. Pero, por otro lado, tras realizar un análisis objetivo, obteniendo los resultados de los tiempos de beat devueltos por el ELLIS y los comparamos con los tiempos en los cuales el cantante comienza a cantar, existen diferencias significativas que a la hora de obtener resultados nos perjudica la etapa de segmentación en su conjunto.

Una vez que se han visto las distintas conclusiones a nivel especifico se ha determinado que la fase 4 del segmentador correspondiente a la aplicación de algoritmos detectores del ritmo no formara parte del sistema. Por lo tanto, se ha demostrado que su utilización carece de sentido a la hora de realizar un sistema de segmentación.

4.1.5 FASE 5: Sistema de corrección y anotación

La última fase del segmentador automático corresponde a la extracción de los tiempos de cada una de las zonas instrumentales y a la correspondiente creación de un fichero de texto donde se muestren. Por lo tanto, en esta sección de la memoria presentaremos la función que permite realizar dicho proceso y unas consideraciones a tener en cuenta para perfeccionar el modelo del segmentador empleado.

Inicialmente el sistema utilizado para la extracción de los tiempos de inicio y fin de las tramas instrumentales se describe de la siguiente forma:

En primer lugar, mediante un bucle se recorre la señal obtenida del decisor de tramos instrumentales para realizar una señalización de los inicios y finales de los intervalos instrumentales, para ello a las tramas de inicio y de fin se le asigna el valor 2 como se muestra en la Figura 4.37.

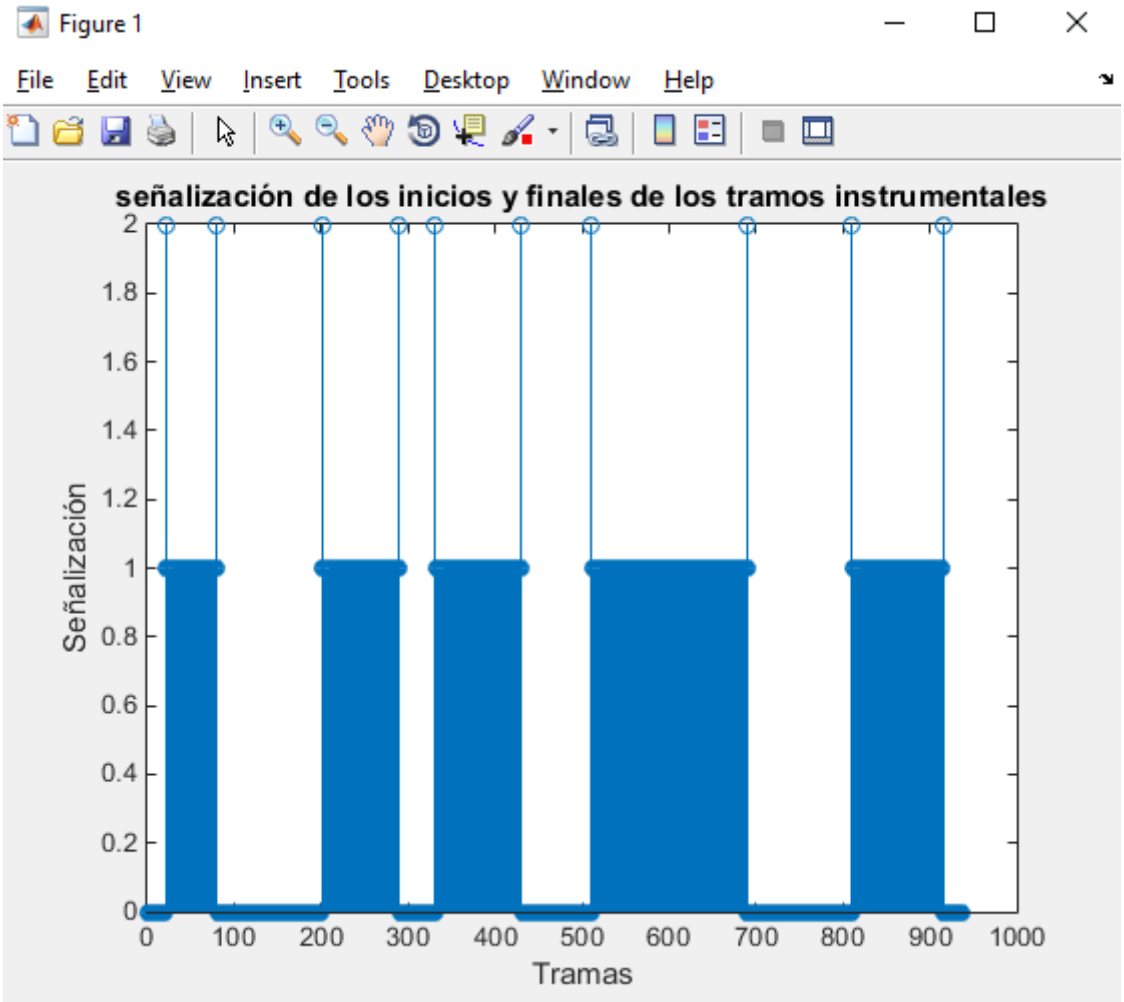

Figura 4.37: Señalización de los inicios y finales de las zonas instrumentales

Una vez se haya realizado la señalización de los inicios y finales de los tramos instrumentales se deben de tener en consideración los siguientes puntos:

- Las tramas correspondientes a los inicios y a los finales se almacenarán en un vector por orden de aparición, de tal forma que si extraemos los valores impares correspondan a los inicios de los tramos instrumentales y si extraemos los valores pares correspondan con los finales de dichos tramos.

Los vectores que se ha obtenido hasta el momento son los siguientes:

- **d:** vector que almacena las tramas que señalan los inicios y finales de las tramas instrumentales (incluida la que pertenece al segundo 0).
- **Inicios:** vector que contiene las tramas que señalan los inicios de los tramos instrumentales, para ello se extraen los valores impares del vector **d.**
- **Finales:** vector que contiene las tramas que señalan los finales de los tramos instrumentales, para ello se extraen los valores pares del vector **d.**

El siguiente paso del proceso sería realizar un cambio del dominio para el eje de abscisas correspondiente a las tramas, para ello se aplica ingeniería inversa.

Conociendo el tamaño de la señal de entrada (muestras) y el número de tramas se puede definir una constante conocida como salto, que indica el número de muestra que compone cada una de las tramas. De tal forma que realizando una multiplicación de las tramas que señalan los inicios y finales de los tramos instrumentales por la constante definida como salto y realizando una división entre la frecuencia de muestreo se podrán obtener los tiempos en segundos de las zonas donde solo hay bases instrumentales.

En las ecuaciones siguientes se puede observar el proceso de ingeniería inversa aplicado para realizar el cambio de tramas a tiempo en segundos.

$$salto = \frac{nº\ de\ muestras\ de\ la\ señal\ de\ entrada\ tras\ el\ muestreo}{nº\ de\ tramas\ de\ la\ señal\ obtenida\ del\ decisor} \qquad (89)$$

$$tiempo\ en\ segundos = (tramas\ de\ inicio\ y\ de\ fin - 1) \times \frac{salto}{fs} \qquad (90)$$

De esta forma únicamente quedaría crear un vector formado por dos columnas, de tal forma que en la primera se coloquen los tiempos de inicio en segundos y en la segunda se coloquen los tiempos de final en segundos.

Los nuevos vectores que se han obtenido son los siguientes:

- **tiempoI:** vector que contiene los tiempos en segundos de los inicios de las zonas instrumentales (incluido el segundo 0).
- **tiempoF:** vector que contiene los tiempos en segundos de los finales de las zonas instrumentales.
- **tiempo:** vector formado por dos columnas, de tal forma que la primera columna contiene los tiempos de "**tiempoI**" y la segunda los tiempos de "**tiempoF**".

Ha sido necesario tomar una serie de consideraciones a la hora de realizar este proceso las cuales son:

- Para evitar que las zonas detectadas como bases instrumentales contengan parte de las bases vocales, se aplica un margen de seguridad definido como M tanto para los tiempos de inicio como para los tiempos de fin. Este margen ha sido seleccionado de tal forma que los resultados fuesen óptimos, M=0.25 segundos.
- Basándonos en las características musicales, carece de sentido que aparezcan tramos musicales cuya duración sea inferior a 0.7 segundos, es por esto que si tras realizar la anotación aparecen tramos inferiores se ignoran.

La función implementada en MATLAB que permite realizar este proceso es:

```
function [decision] = anotar(decision, M, ajuste)
```

Las variables de entrada de la función son:

- **decisión:** corresponde con la señal obtenida de la fase anterior (decisor de tramos instrumentales).
- **M:** se refiere al margen de seguridad que se añade tanto a los inicios como finales de las zonas instrumentales (en segundos).
- **ajuste:** corresponde con el valor mínimo que pueden presentar las zonas instrumentales para ser tenidas en cuenta.

Por último, faltaría realizar una exportación de los datos temporales almacenados en el vector “decisión” a un fichero de texto cuyo nombre coincida con el de la canción que se está analizando. Esta fase se realiza de forma bastante sencilla mediante la función fprintf que incorpora MATLAB.

Es necesario la creación de un fichero de texto porque el sistema de separación NMF que permite obtener las bases instrumentales (armónico/percusiva) utiliza estos tiempos para determinar las zonas instrumentales y poder obtener las bases instrumentales presentes en estas zonas y que se utilizaran en el último modelo de separación NMF que realiza una separación óptima de las distintas fuentes sonoras (vocal/armónico/percusiva).

En la Figura 4.38 se muestra el diagrama de flujo correspondiente a la estructura del sistema e anotación implementado en esta sección.

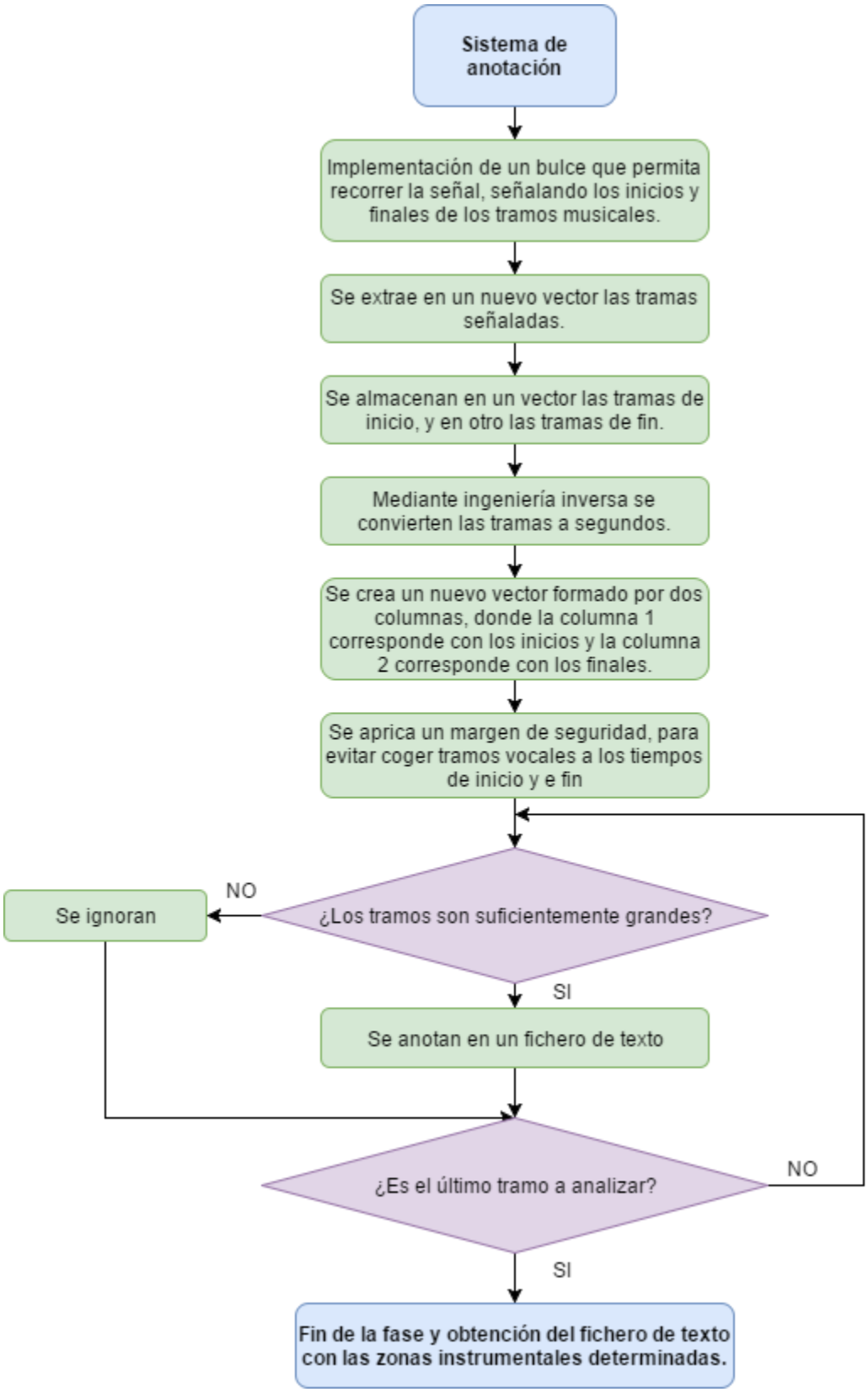

Figura 4.38: Diagrama de flujo del sistema de anotación implementado

En la Figura 4.39 se representa el fichero de texto obtenido tras la realización del sistema completo de segmentación automático para el ejemplo utilizado a lo largo de todo el TFM.

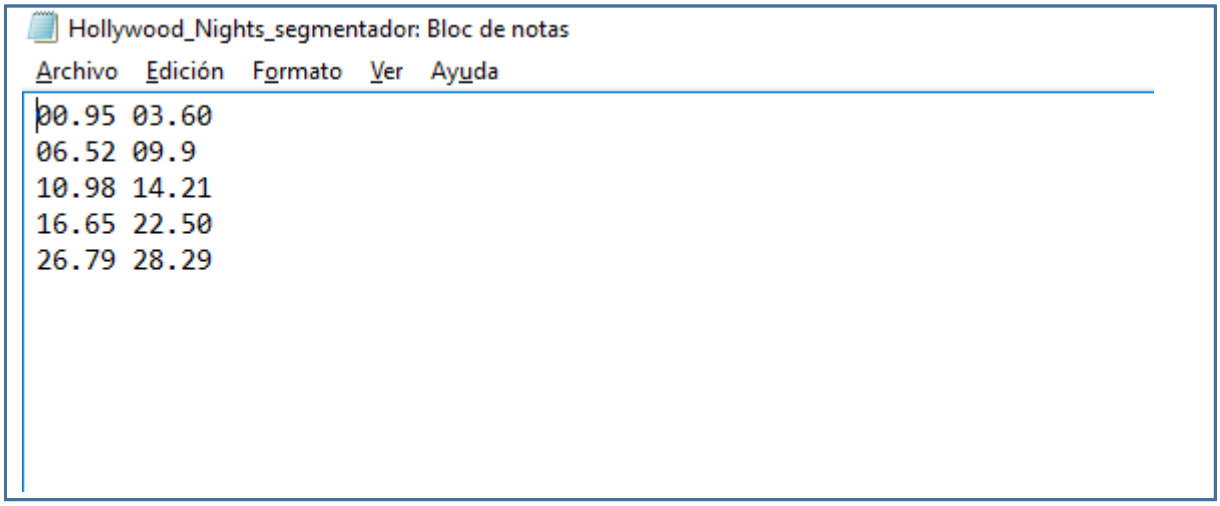

Figura 4.39: Fichero de texto con las zonas instrumentales obtenido del segmentador automático

Si comparamos los resultados obtenidos por nuestro segmentado automático, con los obtenidos de forma manual (anotando a mano, ver Figura 4.40), podemos ver que los resultados de las zonas instrumentales detectadas se aproximan bastante, por lo tanto, el segmentador va a ofrecer resultados muy buenos.

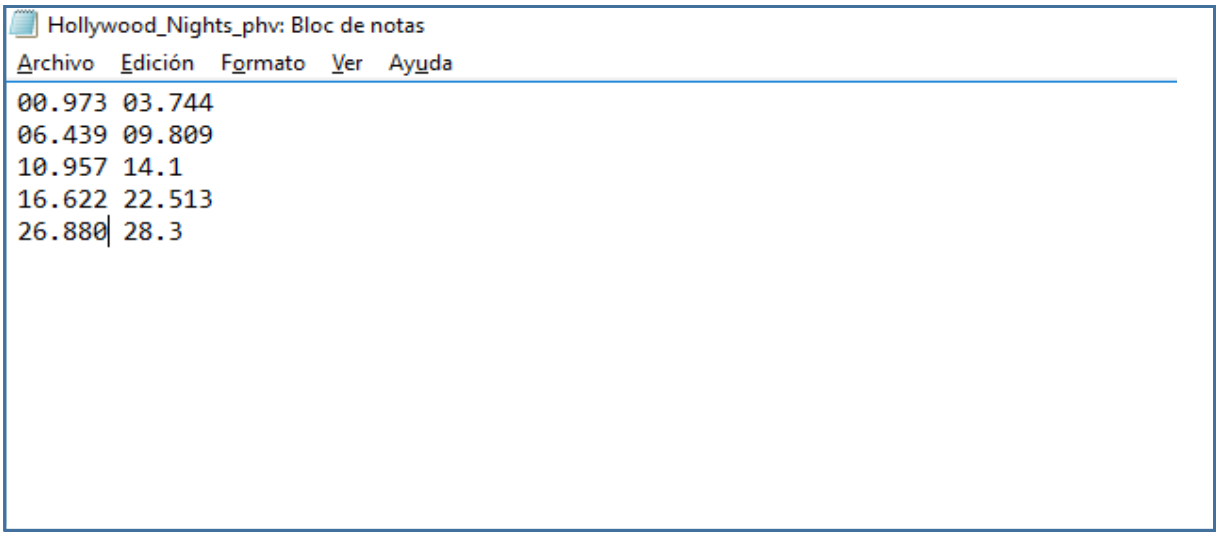

Figura 4.40: Fichero de texto con las zonas instrumentales detectadas de forma manual.

4.1.6 FASE 6: Sistema general y conclusión

Una vez que se ha explicado cada una de las fases que componen el sistema de segmentación y especialmente para que no haya ningún tipo de confusión a la hora de entender el sistema final que se ha determinado como óptimo, se va a llevar a cabo una explicación de las distintas técnicas que se han tomado al final para la realización de cada una de las fases, para ello nos vamos a basar de una imagen donde se muestra todo el proceso desde su inicio hasta su fin.

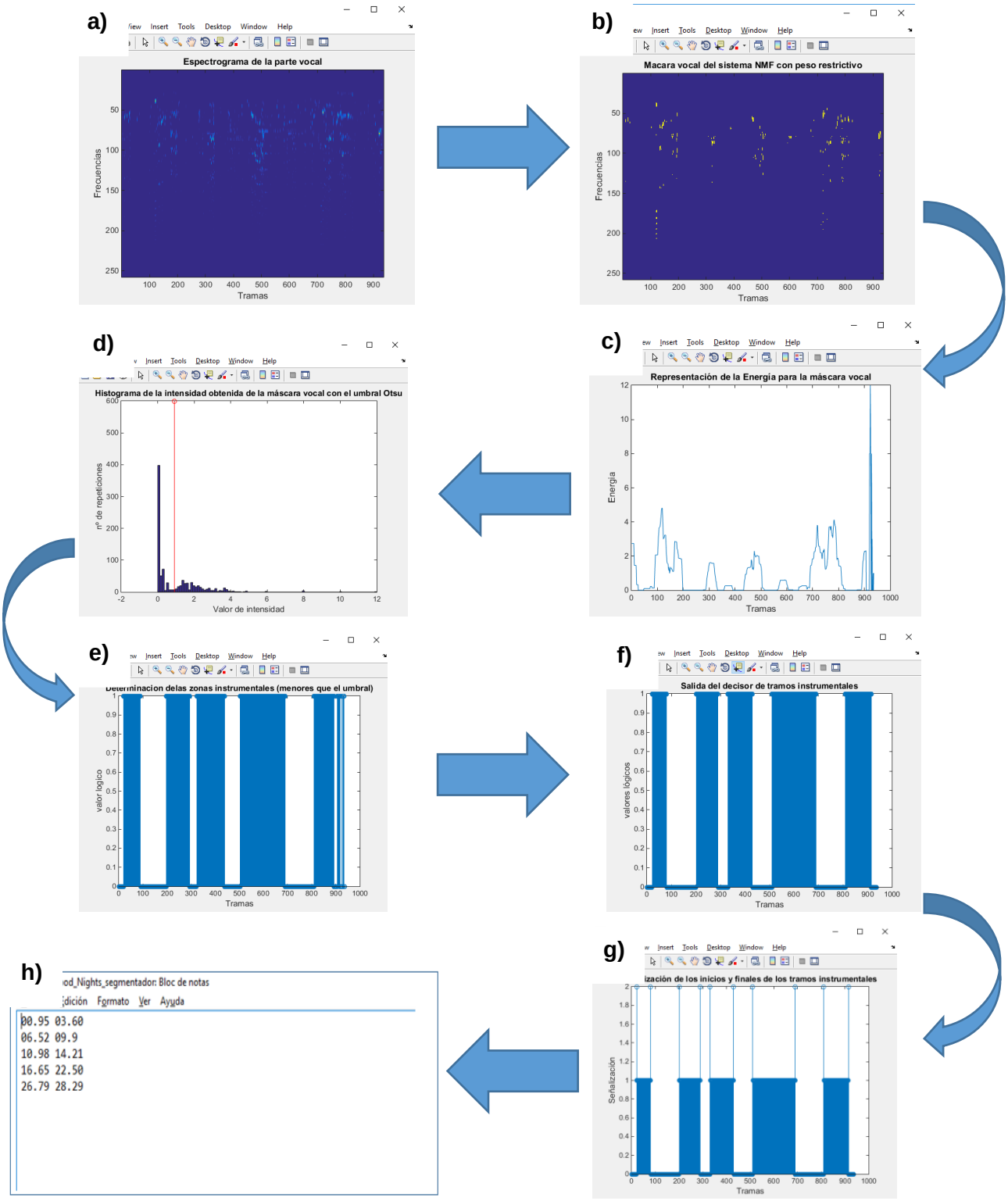

Figura 4.41: Esquema representativo del segmentador automático del TFM. a) Espectrograma de la parte vocal (fase 1). b) Máscara de la parte vocal con peso restrictivo (fase 2). c) Intensidad del espectrograma vocal (fase 2). d) Histograma de la intensidad con umbral de Otsu para distinguir zonas vocales e instrumentales (fase 2). e) primera determinación de las zonas instrumentales (fase 2). f) Aplicación del HMM para mejorar la determinación de las zonas instrumentales (fase 3). g) Indicación de los inicios y finales de las zonas instrumentales (fase 5). h) Anotación de las zonas instrumentales en un fichero de texto para su posterior utilización en el separador H/P/V.

Hay que destacar que existe una relación directa entre cada una de las fases. Pero a partir de la fase 2, las siguientes fases buscan mejorar los resultados de segmentación. En la siguiente tabla se puede observar la cantidad (en %) de las zonas instrumentales obtenidas por el segmentador de forma correcta con forme se van añadiendo las fases al segmentador. Este análisis objetivo se ha llevado a cabo utilizando la base de datos ISMIR 2015. En la sección 5.1.2 “Segmentador automático TFM vs Segmentador manual”, se podrá ver un análisis detallado del segmentador automático para cada una de las canciones que forman la base de datos y las medidas que se utilizan para ello.

FASE	% DE ZONAS INSTRUMENTALES DETECTADAS CORRECTAMENTE
1: Preprocesado NMF	Por si sola no obtiene zonas instrumentales, únicamente la información visual.
2: Discriminación de música y voz	85%
3: Detector de tramos instrumentales	89%
4: Aplicación de algoritmos detectores de beat	No se implementa para el sistema óptimo
5: Sistema de corrección y anotación.	90%

Tabla 4.1: Mejora de los resultados a lo largo de las fases.

Es importante considerar además de las técnicas utilizadas para la implementación final, las que se han analizado y se han demostrado que no conducen a buenos resultados, ya que para futuros proyectos de investigación esto permitirá a otros investigadores no utilizar técnicas que no resulten satisfactorias para el segmentador.

En el caso de haber utilizado la fase 4 de los algoritmos detectores de beat la tabla anterior cambiaría de la siguiente forma.

FASE	% DE ZONAS INSTRUMENTALES DETECTADAS CORRECTAMENTE
1: Preprocesado NMF	Por si sola no obtiene zonas instrumentales, únicamente la información visual.
2: Discriminación de música y voz	85%
3: Detector de tramos instrumentales	89%
4: Aplicación de algoritmos detectores de beat	75%
5: Sistema de corrección y anotación.	76%

Tabla 4.2: Evolución de los resultados si introducimos la fase 4 de los detectores de beat

Claramente se puede observar que los resultados empeoran considerablemente en la fase 4, por lo que estaríamos perjudicando los buenos resultados de las anteriores fases del segmentador, y es por ello la decisión de no utilizarse.

4.2 IMPLEMENTACIÓN DE LA ETAPA DEL SEPARADOR BASADO EN NMF

En esta sección se va a llevar a cabo una explicación detallada de la realización de la fase de separación, basada en los algoritmos NMF.

Para asegurar su entendimiento y además comprobar que el sistema funciona de forma óptima, en primer lugar, se va a llevar a cabo una explicación del sistema de separación NMF únicamente para dos fuentes sonoras, la armónica y la percusiva. En esta parte se van a llevar a cabo dos opciones para analizarlas y llegar a la solución más eficiente. Por lo tanto, se va implementar un sistema de separación NMF para fuentes armónico-percusivas con las siguientes opciones:

a) **Separador NMF para fuentes armónico-percusivas sin inicializar las bases armónicas con bases entrenadas.**
b) **Separador NMF para fuentes armónico-percusivas inicializando las bases armónicas con bases entrenadas correspondientes al piano.**

Es necesario tener en cuenta que ambas opciones que se han mencionado anteriormente se llevan a cabo para conseguir que el sistema de separación NMF funcione de forma eficaz. Esta fase es independiente del resto de los modelos de separación que se van a utilizar en el Trabajo Fin de Master y su único fin es comprobar que el sistema de separación por si solo funciona adecuadamente. Esta fase es necesaria porque de esta forma comprobamos que el sistema de separación funciona bien, por lo tanto, el sistema de separación final de nuestro sistema final tiene más posibilidades de funcionar de forma eficiente.

En la siguiente figura se muestra el esquema que se va a seguir para desarrollar el sistema de separación armónico/percusivo que nos permitirá comprobar el funcionamiento del sistema NMF de separación.

Figura 4.42: Esquema del sistema NMF armónico/percusivo que permite comprobar el funcionamiento del sistema de separación NMF

Una vez que se ha comprobado que el sistema de separación funciona de forma eficiente para las dos fuentes utilizadas (armónico-percusiva) se procederá a desarrollar el sistema que se va a utilizar para llevar a cabo la separación del sistema definido en el Trabajo Fin de Master.

Para el desarrollo del sistema final y teniendo en cuenta además que el segmentador se basa en un modelo NMF se han utiliza en total tres modelos de separación basados en los algoritmos NMF los cuales y como anteriormente se han mencionado son:

- **El primer modelo NMF (segmentador automático)** utilizado realiza una separación vocal/armónico/percusiva mediante la utilización de bases armónicas entrenadas, ya que se desconocen las tramas temporales correspondientes a las zonas instrumentales. Este algoritmo es el que nos permite realizar una segmentación automática de forma óptima como se ha mencionado en el punto 4.1.1.3.
- **El segundo modelo NMF (separador)** utilizado realiza una separación armónica/percusiva (sin parte de voz) en las zonas instrumentales detectadas por el segmentador automático. Este algoritmo es el que permite aprender las bases instrumentales para su posterior utilización.
- **El tercer modelo NMF (separador)** utilizado realiza una separación vocal/armónico/percusiva utilizando las bases instrumentales aprendidas en el segundo modelo NMF. Este algoritmo es el que permite realizar una separación optima de las fuentes sonoras que compone la señal musical, partiendo de las bases instrumentales entrenadas y obteniendo unos resultados satisfactorios.

Para el desarrollo de la **etapa de separación** se han usado los dos últimos modelos NMF comentados anteriormente. El algoritmo NMF se basa en la factorización de una matriz de entrada, es decir, la matriz de entrada es considerada el producto de un conjunto de matrices.

En los modelos desarrollados, se ha tomado como referencia un espectrograma $\boldsymbol{X}$ de entrada, con dimensiones frecuencia-tiempo, para la factorización. Este espectrograma es descompuesto en dos matrices distintas; por un lado, se obtiene una matriz $\boldsymbol{W}$ que se corresponde con la excitación de una fuente sonora en frecuencia; por otro lado, se obtiene una matriz $\boldsymbol{H}$, que corresponde con la activación de esa excitación en tiempo. Para simplificar y facilitar el uso de las distintas ecuaciones se va a trabajar con matrices normalizadas. Por lo tanto, es necesario asociar a cada conjunto de matrices (**H y W**) con

su correspondiente norma ‘e’, la cual se debe diagonalizar (**diag (e)**). Debe existir una diagonal por cada par de matrices, es decir, por cada una de las fuentes sonoras que se quieren separar.

Por lo tanto, es preciso comentar que todas las matrices correspondientes a las bases (W) y a las activaciones (H) están normalizadas. La norma de estas matrices se guarda en un vector definido como “e” y su multiplicación con la matriz se realizada diagonalizando el vector.

Dentro de cada iteración del modelo NMF, se actualizan las matrices H y W mediante las correspondientes ecuaciones de actualización, y después se vuelven a normalizar pasando los valores de norma extraídos al vector “e”.

Todo esto se hace para hacer más fáciles las ecuaciones de actualización ya que las regularizaciones se hacen con respecto a matrices normalizadas. Para terminar las regularizaciones o restricciones se hacen con respecto a matrices normalizadas para asegurar que su proporción en la regularización con respecto a la beta-divergencia sea siempre constante o se puede controlar de antemano.

Para el primer modelo correspondiente al separador, se ha tomado que el espectrograma de entrada está formado por dos espectrogramas diferentes, un espectrograma correspondiente a la parte percusiva y otro correspondiente a la parte armónica. En este modelo solo se realiza separación entre fuentes instrumentales, por lo que se aplica únicamente a aquellas partes de la canción donde sólo se encuentran fuentes de este tipo, para ello se usa la información extraída de los ficheros de texto generados por el segmentador automático. Por lo tanto, esta es la motivación de utilizar un segmentador, ya que, si no dispusiésemos de la información temporal donde esta presentes las zonas instrumentales, el sistema carecería de información adicional para realizar la extracción de las bases armónico-percusivas y los resultados de separación no serían tan eficaces.

Estas dos matrices contienen las bases armónicas y percusivas, donde se encuentra la información que indica si en un determinado punto de la canción aparecen sonidos armónicos, sonidos percusivos o ambos.

En este modelo, ambas bases son actualizadas mediante las llamadas reglas de actualización multiplicativas, es decir, a través de un número determinado de iteraciones, la información de estas bases va actualizándose. Para el cálculo de esta información se

minimiza una función de coste global, llamada D, que se compone de tres partes. La primera parte, es el llamado coste β-divergencia, que hace referencia a la similitud de la señal original con la señal tras cada paso por el algoritmo. La segunda parte, es el costo percusivo, que toma en cuenta la cantidad de sonido percusivo que habrá en un punto de la canción. La última parte, es el costo armónico, que cuenta con la misma función que el costo percusivo, pero referido a los sonidos armónicos.

Una vez estas bases son calculadas, son usadas en el segundo modelo para realizar la separación entre sonidos armónico/percusivos y vocales. Se calculan nuevas bases, esta vez en las zonas instrumentales/vocales. Para ello, en este modelo se introducen un nuevo espectrograma referente a la información vocal. Se usa un modelo excitación/filtro para estimar la parte vocal de estas zonas, es decir, a cada excitación vocal corresponde un filtro determinado que permite su extracción, para finalmente obtener el trayecto de la voz principal de la canción. En este segundo modelo, se actualizan las bases percusivas y armónicas, así como las bases correspondientes a las activaciones de filtros y excitaciones vocales usando las reglas de actualización multiplicativas ya comentadas. Esto se lleva a cabo únicamente en las zonas vocales, ya que en las zonas instrumentales se han extraído las bases armónica y percusivas.

A continuación, se va a realizar una explicación detallada de las distintas fases que se deben de llevar a cabo para conseguir un sistema de separación NMF que funcione de forma correcta y al fin al cabo obtener un sistema Segmentador NMF/Separador NMF que consiga extraer las distintas fuentes sonoras presentes en la señal musical. Un esquema detallado de los distintos pasos se puede ver en la Figura 4.43.

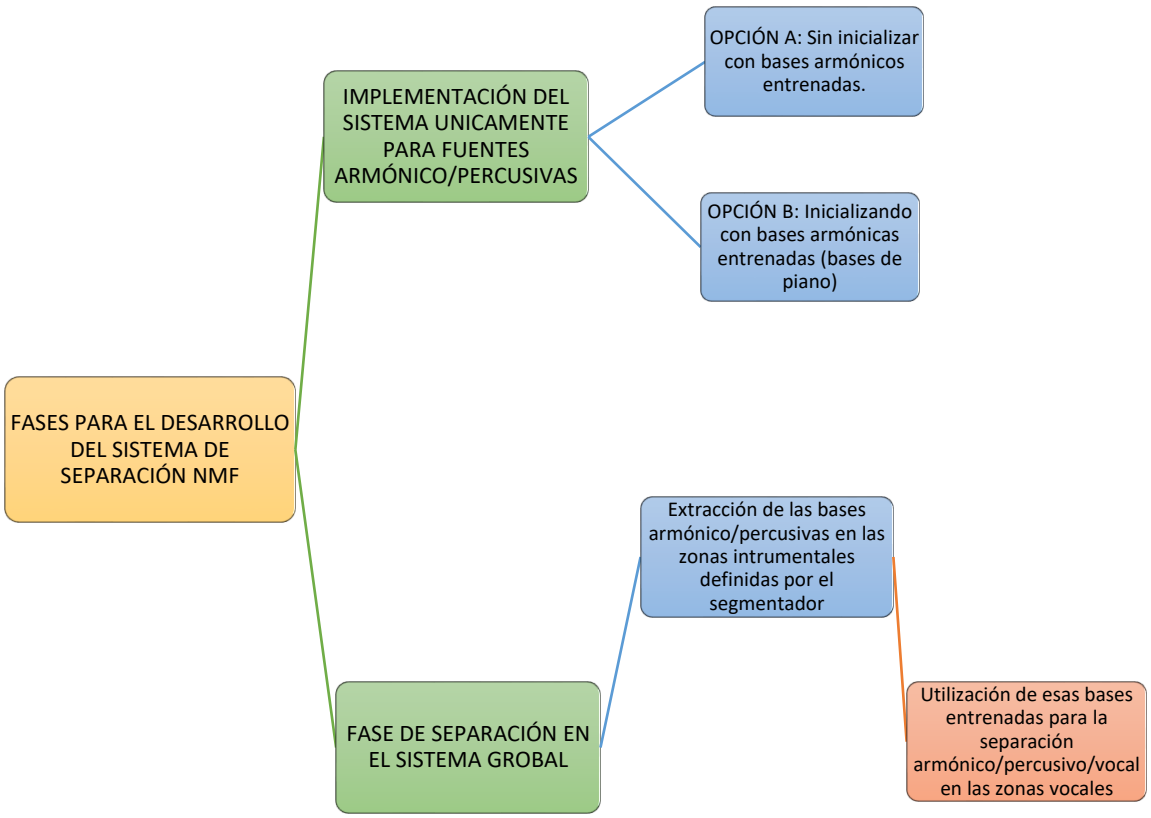

Figura 4.43: Esquema de las distintas fases a desarrollar en este apartado del Trabajo Fin de Master.

4.2.1 SISTEMA DE SEPARACIÓN UNICAMENTE PARA FUENTES ARMÓNICO/PERCUSIVAS

Como se ha comentado anteriormente es necesario realizar un análisis del modelo de separación que vamos a llevar a cabo sin tener en cuenta los resultados del segmentador. Es decir, este modelo de separación que se va a analizar no depende de los resultados del segmentador, y esto se hace necesario para demostrar que el sistema de separación NMF funciona de forma correcta independientemente del segmentador.

Imaginar que los resultados del segmentador son perfectos y conseguimos obtener las zonas instrumentales de una forma muy precisa. De nada nos serviría si el sistema de separación NMF no funcionase de forma correcta. Sin embargo, con este análisis previo se puede demostrar que el sistema de separación NMF funciona de forma correcta. De esta forma asegurando que tanto el sistema de segmentación como el de separación funcionan de forma eficiente por separado, podemos asegura que van a funcionar de forma eficaz de manera conjunta.

En esta sección el sistema de separación únicamente nos permitirá extraer dos fuentes sonoras, la armónica y la percusiva. Es por ello que las señales musicales que vamos a utilizar para nuestro análisis únicamente deben de estar formadas por las dos fuentes mencionadas anteriormente.

Para desarrollar este modelo se van a tomar dos opciones como se puede ver en el esquema mostrado en la figura 4.43, los cuales son:

- **OPCIÓN A: Sin inicializar con bases armónicas entrenadas.**
- **OPCIÓN B: Inicializando con bases armónicas entrenadas (bases de piano).**

Tras realizar un análisis detallado del sistema NMF a implementar podremos observar que la opción B es la más efectiva, ya que con ella partimos de unas bases que se consideran armónicas, por lo tanto, teniendo en cuenta que de alguna forma ya tenemos caracterizado el comportamiento armónico será más sencillo y con resultados más eficaces lograr separar las dos fuentes sonoras que entran en juego en este modelo.

Las dos opciones que se describen a continuación utilizan el mismo modelo NMF con la única diferencia de la inicialización de las bases armónicas, por ello se realizara una explicación detalla del modelo en la opción a y se observaran las diferencias al inicializar con las bases armónicas propias del piano en la opción b.

4.2.1.1 OPCIÓN A: Sin inicializar con bases armónicas entrenadas.

Cómo ya hemos hablado en la sección del estado del arte, existen diversos métodos que permiten la separación de fuentes sonoras en una pista ya mezclada. Para este TFM, se ha implementado una modificación del algoritmo NMF (Non-Negative Matrix Factorization). Se han asumido una serie de restricciones que ofrecen ciertas ventajas respecto al algoritmo original, en concreto, consideraciones de suavidad (smothness) y dispersión (sparseness).

Este modelo se basa en una modificación del algoritmo NMF que no requiere ningún postprocesado o entrenamiento para distinguir entre bases armónicas y percusivas [53] dado que la información se encuentra en el proceso de descomposición.

Se asume que los sonidos armónicos poseen cualidades de dispersión espectral y suavidad temporal, mientras que los sonidos percusivos poseen cualidades de dispersión temporal y suavidad espectral.

Entendemos los sonidos percusivos como aquellos producidos al golpear instrumentos, mientras que los sonidos armónicos se corresponden con aquellos que surgen como variaciones de frecuencias como es el caso de la guitarra.

El método aquí desarrollado presenta una serie de ventajas.

- Se trata de un método robusto para varias fuentes espectrales.
- Es un algoritmo eficiente y veloz en el procesamiento.
- No hay necesidad de definir umbrales.

Este método no necesita de entrenamiento ya que la información armónica/percusiva, es extraída del espectro tiempo-frecuencia que se usa en la factorización. Además, distingue automáticamente entre bases percusivas y bases armónicas.

En sonidos armónicos típicamente encontramos un espectrograma disperso en frecuencia; así mismo, para sonidos percusivos el espectrograma se caracteriza por ser impulsivo en frecuencia y disperso en tiempo.

Estas características permiten modelar sonidos armónicos usando la dispersión en frecuencia, picos espectrales, y la suavidad en tiempo, poca variación en la amplitud. Para los sonidos percusivos encontramos el caso contrario, en frecuencia la energía decrece lentamente mientras en tiempo la mayor parte de la energía se encuentra concentrada en pequeñas regiones.

En la Figura 4.44 se puede observar cómo se cumplen las propiedades que se han mencionado para sonidos armónicos y percusivos.

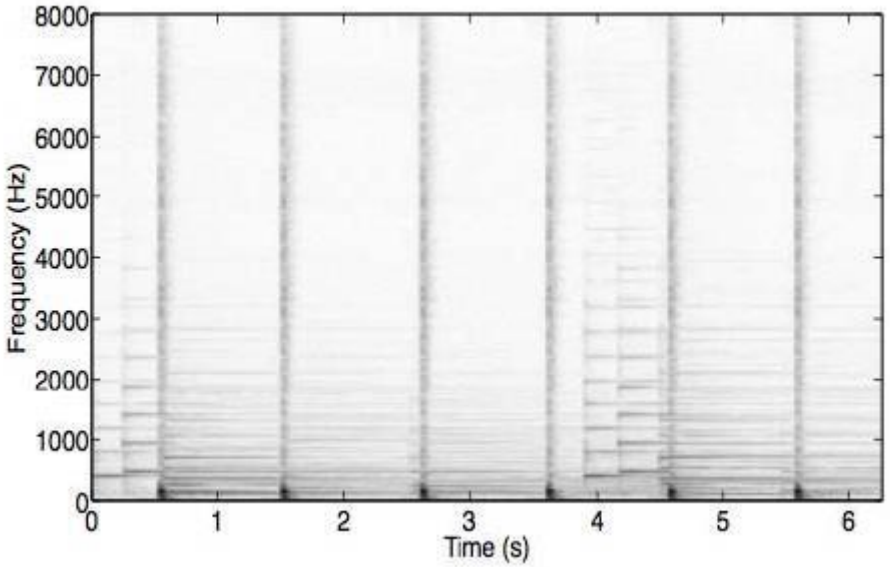

Figura 4.44: Figura de las restricciones de dispersión y suavidad para sonidos armónicos y percusivos.

El resultado de la descomposición de un espectrograma de entrada $\boldsymbol{X}$ mediante la factorización en una matriz no negativa, es dos matrices no negativas que notaremos como $\boldsymbol{W}$ y $\boldsymbol{H}$.

$$\boldsymbol{X} \approx \boldsymbol{W} * \boldsymbol{H} \tag{91}$$

El espectrograma magnitud $\boldsymbol{X}$ de una señal musical $\boldsymbol{x(t)}$, se compone de tramas $\boldsymbol{T}$ junto con bins de frecuencia $\boldsymbol{F}$ y un conjunto de unidades tiempo-frecuencia $X_{f,t}$. Cada $X_{f,t}$ se define con un bins de frecuencia f-th y un frame de tiempo t-th, calculado usando la magnitud de la STFT, Transformada de Fourier de Tiempo Reducido, usando una ventana $Hamming\ w(n)$ de N muestras y un desplazamiento temporal J.

La columna i-ésima de la matriz W es una base de frecuencia que representa el patrón espectral de una fuente de sonido que está activa en el espectrograma. Así, la fila i-ésima de la matriz H, es considerada la ganancia o activación temporal que representa el intervalo de tiempo en el cuál, la base de frecuencia i-ésima está activa.

El algoritmo original del NMF, no puede usarse para distinguir entre los dos tipos de bases mencionadas, ya que sólo asegura la convergencia a mínimos locales, que permite reconstruir el espectrograma, pero no distinguir entre las bases. Es por ello que se utilice

una modificación del algoritmo, aplicando las restricciones que nos caracterizan a los distintos tipos de fuentes sonoras.

Es necesario normalizar la entrada para llevar a cabo la correcta separación entre fuentes armónicas/percusivas. Por lo que el espectrograma $X_{n\beta}$, se calcula con el número de tramas, el número de bins y normalización de la norma beta. Esta normalización es necesaria para que las ecuaciones que definen las restricciones, como las ecuaciones de reconstrucción tengan una implementación más sencilla.

$$X_{n\beta} = \frac{X}{\sigma . X_{\beta}} = \frac{X}{\left(\frac{\sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t}^{\beta}}{FT}\right)^{\frac{1}{\beta}}} \tag{92}$$

Se define una función para factorizar un espectrograma Xn como mezcla de dos espectrogramas separados X_p y X_h. Cada espectrograma separado exhibe características espectro-temporales para los sonidos armónicos/percusivos.

$$X_{n\beta} \approx \hat{X}_n = \hat{X}_p + \hat{X}_h = W_{P_{F,Rp}} diag\left(ep_{Rp}\right) H_{P_{Rp,T}} + W_{H_{F,Rh}} diag(eh_{Rh}) H_{H_{Rh,T}} \tag{93}$$

Donde:

- $\hat{X}_n$: espectrograma estimado y normalizado de la función mezcla.
- $\hat{X}_p$: espectrograma estimado y normalizado de la fuente percusiva.
- $\hat{X}_h$: espectrograma estimado y normalizado de la fuente armónica.
- $W_{P_{F,Rp}}$: matriz de bases percusivas.
- $H_{P_{Rp,T}}$: matriz de activaciones percusivas.
- $W_{H_{F,Rh}}$: matriz de bases armónicas.
- $H_{H_{Rh,T}}$: matriz de activaciones armónicas.
- $diag\left(ep_{Rp}\right)$: diagonal de la norma de la parte percusiva.
- $diag(eh_{Rh})$: diagonal de la norma de la parte armónica.
- Rp: número de componentes percusivas utilizadas en la factorización.
- Rh: número de componentes armónicas utilizadas en la factorización.

Una vez detalladas las distintas matrices y vectores que componen nuestro sistema de forma matemática es necesario aclarar una serie de cosas:

- Siempre se van a trabajar con matrices y vectores normalizados, ya que de esta forma las ecuaciones de actualización y reconstrucción van a ser más sencillas.
- El uso de las normas 'e', las cuales se diagonalizaran posteriormente, en ese sistema nos permite poder operar con esas matrices normalizas.
- Es necesario que exista una norma 'e' por cada fuente sonora, en este caso únicamente existen dos fuentes, es por ello que se utilicen dos normas 'e'.

Por lo tanto, el objetivo del sistema es minimizar una función de coste global para estimar $W_{P_{F,Rp}}$, $H_{P_{Rp,T}}$, $W_{H_{F,Rh}}$ y $H_{H_{Rh,T}}$. Esta función de coste global depende del coste β-divergencia y de los costes procedentes de las restricciones que se asocian con la parte armónica y percusiva, es lo que se conoce como coste armónico y coste percusivo.

Por lo tanto, el siguiente paso es definir los distintos costes que nos van a permitir estimar las matrices de las activaciones y de las bases para las distintas fuentes sonoras presentes.

- **Coste β-divergencia:** El espectrograma normalizado de entrada $X_{n\beta}$ se construye minimizando el coste β-divergencia $d\beta(x|y)$ para los dos espectrogramas $\hat{X}_p$ y $\hat{X}_h$. Considerando que, $x = X_{n\beta}$ e $y = \hat{X}_p + \hat{X}_h$.

  En la ecuación 94 se puede observar las distintas funciones que existen en función del valor de β. Para la realización de este Trabajo Fin de Master, se ha seleccionado un valor de β=1,3. Este valor ha sido seleccionado tras la realización de diferentes pruebas, ya que se ha demostrado que actúa de la forma más eficiente.

$$d\beta(x|y) = \begin{cases} \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{1}{\beta(\beta-1)}\left(x_{f,t}^{\beta} + (\beta-1)y_{f,t}^{\beta} - \beta x_{f,t}\, y_{f,t}^{\beta-1}\right) & \beta \in (0,1)\cup(1,2] \\ \sum_{f=1}^{F}\sum_{t=1}^{T} x_{f,t} log\frac{x_{f,t}}{y_{f|t}} - x_{f,t} + y_{f,t} & \beta = 1 \\ \sum_{f=1}^{F}\sum_{t=1}^{T} \frac{x_{f,t}}{y_{f,t}} - log\frac{x_{f,t}}{y_{f,t}} - 1 & \beta = 0 \end{cases} \quad (94)$$

En este modelo, la aproximación de NMF usa una función objetiva que considera el costo β-divergencia y las características espectro-temporales comunes en las señales percusivas y armónicas. Por lo tanto, el siguiente paso es definir las restricciones que caracterizan a la fuente armónica y a la percusiva.

- **Coste percusivo:** se toman las consideraciones de suavidad en frecuencia y dispersión en tiempo. Los sonidos percusivos suelen representarse como señales de banda ancha, con energía confinada en cortos intervalos de tiempo. Se usa una definición de suavidad espectral, SSM, asociada con la matriz Wp. Un alto coste, se asocia con ganancias distintas de cero, suponiendo que los sonidos percusivos pueden representarse como transitorios con energías concentradas en cortos intervalos de tiempo. Este hecho se asocia con Hp.

$$SSM = \frac{T}{Rp}\sum_{rp=1}^{Rp}\frac{1}{\sigma W_{P_{F,Rp}}{}^{2}}\sum_{f=2}^{F}\left(W_{P_{f-1,rp}} - W_{P_{f,rp}}\right)^{2} \tag{95}$$

$$TSP = \frac{F}{R_p}\sum_{r_p=1}^{R_p}\sum_{t=1}^{T}\left|\frac{H_{P_{rp,t}}}{\sigma H_{P_{rp}}}\right| \tag{96}$$

$$\text{Con } \sigma H_{P_{rp}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} H_{P_{rp,t}}^{2}} \text{ y } \sigma W_{P_{rp}} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} W_{P_{rp,t}}^{2}}$$

- **Coste armónico:** Para los sonidos armónicos ocurre de forma similar que en el caso anterior. Sin embargo, en este caso se toman consideraciones de dispersión en frecuencia y suavidad en tiempo.

$$SP = \frac{T}{R_h}\sum_{r_h=1}^{R_h}\sum_{f=1}^{F}\left|\frac{W_{H_{f,rh}}}{\sigma W_{H_{rh}}}\right| \tag{97}$$

$$TSM = \frac{F}{R_h}\sum_{r_h=1}^{R_h}\frac{1}{\sigma H_{H_{rh}}{}^{2}}\sum_{t=2}^{T}(H_{H_{rh,t-1}} - H_{H_{rh,t}})^{2} \tag{98}$$

- **Algoritmo global NMF armónico/percusivo:** la función global D, incluido el coste β-divergencia, se define en la ecuación 99.

$$D = d\beta\left(X_{n\beta}|\left(X_p + X_p\right)\right) + K_{SSM}SMM + K_{TSP}TSP + K_{SSP}SSP + K_{TSM}TSM \quad (99)$$

Donde:

- K_{SSM}: es el peso que se le asignara la a restricción de suavidad para la matriz de las bases percusivas.
- K_{TSP}: es el peso que se le asignara la a restricción de dispersión para la matriz de las activaciones percusivas.
- K_{SSP}: es el peso que se le asignara la a restricción de dispersión para la matriz de las bases armónicas.
- K_{TSM}: es el peso que se le asignara la a restricción de suavidad para la matriz de las activaciones armónicas.

Esta es la clave que nos permitirá dar más peso a unas restricciones que a otras y lo que nos llevará a conseguir una separación lo más eficiente posible. Estos pesos se ajustan llevando a cabo distintas pruebas y análisis, de tal forma que observando los resultados obtenidos con unos pesos u otros podamos determinar cuáles son los pesos que se considerarían óptimos.

El siguiente paso es usar las llamadas “reglas de actualización multiplicativas”, de forma que $\boldsymbol{D}$ no aumente, mientras se garantiza la no negatividad mediante un algoritmo de gradiente descendente, usando una correcta elección del tiempo de salto y estimación para cada parámetro escalar Z. Se expresan las derivadas parciales de la función objetiva $\frac{\partial D}{\partial Z}$, como la diferencia entre $\left[\frac{\partial D}{\partial Z}\right]^+$ y $\left[\frac{\partial D}{\partial Z}\right]^-$. En este primer sistema que se ha desarrollado como se ha mencionado anteriormente con la idea de probar el sistema NMF por sí solo, únicamente en un proceso iterativo se obtienen las matrices de bases y activaciones de las fuentes armónico y percusivas.

$$Z = Z \odot \frac{\left[\frac{\partial D}{\partial Z}\right]^-}{\left[\frac{\partial D}{\partial Z}\right]^+} \quad (100)$$

Sustituyendo las bases percusivas Wp y de ganancia Hp dentro de la ecuación (100), se obtienen las siguientes reglas de actualización multiplicativas para sonidos percusivos.

$$W_p = W_p \odot \frac{\left[\frac{\partial d\beta}{\partial W_p}\right]^- + K_{SMM}\left[\frac{\partial SSM}{\partial W_p}\right]^-}{\left[\frac{\partial d\beta}{\partial W_p}\right]^+ + K_{SMM}\left[\frac{\partial SSM}{\partial W_p}\right]^+} \tag{101}$$

$$H_p = H_p \odot \frac{\left[\frac{\partial d\beta}{\partial H_p}\right]^- + K_{TSM}\left[\frac{\partial TSM}{\partial H_p}\right]^-}{\left[\frac{\partial d\beta}{\partial H_p}\right]^+ + K_{TSM}\left[\frac{\partial TSM}{\partial H_p}\right]^+} \tag{102}$$

$$\left[\frac{\partial d\beta}{\partial W_p}\right]^- = \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] diag(ep)' \tag{103}$$

$$\left[\frac{\partial d\beta}{\partial W_p}\right]^+ = \left[\left(\hat{X}_n\right)^{\beta-1}\right] \left(diag(ep)H_p\right)' \tag{104}$$

$$\left[\frac{\partial d\beta}{\partial H_p}\right]^- = \left(diag(ep)W_p\right)' \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{105}$$

$$\left[\frac{\partial d\beta}{\partial H_p}\right]^+ = \left(diag(ep)W_p\right)' \left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{106}$$

$$\left[\frac{\partial SSM}{\partial W_p}\right]^+_{f,rp} = \frac{1}{2(F-1)\left(\frac{rp}{F}\right)} 4W_p \tag{107}$$

$$\left[\frac{\partial SSM}{\partial W_p}\right]^-_{f,rp} = \frac{2}{2(F-1)\left(\frac{rp}{F}\right)} \left[\left(W_{P_{f-1,rp}} + W_{P_{f+1,rp}}\right) + \sum_{f=2}^{F}\left(W_{P_{f,rp}} - W_{P_{f-1,rp}}\right)^2\right] \tag{108}$$

$$\left[\frac{\partial TSP}{\partial H_p}\right]^+_{rp,t} = \frac{1}{\sqrt{T} * rp} \tag{109}$$

$$\left[\frac{\partial TSP}{\partial H_p}\right]^-_{rp,t} = 0 \tag{110}$$

Sustituyendo las bases armónicas Wh y de ganancia y Hh dentro de la ecuación (100), se obtienen las siguientes reglas de actualización multiplicativas para sonidos armónicos.

$$W_H = W_H \odot \frac{\left[\frac{\partial d\beta}{\partial W_H}\right]^- + K_{SSP}\left[\frac{\partial SSP}{\partial W_H}\right]^-}{\left[\frac{\partial d\beta}{\partial W_H}\right]^+ + K_{SSP}\left[\frac{\partial SSP}{\partial W_H}\right]^+} \tag{111}$$

$$H_H = H_H \odot \frac{\left[\frac{\partial d\beta}{\partial H_H}\right]^- + K_{TSM}\left[\frac{\partial TSM}{\partial H_H}\right]^-}{\left[\frac{\partial d\beta}{\partial H_H}\right]^+ + K_{TSM}\left[\frac{\partial TSM}{\partial H_H}\right]^+} \tag{112}$$

$$\left[\frac{\partial d\beta}{\partial W_h}\right]^- = \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] diag(eh)' \tag{113}$$

$$\left[\frac{\partial d\beta}{\partial W_h}\right]^+ = \left[\left(\hat{X}_n\right)^{\beta-1}\right] (diag(eh)H_h)' \tag{114}$$

$$\left[\frac{\partial d\beta}{\partial H_h}\right]^- = (diag(ep)W_h)' \left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{115}$$

$$\left[\frac{\partial d\beta}{\partial H_h}\right]^+ = (diag(ep)W_h)' \left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{116}$$

$$\left[\frac{\partial TSM}{\partial H_H}\right]^+_{rh,t} = \frac{1}{2(T-1)\left(\frac{rh}{T}\right)} 4H_h \tag{117}$$

$$\left[\frac{\partial TSM}{\partial H_h}\right]^-_{rh,t} = \frac{2}{2(T-1)\left(\frac{rh}{T}\right)} \left[\left(H_{H_{rh,t-1}} + H_{H_{rh,t+1}}\right) + \sum_{t=2}^{T}\left(H_{H_{rh,t}} - H_{H_{rh,t-1}}\right)^2\right] \tag{118}$$

$$\left[\frac{\partial SSP}{\partial W_h}\right]^+_{f,rh} = \frac{1}{\sqrt{F} * rh} \tag{119}$$

$$\left[\frac{\partial SSP}{\partial W_h}\right]^-_{f,rh} = 0 \tag{120}$$

Estas serían las actualizaciones de las normas para las matrices de activaciones y para las matrices de las bases.

$$ep = ep \odot \sqrt{\sum {W_p}^2} \tag{121}$$

$$eh = eh \odot \sqrt{\sum {W_e}^2} \tag{122}$$

$$ep = ep \odot \sqrt{\sum {H_p}^2} \tag{123}$$

$$eh = eh \odot \sqrt{\sum {H_h}^2} \tag{124}$$

La Tabla 4.3 muestra una aclaración para afianzar el funcionamiento de este primer modelo NMF que permite obtener las bases armónico/percusivas.

Variables de entrada	Funciones de coste	Restricciones	Peso de las restricciones	Matrices que evolucionan	Variables de salida
$\boldsymbol{x}$, señal musical, solo armónico y percusiva	Coste β-divergencia	SSM, suavidad en frecuencia (percusivo)	K_{SSM} , peso de la restricción SSM	W_p	W_p
$\boldsymbol{W_p}$, inicializada aleatoriamente	Coste armónico	TSP, dispersión en el tiempo (percusivo)	K_{TSM}, peso de la restricción TSM	H_p	H_p
$\boldsymbol{H_p}$, inicializada aleatoriamente	Coste percusivo	SSP, dispersión en frecuencia (armónico)	K_{TSP}, peso de la restricción TSP	W_h	W_h
$\boldsymbol{W_h}$, inicializada aleatoriamente		TSM, suavidad en el tiempo (armónico)	K_{SSP} , peso de la restricción SSP	H_h	H_h
$\boldsymbol{H_h}$, inicializada aleatoriamente			K_{SM}, valor común para K_{SSM} y K_{TSM}	ep (vector norma)	ep
$\boldsymbol{R_p}$, numero de bases percusivas			K_{SP}, valor común para K_{TSP} y K_{SSP}	eh (vector norma)	eh
$\boldsymbol{R_h}$, número de bases armónicas					
maxIter, número de iteraciones					

Tabla 4.3: Aclaración del modelo NMF para obtener las bases armónicas percusivas, donde ninguna de las matrices se mantiene fija y todas se inicializan aleatoriamente.

Para la actualización iterativa de las matrices Wp, Hp, Wh y Hh, se usan MaxIteriteraciones (la variable maxIter optimizada es 100, es decir, debe haber 100 iteraciones para que el algoritmo obtenga una separación óptima), en las cuales las bases armónicas y percusivas van optimizando su información hasta que llegan a converger.

Es necesario que quede claro que dentro del proceso iterativo en primer lugar se realiza la actualización de las distintas matrices, en segundo lugar, se realiza la normalización de cada matriz (independientemente de que se trate de una matriz de activaciones o de bases), además de la actualización de los vectores que representan las normas (ep, eh), y por último se lleva a cabo una reconstrucción con la ecuación global 93.

El último paso corresponde a la fase de reconstrucción. Tras el cálculo de las distintas bases factorizadas antes mencionadas, se calculan dos espectrogramas de salida, Xp, Xh correspondientes a la parte percusiva y armónica respectivamente.

$$X_p = W_p diag(ep) H_p \tag{125}$$

$$X_h = W_h diag(eh) H_h \tag{126}$$

La señal percusiva $xp(t)$ se sintetiza usando el espectrograma Xp como el producto de las bases factorizadas Wp, la diagonal de la norma ep y sus activaciones Hp. Para asegurar la reconstrucción, se genera una máscara Mp. La información de fase se calcula multiplicando la máscara por el espectrograma complejo Xc relacionado con la señal mezclada $x(t)$. Para la señal armónica $xh(t)$ se sigue el mismo procedimiento.

$$M_p = \frac{X_p}{X_p + X_h} \tag{127}$$

$$M_h = \frac{X_h}{X_p + X_h} \tag{128}$$

$$x_p(t) = IDFT(M_p X_c) \tag{129}$$

$$x_h(t) = IDFT(M_h X_c) \tag{130}$$

La Tabla 4.4 muestra una aclaración para afianzar el funcionamiento de la reconstrucción final de las señales obtenidas por el separador (armónica-percusiva).

Variables de entrada	Espectrogramas de las fuentes	Máscaras	Señales de salida
$\boldsymbol{W_p, diag(ep), H_p}$	X_p	M_p	$x_p(t)$
$\boldsymbol{W_h, diag(eh), H_h}$	X_h	M_h	$x_h(t)$

Tabla 4.4: Aclaración para afianzar el funcionamiento de la reconstrucción final de las señales obtenidas por el separador (armónica-percusiva).

Para llevar a cabo un mejor entendimiento del sistema se desarrolla un esquema de fases donde se pueden apreciar todos los pasos explicados anteriormente.

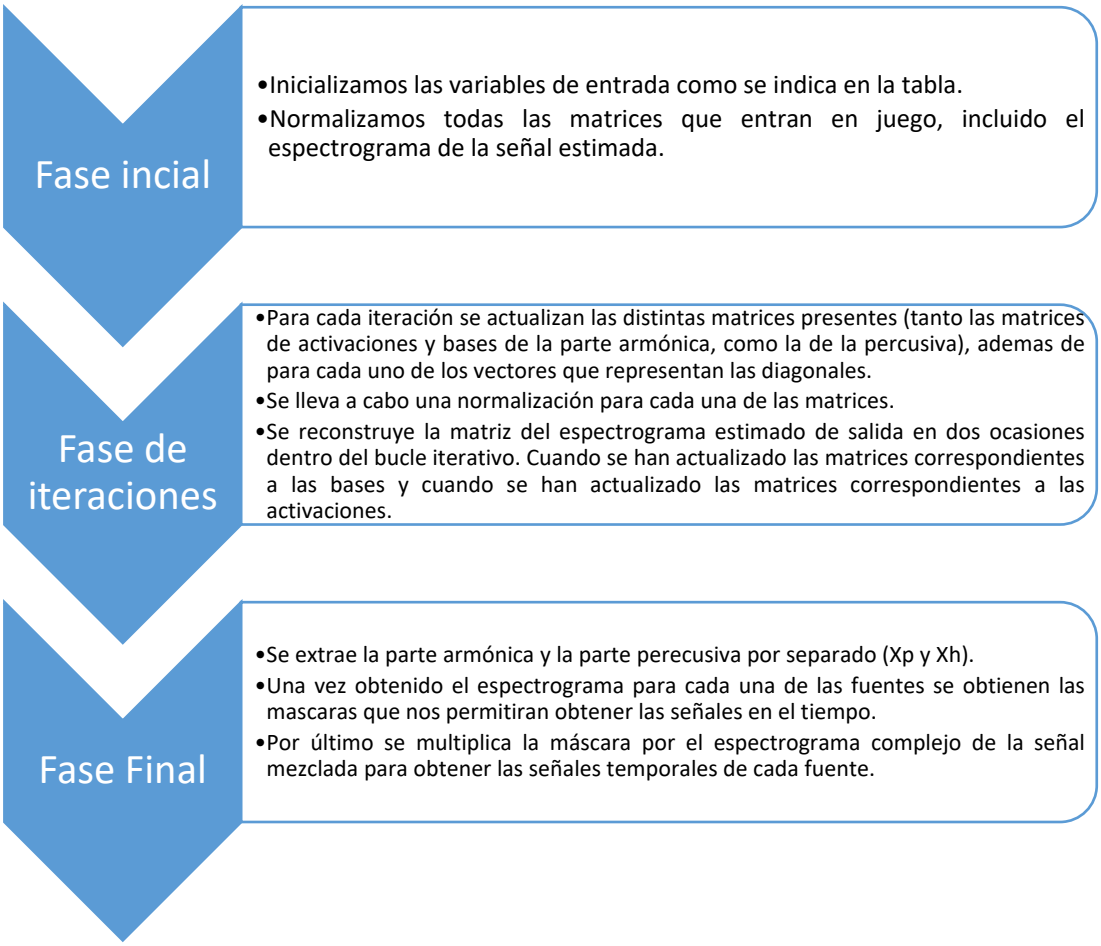

Figura 4.45: Esquema aclaratorio del proceso de separación NMF para llevar a cabo una separación armónico/percusiva.

Tanto para esta opción, como para la opción B vamos a mostrar una serie de imágenes que nos permitirán comprender la mejora existente entre elegir la opción A y

elegir la opción B. Se podrían mostrar una gran cantidad de imágenes que nos hiciesen ver la gran diferencia entre los resultados que se aprecian visualmente en un modelo y en otro. Por ejemplo, se podrían mostrar las matrices de activaciones o las matrices de bases para la fuente armónica y percusiva y ver si realmente tienen un comportamiento percusivo o armónico.

Pero para realizar un análisis de una forma más precisa y concreta se van a mostrar la forma del espectrograma perteneciente a la fuente armónica y la forma del espectrograma perteneciente a la fuente percusiva. Para la opción A (sin inicializar con bases armónicas entrenadas) los resultados visuales pertenecientes al espectrograma de la matriz armónica y al espectrograma de la matriz percusiva son:

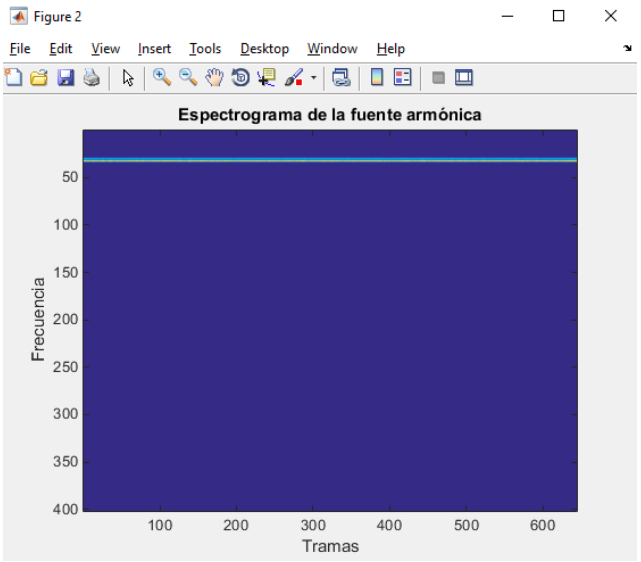

Figura 4.46: Espectrograma de la fuente armónica (opción A).

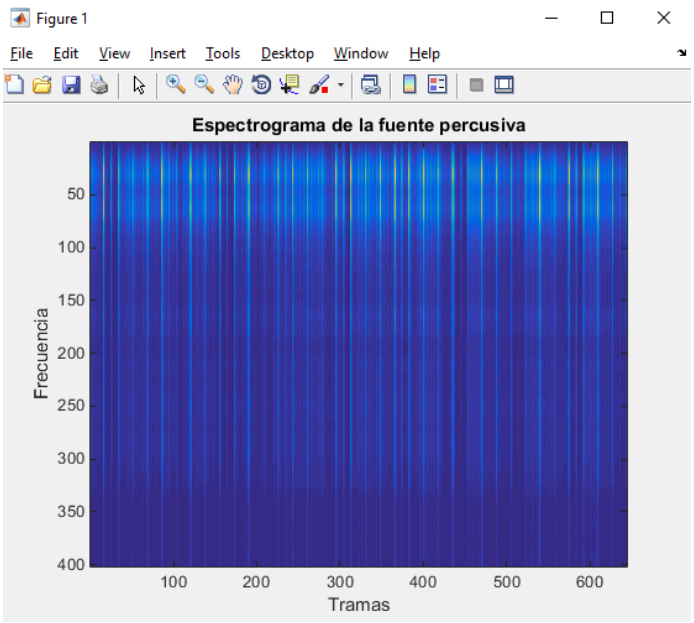

Figura 4.47: Espectrograma de la fuente percusiva (opción A).

Como se puede observar en las imágenes anteriores el espectrograma de la fuente armónica no llega a cumplir las características de los sonidos armónicos de una forma efectiva. Aunque por otro lado el espectrograma de percusivos si represente un comportamiento claramente percusivo, pero no exacto.

4.2.1.2 OPCIÓN B: Inicializando con bases armónicas entrenadas (bases de piano)

Este modelo es similar al modelo explicado anteriormente. Las únicas diferencias que existen con respecto al anterior son las siguientes:

- En este caso la matriz de las bases armónicas W_h se inicializan con una matriz de bases previamente entrenada. Esta matriz está formada con las bases correspondientes al piano, ya que es considerado uno de los instrumentos cruciales a la hora de estudiar el comportamiento armónico de dichos instrumentos.
- Por lo tanto, en este caso no es necesario aplicar ninguna restricción a la matriz W_h. En el caso anterior se aplicaba la restricción de dispersión (SP) a la matriz de las bases armónicas W_h. Esto era necesario para caracterizar de alguna forma el comportamiento armónico de la matriz. Sin embargo, en este caso no es necesario aplicar ninguna restricción para esa matriz, ya que partimos de un comportamiento armónico. Por lo tanto, únicamente se deja evolucionar (se actualiza) sin la necesidad de añadir restricciones.
- Como es de esperar y debido a que añadimos información adicional al sistema los resultados obtenidos serán mejores como se podrá observar.

A nivel matemático todas las ecuaciones utilizadas serian similares excepto la 111, que quedaría como se muestra en la ecuación 131.

$$W_H = W_H \odot \frac{\left[\frac{\partial d\beta}{\partial W_H}\right]^-}{\left[\frac{\partial d\beta}{\partial W_H}\right]^+} \tag{131}$$

Por lo tanto, la ecuación 119 y 120 no serían necesarias ya que no existe ningún tipo de restricción para la matriz de bases armónicas W_h.

El proceso sería similar al anterior por lo tanto únicamente hay que tener en cuenta que a la hora de inicializar las bases armónicas se utiliza una matriz previamente entrenada de bases armónicas de piano, y que a la hora de llevar a cabo la reconstrucción de dicha

matriz no se utiliza ningún tipo de reconstrucción, por lo tanto, el coste perteneciente a la matriz de las bases armónicas es 0, en otras palabras, el valor de SP=0.

De forma similar a como se ha realizado en el modelo de la opción A se van a mostrar una serie de resultados visuales para poder realizar una comparación entre las dos opciones. Se ha elegido mostrar el espectrograma de la fuente armónica y el espectrograma de la fuente percusiva al igual que para la opción A.

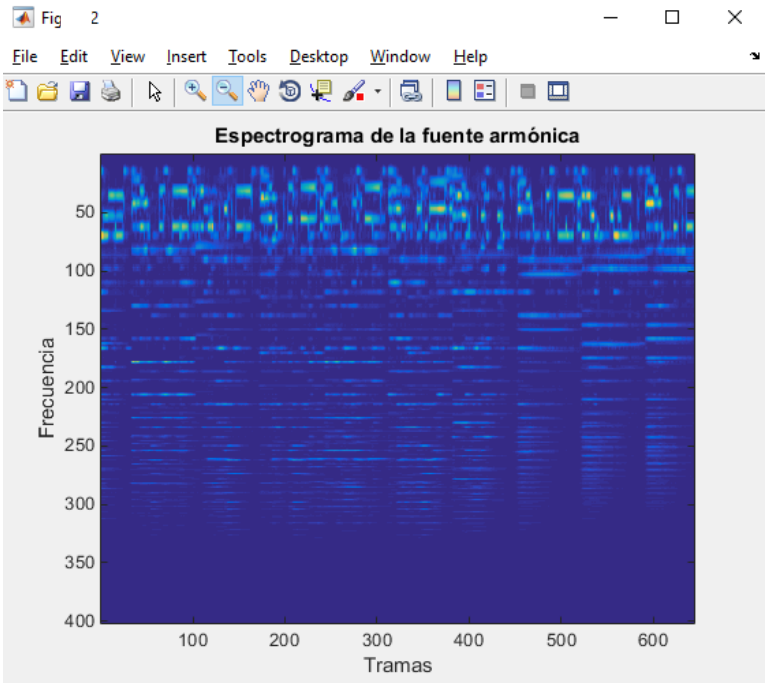

Figura 4.48: Espectrograma de la fuente armónica (opción B).

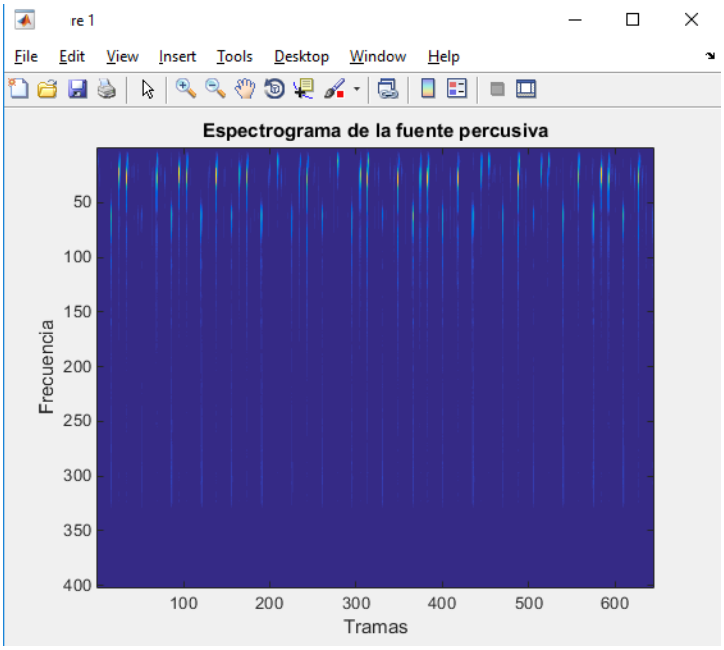

Figura 4.49: Espectrograma de la fuente percusiva (opción B).

Como conclusión final se puede observar que, llevando a cabo una inicialización de las bases armónicas, en este caso con una matriz de bases armónicas de piano, el funcionamiento del sistema se lleva a cabo de una forma más efectiva y precisa.

Por lo tanto, se puede observar que, en la segunda opción, al añadir información adicional, estamos ayudando al algoritmo a extraer mejor las dos fuentes sonoras (armónica y percusiva) por separado.

En definitiva, en este caso encontramos que los sonidos armónicos presentan un espectrograma disperso en frecuencia e impulsivo en el tiempo; así mismo, para sonidos percusios el espectrograma es impulsivo en frecuencia y disperso en tiempo. Al fin y albo este es el hecho que corresponde con la teoría definida anteriormente sobre el comportamiento de los sonidos armónicos y percusivos.

Como se verá en la sección de resultados las medidas objetivas aumentan de forma considerable utilizando la segunda opción, algo que se considera lógico al trabajar con información adicional añadida por las bases correspondientes a los sonidos armónicos.

Para que todo quede claro con la opción A conseguimos que el espectrograma de los sonidos percusivos tenga un espectro claramente percusivo, pero añadiendo información adicional, por lo que no se realiza de forma precisa. Sin embargo, el espectrograma de los sonidos armónicos no llega a tener un comportamiento exactamente armónico como se define en la teoría. Esto se debe a que, aunque se estén utilizando las restricciones que caracterizan a los sonidos armónicos, estamos partiendo de una matriz que se inicializa de forma aleatoria. Sin embargo, en la opción B tanto el espectrograma de los sonidos percusivos, como el espectrograma de los sonidos armónicos presentan unas características acordes a la definición de sonido armónico y percusivo, esto se debe a que además de utilizar restricciones para caracterizar a los dos tipos de sonidos, partimos de una información adicional que nos muestra cómo se comportan esos sonidos. En concreto se utiliza una matriz de bases armónica entrenada (correspondiente al piano) que como se ha mencionado anteriormente facilita el funcionamiento del sistema basado en NMF.

4.2.2 FASE DE SEPARACIÓN NMF EN EL SISTEMA GROBAL

Una vez que se ha llevado a cabo una comprobación de que el sistema de separación NMF funciona de forma correcta por separado, podemos realizar la unión entre el sistema de segmentación y el sistema de separación NMF. Comprobando que ambos sistemas funcionan de forma eficiente por separado podemos asegurar que van a funcionar de forma eficaz en conjunto.

En esta sección se va a explicar de forma detallada en que consiste el sistema de separación NMF que nos permita obtener las tres fuentes sonoras que caracterizan a cualquier señal musical, dichas fuentes son: la fuente vocal, la fuente armónica y la fuente percusiva.

El sistema general utilizado para la separación de las tres fuentes sonoras está compuesto por dos modelos, como se ha mencionado anteriormente. El primer modelo nos permitirá realizar un entrenamiento de las bases armónico y percusivas que están presentes en las zonas instrumentales determinadas por el segmentador. Por lo tanto, es necesario que el segmentador funcione de forma correcta para que el primer modelo del separador pueda obtener el entrenamiento de las bases armónico y percusivas de una forma eficiente y precisa.

Una vez que ya se disponen de las bases armónico y percusivas entrenadas en las zonas instrumentales se pueden utilizar en un segundo modelo para la extracción de la parte armónica, percusiva y vocal de las zonas vocales. Por lo tanto, es necesario distinguir entre dos zonas importantes que son las que nos proporciona el segmentador:

- **Zonas instrumentales:** estas zonas son las que están compuestas únicamente de fuentes armónicas y percusivas. Por lo tanto, estas son las zonas utilizadas por el primer modelo de separación para llevar a cabo un entrenamiento de las bases armónicas y percusivas. En estas zonas se puede asegurar que no existe parte vocal, por lo tanto, es más sencillo caracterizar y entrenar dichas bases.
- **Zonas vocales:** estas zonas son aquellas donde además de aparecer fuentes armónico y percusivas está presente la fuente vocal. Por lo tanto, son en estas zonas donde se lleva a cabo el segundo modelo del separador NMF que nos va a permitir extraer las fuentes armónico, percusivo y vocal de las zonas instrumentales. Para ello utilizara las bases armónico y percusivas que han sido entrenadas en el modelo anterior.

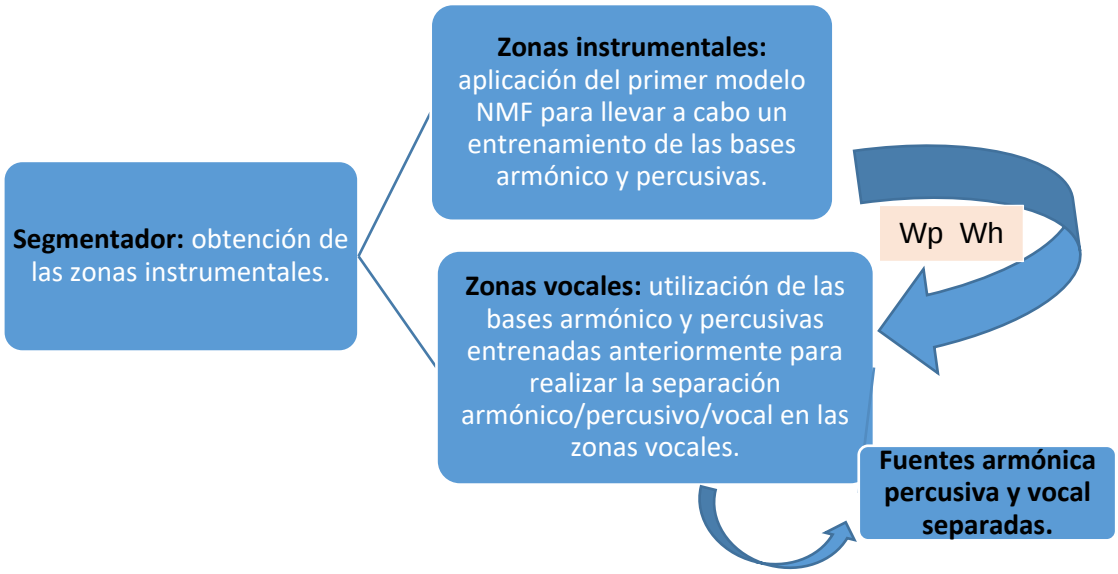

Figura 4.50: Esquema representativo para la fase de separación, incluyendo el propósito de la fase de segmentación.

4.2.2.1 Modelo NMF para extraer las bases instrumentales de las zonas instrumentales

Este modelo es el que se lleva a cabo en las zonas instrumentales para obtener las matrices armónico y percusivas entrenadas, y de esta forma añadir en el segundo modelo información adicional para conseguir una separación armónica/percusivo/vocal.

Más concretamente se va a llevar a cabo un modelo de separación basado en NMF únicamente en las zonas instrumentales determinadas por el segmentador. Por lo tanto, lo que se busca es obtener las bases armónico y percusivas entrenadas.

Para llevar a cabo este modelo se va a utilizar el mismo modelo explicado en la sección 4.2.1.2 (OPCIÓN B: Inicializando con bases armónicas entrenadas (bases de piano)). Se utiliza el mismo modelo ya que se ha demostrado anteriormente que el sistema NMF implementado para dos fuentes sonoras funciona de forma óptima cuando inicializamos la matriz de las bases armónicas con bases de piano.

Como ya se ha realizado una explicación detallada del sistema queda decir que el objetivo de este modelo es obtener las matrices de las bases armónicas (Wh) y percusivas (Wp). Estas matrices son las que se utilizaran en el segundo modelo para inicializar las matrices de las bases armónicas y percusivas, y de esta forma partir de una información adicional que nos va a ayudar a realizar una separación más óptima en las zonas vocales.

4.2.2.2 *Modelo NMF para realizar una separación armónica/percusivo/vocal de las zonas vocales*

Una vez que se ha empleado un modelo NMF inicial basado en el concepto de $\lambda_{DINÁMICO}$ para utilizarlo de ayuda a la hora de implementar el segmentador automático, que nos determina las zonas donde únicamente existen bases instrumentales y además una vez que se ha utilizado el primer modelo NMF para realizar una separación armónico/percusiva que permita extraer las bases instrumentales (armónico y percusivas), queda implementar el segundo modelo NMF [53] para conseguir la separación final de las fuentes sonoras (armónico/percusivo/vocal).

Ya disponemos de las bases armónicas y percusivas entrenadas en la fase anterior, por lo tanto, ya le estamos suministrando información adicional al sistema NMF para que puede llevar a cabo con mayor facilidad la separación armónico-percusivo-vocal.

Además del primer modelo desarrollado, se usa una segunda modificación del mismo para el desarrollo de este TFM. En la factorización del espectrograma $X_{n\beta}$, se añade el espectrograma $\hat{X}_v$ estimado, como se indica en la ecuación 132.

$$\begin{aligned} X_{n\beta} \approx \hat{X}_n &= \hat{X}_v + \hat{X}_p + \hat{X}_h \\ &= \left(W_{F0_{F,Re}} diag(eFO_{Re}) H_{F0_{Re,T}} * W_{F_{F,Rf}} diag\left(eF_{Rf}\right) H_{F_{Rf,T}}\right) \\ &+ W_{P_{F,Rp}} diag\left(ep_{Rp}\right) H_{P_{Rp,T}} + W_{H_{F,Rh}} diag(eh_{Rh}) H_{H_{Rh,T}} \end{aligned} \quad (132)$$

Donde:

- $\hat{X}_n$: espectrograma estimado y normalizado de la función mezcla.
- $\hat{X}_p$: espectrograma estimado y normalizado de la fuente percusiva.
- $\hat{X}_h$: espectrograma estimado y normalizado de la fuente armónica.
- $\hat{X}_v$: espectrograma estimado y normalizado de la fuente vocal.
- $W_{P_{F,Rp}}$: matriz de bases percusivas.
- $H_{P_{Rp,T}}$: matriz de activaciones percusivas.
- $W_{H_{F,Rh}}$: matriz de bases armónicas.
- $H_{H_{Rh,T}}$: matriz de activaciones armónicas.
- $W_{F0_{F,Re}}$: matriz de bases de excitaciones vocales.
- $H_{F0_{Re,T}}$: matriz de activaciones de excitaciones vocales.
- $W_{F_{F,Rf}}$: matriz de bases de filtros vocales.
- $H_{F_{Rf,T}}$: matriz de activaciones de filtros vocales.
- $diag\left(ep_{Rp}\right)$: diagonal de la norma de la parte percusiva.

- $diag(eh_{Rh})$: diagonal de la norma de la parte armónica.
- $diag(eFO_{Re})$: diagonal de la norma de las excitaciones vocales.
- $diag(eF_{Rf})$: diagonal de la norma de los filtros vocales.
- Rp: número de componentes percusivas utilizadas en la factorización.
- Rh: número de componentes armónicas utilizadas en la factorización.
- Re: número de componentes de la matriz de excitaciones vocales.
- Rf: número de componentes de la matriz de filtros vocales.

$W_{F0_{F,Re}}$ constituye la matriz de excitaciones vocales y $H_{F0_{Re,T}}$ contiene sus activaciones. Así mismo, $W_{F_{F,Rf}}$ constituye la matriz de filtros vocales y $H_{F_{Rf,T}}$ sus activaciones.

La matriz de bases de excitaciones vocales $W_{F0_{F,Re}}$ se inicializa con el modelo de excitación global (donde cada columna contiene la excitación, en frecuencia, de un pitch) y se deja fija durante todo el proceso, en cambio, la matriz de bases de fonema $W_{F_{F,Rf}}$ se deja fija inicializándolas previamente con bases de fonema entrenadas en una base de datos de voz.

Las bases $H_{F0_{Re,T}}$ y $H_{F_{Rf,T}}$ se irán actualizando mediante las reglas de actualización multiplicativas, al igual que las bases percusivas y armónicas. Sin embargo, es importante tener en cuenta que las bases $H_{F0_{Re,T}}$ y $H_{F_{Rf,T}}$ se deben de poner a 0 en las zonas instrumentales, ya que no existe voz en esas zonas.

Los costes presentes en este modelo general de separación, además del coste propio de la **β-divergencia** son:

- **Coste percusivo:** se toman las consideraciones de suavidad en frecuencia y dispersión en tiempo. Los sonidos percusivos suelen representarse como señales de banda ancha, con energía confinada en cortos intervalos de tiempo. Se usa una definición de suavidad espectral, SSM, asociada con la matriz Wp. Un alto coste, se asocia con ganancias distintas de cero, suponiendo que los sonidos percusivos pueden representarse como transitorios con energías concentradas en cortos intervalos de tiempo. Este hecho se asocia con Hp.

$$SSM = \frac{T}{Rp}\sum_{rp=1}^{Rp}\frac{1}{\sigma W_{P_{F,Rp}}{}^{2}}\sum_{f=2}^{F}\left(W_{P_{f-1,rp}} - W_{P_{f,rp}}\right)^{2} \quad (133)$$

$$TSP = \frac{F}{R_p}\sum_{r_p=1}^{R_p}\sum_{t=1}^{T}\left|\frac{H_{P_{rp,t}}}{\sigma H_{P_{rp}}}\right|$$

$$\text{Con } \sigma H_{P_{rp}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} H_{P_{rp,t}}^{2}} \text{ y } \sigma W_{P_{rp}} = \sqrt{\frac{1}{F}\sum_{f=1}^{F} W_{P_{rp,t}}^{2}} \quad (134)$$

- **Coste armónico:** para los sonidos armónicos ocurre de forma similar que en el caso anterior. Sin embargo, en este caso se toman consideraciones de dispersión en frecuencia y suavidad en tiempo.

$$SP = \frac{T}{R_h}\sum_{r_h=1}^{R_h}\sum_{f=1}^{F}\left|\frac{W_{H_{f,rh}}}{\sigma W_{H_{rh}}}\right| \quad (135)$$

$$TSM = \frac{F}{R_h}\sum_{r_h=1}^{R_h}\frac{1}{\sigma H_{H_{rh}}{}^{2}}\sum_{t=2}^{T}(H_{H_{rh,t-1}} - H_{H_{rh,t}})^{2} \quad (136)$$

- **Coste vocal:** para los sonidos vocales es importante tener en cuenta que únicamente la voz puede pronunciar o cantar una única nota por cada instante, por lo tanto, se aplicara la restricción de monofonía a las matrices de activaciones de los filtros y las excitaciones vocales.

$$SAHF0 = \frac{1}{Re(Re-1)}\sum_{i=1}^{Re}\sum_{j=1}^{T} HF0 * HF0^{T} - Trace(HF0 * HF0^{T}) \quad (137)$$

$$SAHF = \frac{1}{Rf(Rf-1)}\sum_{i=1}^{Rf}\sum_{j=1}^{T} HF * HF^{T} - Trace(HF * HF^{T}) \quad (138)$$

Donde Trace se refiere al cálculo de la diagonal de las matrices entre paréntesis.

Se definen por tanto las variables $KSAHFO$ y $KSAHF$. Que hacen referencia al peso de las restricciones del single-activation que se van a aplicar a las matrices de activaciones de los filtros y las excitaciones vocales.

Por otro lado, es importante destacar que $W_{P_{F,Rp}}$ y $W_{H_{F,Rh}}$se inicializan con las bases percusivas y armónicas aprendidas en el primer modelo NMF de separación (armónico/percusivo), mientras que $H_{P_{Rp,T}}$ y $H_{H_{Rh,T}}$ se inicializan con las activaciones aprendidas en el modelo NMF de separación (armónico/percusiva) anterior. Se inicializan de esta forma para las zonas instrumentales detectadas por el segmentador automático, y además se quedan fijas. Por otro lado, para aquellas zonas no instrumentales se inicializan aleatoriamente y se dejan evolucionar.

El cometido de esta segunda modificación, es separar la parte vocal de los sonidos armónicos/percusivos de forma más precisa, resultando pistas instrumentales más limpias de voz.

Se define la nueva función global D como se muestra en la ecuación (139):

$$D = d\beta\left(X_{n\beta}|\left(X_p + X_p\right)\right) + K_{SSM}SMM + K_{TSP}TSP + K_{SSP}SSP + K_{TSM}TSM + K_{SAHFO}SAHFO + K_{SAHF}SAHF \quad (139)$$

A continuación, se obtienen las siguientes reglas de actualización multiplicativas para la base $H_{F0_{Re,T}}$.

$$H_{F0} = H_{F0} \odot \frac{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{-} + K_{SAHFO}\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{-}}{\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{+} + K_{SAHFO}\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{+}} \quad (140)$$

$$\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{-} = (diag(eF0)W_{F0})'\left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \quad (141)$$

$$\left[\frac{\partial d\beta}{\partial H_{F0}}\right]^{+} = (diag(ep)W_{F0})'\left[\left(\hat{X}_n\right)^{\beta-1}\right] \quad (142)$$

$$\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{-} = \frac{2}{Re(Re-1)}H_{F0} \quad (143)$$

$$\left[\frac{\partial SAHFO}{\partial H_{F0}}\right]^{+} = \frac{2}{Re(Re-1)} 1_{Re,Re} H_{F0} \tag{144}$$

A continuación, se obtienen las siguientes reglas de actualización multiplicativas para la base $H_{F_{Rf,T}}$.

$$H_F = H_F \odot \frac{\left[\frac{\partial d\beta}{\partial H_F}\right]^{-} + K_{SAHF}\left[\frac{\partial SAHF}{\partial H_F}\right]^{-}}{\left[\frac{\partial d\beta}{\partial H_F}\right]^{+} + K_{SAHF}\left[\frac{\partial SAHF}{\partial H_F}\right]^{+}} \tag{145}$$

$$\left[\frac{\partial d\beta}{\partial H_F}\right]^{-} = (diag(eF)W_F)'\left[\left(\hat{X}_n\right)^{\beta-2} \odot X_{n\beta}\right] \tag{146}$$

$$\left[\frac{\partial d\beta}{\partial H_F}\right]^{+} = (diag(eF)W_F)'\left[\left(\hat{X}_n\right)^{\beta-1}\right] \tag{147}$$

$$\left[\frac{\partial SAHF}{\partial H_F}\right]^{-} = \frac{2}{Rf(Rf-1)} H_F \tag{148}$$

$$\left[\frac{\partial SAHF}{\partial H_F}\right]^{+} = \frac{2}{Rf(Rf-1)} 1_{Rf,Rf} H_F \tag{149}$$

Estas serían las actualizaciones de las normas para las matrices de activaciones y para las matrices de las bases.

$$ep = ep \odot \sqrt{\sum {W_p}^2} \tag{150}$$

$$eh = eh \odot \sqrt{\sum {W_e}^2} \tag{151}$$

$$eF0 = eF0 \odot \sqrt{\sum {W_{F0}}^2} \tag{152}$$

$$eh = eh \odot \sqrt{\sum {H_h}^2} \tag{153}$$

$$eF0 = eF0 \odot \sqrt{\sum {W_{F0}}^2} \tag{154}$$

$$eF = eF \odot \sqrt{\sum {W_F}^2} \tag{155}$$

$$ep = ep \odot \sqrt{\sum {H_p}^2} \tag{156}$$

La Tabla 4.5 muestra una aclaración para afianzar el funcionamiento del segundo modelo NMF que permite obtener la separación de las fuentes sonoras (H/P/V).

Variables de entrada	Funciones de coste	Restricciones	Peso de las restricciones	Matrices fijas	Matrices que evolucionan	Variables de salida
$\boldsymbol{x}$, señal global completa	Coste β-divergencia	SSM, suavidad en frecuencia (percusivo)	K_{SSM}, peso de la restricción	W_p, en las zonas musicales	W_p, en las zonas vocales	W_p
$\boldsymbol{W_p}$, igual que la obtenida anteriormente	Coste armónico	TSP, Dispersión en el tiempo (percusivo)	K_{TSM}, peso de la restricción TSM	H_p, en las zonas musicales	H_p, en las zonas vocales	H_p
$\boldsymbol{H_p}$, igual que la obtenida anteriormente	Coste percusivo	SSP, dispersión en frecuencia (armónico)	K_{TSP}, peso de la restricción TSP	W_h, en las zonas musicales	W_h, en las zonas vocales	W_h
$\boldsymbol{W_h}$, igual que la obtenida anteriormente	Coste vocal	TSM, suavidad en el tiempo (armónico)	K_{SSP} , peso de la restricción	H_h, en las zonas musicales	H_h, en las zonas vocales	H_h
$\boldsymbol{H_h}$, igual que la obtenida anteriormente		SAHFO, Monofonía en el tiempo (vocal).	K_{SM}, valor común para K_{SSM} y K_{TSM}	W_F, inicializada con bases vocales	$H_{F\square}$, en las zonas vocales	W_{F0}
Re, número de excitaciones utilizadas en		SAHF, Monofonía en el tiempo (vocal)	K_{SP}, valor común para K_{TSP} y K_{SSP}	H_{FO}, con valor 0 en las zonas musicales	H_f, en las zonas vocales	H_{FO}
Rf**,** número de bases vocales utilizadas en la			K_{SAHFO}	H_f, con valor 0 en las zonas musicales	W_{F0}	H_f
$\boldsymbol{Rp}$, numero de bases percusivas			K_{SAHF}		ep vector norma	W_F
$\boldsymbol{Rh}$, número de bases armónicas					eh vector norma	
$\boldsymbol{W_{FO}}$, inicializada con el modelo de excitación global					ef vector norma	
$\boldsymbol{H_{FO}}$					eF0 vector norma	
$\boldsymbol{W_f}$, inicializada con bases de fonemas						
$\boldsymbol{H_f}$						

Tabla 4.5: Aclaración del segundo modelo NMF para obtener las fuentes H/P/V, partiendo de las babes armónico y percusivas obtenidas del primer modelo

Es necesario que quede claro que dentro del proceso iterativo en primer lugar se realiza la actualización de las distintas matrices, en segundo lugar, se realiza la

normalización de cada matriz (independientemente de que se trate de una matriz de activaciones o de bases), además de actualizar los vectores que representan las diagonales de la norma (ep, eh, eF y eF0).

4.2.2.3 Reconstrucción de la señal

Tras el cálculo de las distintas bases factorizadas antes mencionadas, se calculan tres espectrogramas de salida, Xp, Xh y Xv correspondientes a la parte percusiva, la parte armónica y armónica respectivamente.

$$X_p = W_p diag(ep) H_p \tag{157}$$

$$X_h = W_h diag(eh) H_h \tag{158}$$

$$X_v = W_{F0} diag(eFO) H_{F0} * W_F diag(eF) H_F \tag{159}$$

La señal percusiva $xp(t)$ se sintetiza usando el espectrograma Xp como el producto de las bases factorizadas Wp, la diagonal de la norma ep y sus activaciones Hp. Para asegurar la reconstrucción, se genera una máscara Mp. La información de fase se calcula multiplicando la máscara por el espectrograma complejo Xc relacionado con la señal mezclada $x(t)$. Para la señal armónica $xh(t)$ y la señal vocal $xv(t)$ se sigue el mismo procedimiento.

$$M_p = \frac{X_p}{X_p + X_h + X_v} \tag{160}$$

$$M_h = \frac{X_h}{X_p + X_h + X_v} \tag{161}$$

$$M_v = \frac{X_v}{X_p + X_h + X_v} \tag{162}$$

$$x_p(t) = IDFT(M_p X_c) \tag{163}$$

$$x_h(t) = IDFT(M_h X_c) \tag{164}$$

$$x_v(t) = IDFT(M_v X_c) \tag{165}$$

La Tabla 4.6 muestra una aclaración para afianzar el funcionamiento de la reconstrucción final de las señales obtenidas por el separador (armónica-percusiva-vocal).

Variables de entrada	Espectrogramas de las fuentes	Mascaras	Señales de salida
$\boldsymbol{W_p, diag(ep), H_p}$	X_p	M_p	$x_p(t)$
$\boldsymbol{W_h, diag(eh), H_h}$	X_h	M_h	$x_h(t)$
$\boldsymbol{W_{F0}, diag(eFO), H_{F0}, W_F, diag(eF), H_F}$	X_v	M_v	$x_v(t)$

Tabla 4.6: Aclaración para afianzar el funcionamiento de la reconstrucción final de las señales obtenidas por el separador (armónica-percusiva).

Para llevar a cabo un mejor entendimiento del sistema se desarrolla un esquema de fases donde se pueden apreciar todos los pasos explicados anteriormente.

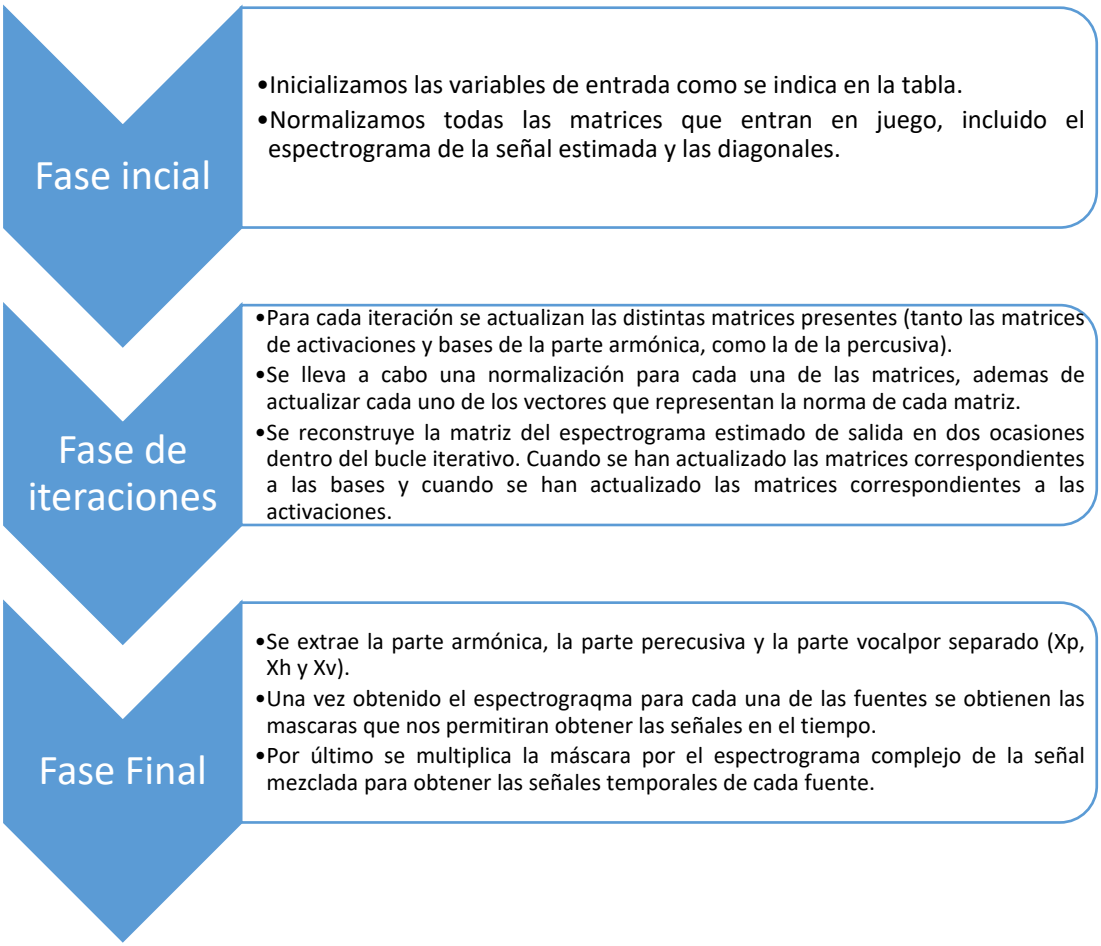

Figura 4.51 : Esquema aclaratorio del proceso de separación NMF para llevar a cabo una separación armónico/percusiva.

5 RESULTADOS Y DISCUSIÓN

Una vez se ha implementado el sistema final compuesto por la segmentación automática y la separación de las diferentes fuentes sonoras, es necesario realizar una valoración de los resultados obtenidos en cada una de las etapas para calificar mediante una serie de medidas objetivas la implementación del sistema.

Es importante tener en cuenta que el Trabajo Fin de Master tiene el objetivo principal de llevar a cabo un sistema de forma totalmente automática que sea capaz de segmentar en primer lugar entre zonas vocales y zonas instrumentales, y de separar en segundo lugar las tres fuentes sonoras que componen una señal musical (armónico, percusivo, vocal).

Este capítulo se divide en dos etapas fundamentales que permitirán contemplar los resultados obtenidos para cada una de las fases del sistema final, estas etapas son:

- **Segmentador automático:** en esta etapa de los resultados se realizara una evaluación de los resultados obtenidos para el segmentador, realizando en primer lugar una comparación con los resultados obtenidos del sistema manual (tramos musicales detectados por el sistema auditivo humano e implementados en un fichero de texto), en segundo lugar una comparación con los resultados obtenidos en el Trabajo Fin de Grado realizado en el año 2015 (encargado de la segmentación, separación y transcripción de señales musicales correspondientes al Flamenco) y los obtenidos por Justin Salomon y Emilia Gomez [29], y por último una comparación con uno de los sistemas encargado de realizar la extracción de la línea melódica en una grabación monoaural desarrollado por Bernhard Lehner [32], comprendido en el estado del arte y en la bibliografía descrita en el Trabajo Fin de Master.
- **Sistema de separación basado en modelos NMF:** en esta etapa de los resultados se realizará una evaluación de los resultados obtenidos por el sistema de separación basado en NMF. En primer lugar, se realiza una evaluación del sistema NMF únicamente con fuentes armónicas y percusivas. En este punto se evaluará el sistema NMF armónico/percusivo óptimo con el sistema NMF armónico/percusivo sin mejoras para observar los resultados y afirmar que el sistema de separación basado en NMF funciona de forma correcta independientemente de la etapa del segmentador (ya que en este caso no existe fuente vocal, por lo que no es necesario determinar las zonas instrumentales). Por otro lado, se realizará una evaluación del sistema NMF armónico/percusivo/vocal implementado en el TFM y el implementado en el TFG. Por último, se comparará con uno de los sistemas que obtuvo mejores resultados en la separación de señales musicales por Durrieu [64].

5.1 Segmentador Automático

En esta sección se llevarán a cabo la implementación de los distintos resultados objetivos para evaluar la etapa de segmentación. Para no tener que incluir un apartado de "Setup" por cada una de los análisis llevados a cabo (ya que el sistema óptimo que se compara en todos los casos es el mismo) se introducirá un único punto que lo defina en este apartado.

5.1.1 Setup

Para realizar una segmentación lo más óptima posible de forma global, independientemente del género musical, es necesario implementar el segmentador tal y como se ha descrito en la memoria del TFM. Por lo tanto, el segmentador se compondrá de 4 fases. En la primera se realizará una primera detección de las zonas vocales utilizando un modelo de separación NMF, en el cual es importante dar mayor peso a las restricciones vocales que a las armónicas y las percusivas, al fin y al cabo, lo que nos interesa es quedarnos únicamente con la parte vocal. Es importante tener en cuenta la inicialización de las bases armónicas con unas previamente entrenadas. En la segunda fase se lleva a cabo una discriminación entre zonas vocales y zonas instrumentales, para ello se aplica un peso restrictivo a la máscara vocal obtenida de la fase anterior (el valor del peso será cercano al 0.8, ya que de esta forma nos aseguramos que nos quedamos con la parte vocal) y mediante un umbral definido por el algoritmo de Otsu podemos discriminar zonas instrumentales de las vocales aplicándoselo a la energía obtenida de la máscara a la que se le ha aplicado el peso restrictivo. La tercera fase nos permite determinar las zonas instrumentales y para ello aplicábamos el clasificador HMM. En la última fase se tendrán en cuenta características musicales, de tal forma que se seleccionarán aquellos tramos instrumentales que superen los 0.7 segundos de duración, considerada la duración mínima en apreciar zona instrumental sin simular el efecto de los artefactos o ruidos aislados.

5.1.2 Segmentador automático TFM vs Segmentador manual

5.1.2.1 Base de datos

La base de datos que se utilizará para esta evaluación corresponde con un conjunto de 9 canciones pertenecientes al género pop. Esta base de datos fue utilizada en el ISMIR (The International Society of Music Information Retrieval) en el año 2015. Está formada por 9 canciones de 30 segundos de duración donde están presentes todas las fuentes (armónica, percusiva y vocal).

NOMBRE DE LA CANCIÓN
Hollywood_Nights_phv
Hotel_California_phv
Hurts_So_Good_phv
La_Bamba_phv
Make_It_Wit_Chu_phv
Ring_Of_Fire_phv
Rooftops_phv
Sultans_Of_Swing_phv
Under_Pressure_phv

Tabla 5.1: Base de datos del ISMIR 2015

5.1.2.2 Métricas utilizas

Las métricas que se han tenido en consideración para obtener los resultados de esta base de datos son, la tasa de exactitud (ACC), la tasa de éxito (HR) y la tasa de falsas alarmas (FA). Estas medidas están dadas por:

$$ACC = \frac{N_v^{corr} + N_u^{corr}}{N_v + N_u} \times 100 \tag{166}$$

$$HR = \frac{N_v^{corr}}{N_v} \times 100 \tag{167}$$

$$FA = \frac{N_u^{corr}}{N_u} \times 100 \tag{168}$$

donde N_v y N_u corresponden con el número total de tramas de voz y de tramas musicales, respectivamente. N_v^{corr} corresponden con el número de tramas de voz y de tramas musicales clasificadas correctamente. N_u^{corr} corresponde con el número de tramas de música clasificadas erróneamente.

Esencialmente, ACC representa el porcentaje de tramas de voz y de música detectado correctamente. HR representa el porcentaje de tramas de voz detectado de forma correcta. FA representa el porcentaje de tramas de música clasificadas erróneamente como tramas de voz.

5.1.2.3 Resultados

Para estos resultados se ha utilizado la base de datos perteneciente al ISMIR 2015 y debido a que la comparación es con el segmentador manual (cuyos resultados son óptimos) únicamente se mostrará la tabla de los resultados obtenidos por el segmentador automático del TFM.

Nombre de la canción	ACC	HR	FA
Hollywood_Nights_phv	0.9645	0.9605	0.0340
Hotel_California_phv	0.8993	0.8914	0.1004
Hurts_So_Good_phv	0.9210	0.8412	0.0474
La_Bamba_phv	0.8755	0.8647	0.1475
Make_It_Wit_Chu_phv	0.9105	0.9004	0.0027
Ring_Of_Fire_phv	0.8754	0.8459	0.1047
Rooftops_phv	0.8214	0.8154	0.1742
Sultans_Of_Swing_phv	0.9145	0.9104	0.0047
Under_Pressure_phv	0.9314	0.9247	0.0094
VALOR MEDIO	**0.9015**	**0.8838**	**0.0694**

Tabla 5.2: Medidas objetivas obtenidas para el segmentador automático del TFM.

Como se puede ver en la tabla anterior los resultados son satisfactorios, ya que por un lado la tasa de exactitud presenta un valor medio elevado (90.15%), la tasa de éxito también presenta un valor medio elevado (88.38%) y en cambio el valor medio de la tasa de faltas alarmas (6.94%) presenta un valor reducido.

Se han conseguido unos resultados que permiten demostrar que el funcionamiento del sistema de segmentación consigue superar os objetivos previamente planteados para este Trabajo Fin de Master.

5.1.3 Segmentador automático TFM vs Segmentador automático TFG

5.1.3.1 Base de datos

Para este análisis se ha utilizado una base de datos que se utilizó también en el TFG con el objetivo de poder compararnos con el trabajo realizado anteriormente y analizarlo. Además, nos permitirá compararnos con un segmentador automático presentado por Justin y Emilia Gómez [29]. Esta base de datos se obtuvo del MIREX 2010-2011, un concurso anual en el cual diferentes participantes de todo el país presentan sus algoritmos MIR y son evaluados partiendo de las mismas bases de datos, con el fin de comparar el rendimiento de cada uno de los modelos presentados. En MIREX 2010-2011 fueron utilizadas cuatro bases de datos, de las cuales solo ADC2004 fue distribuida públicamente para estar a disposición de quien la quisiera. Esta base de datos se compone

de 20 pistas (con antigüedad de unos 20 años aproximadamente) variando entre los géneros de pop, jazz y opera. Incluye grabaciones reales, canto sintetizado y audio generado a partir de archivos MIDI. Con un tiempo total de juego de unos 369 segundos.

Etiqueta	Canción
B01	**daisy1**
B02	**daisy2**
B03	**daisy3**
B04	**daisy4**
B05	**jazz1**
B06	**jazz2**
B07	**jazz3**
B08	**jazz4**
B09	**midi1**
B10	**midi2**
B11	**midi3**
B12	**midi4**
B13	**opera_fem2**
B14	**opera_fem3**
B15	**opera_fem4**
B16	**opera_fem5**
B17	**pop1**
B18	**pop2**
B19	**pop3**
B20	**pop4**

Tabla 5.3: Pistas pertenecientes a la base de datos ADC2005

5.1.3.2 Métricas utilizas

Para esta base de datos había que tener en cuenta que el trabajo [29] también implementaba un transcriptor de la línea melódica (fuente vocal) por lo tanto las medidas objetivas utilizadas en este trabajo varían un poco a las mencionadas anteriormente para comparar el sistema automático con el manual. Se han tomado en consideración por lo tanto las mismas métricas que se utilizó en el TFG, para que de esta forma podamos ver el progreso llevado a cabo y los distintos resultados entre el TFM, el TFG y el sistema presentado por Justin y Emilia Gómez.

Las métricas que se utilizaron en el concurso MIREX 2010-2011 correspondientes a la evaluación de los resultados del segmentador automático son:

- **Voicing Recall Rate:** la proporción de tramas que son etiquetadas como voz por el algoritmo y que realmente se consideran tramas de voz.
- **Voicing False Alarm Rate:** la proporción de tramas que son etiquetadas como voz por el algoritmo y que realmente se consideran tramas de música.

$$Voicing\ Recall\ Rate = \frac{N_v^{corr}}{N_v} \times 100 \tag{169}$$

$$Voicing\ False\ Alarm\ Rate = \frac{N_v^{err}}{N_v} \times 100 \quad (170)$$

donde N_v corresponde con el número total de tramas de voz. N_v^{corr} corresponde con el número de tramas de voz clasificadas correctamente. N_v^{err} corresponde con el número de tramas de voz clasificadas erróneamente.

5.1.3.3 Resultados

En el siguiente apartado se muestran por un lado los resultados de las medidas utilizadas para evaluar la base de datos ADC2004 proporcionadas por el segmentador automático implementado en el TFM. Por otro lado, se realiza una comparación de los resultados obtenidos con uno de los segmentadores automáticos presentado por Justin Salomon y Emilia Gómez. Por otro lado, se llevará a cabo una comparación con los resultados obtenidos en el TFG. Por lo tanto, se llevará a cabo una visión general de la calidad del segmentador que se ha implementado en el TFM, en comparación con otros sistemas.

En la siguiente tabla podemos observar los resultados de las medidas objetivas obtenidas del segmentador automático implementado en el TFM para la base de datos ADC2004.

Etiqueta		Voicing Recall Rate	Voicing False Alarm Rate
B01		0.8451	0.1947
B01	DAISY	0.8742	0.0267
B03		0.8014	0.0911
B04		0.7994	0.1780
B05		0.7842	0.1847
B06	JAZZ	0.7914	0.1473
B07		0.7454	0.1749
B08		0.8147	0.0748
B09		0.7587	0.1993
B10	MIDI	0.7487	0.1774
B11		0.8994	0.0140
B12		0.7844	0.0997
B13		0.9410	0.0040
B14	OPERA	0.9112	0.0082
B15		0.9004	0.0067
B16		0.8994	0.0091
B17		0.8441	0.0423
B18	POP	0.8957	0.0574
B19		0.8047	0.1664
B20		0.8982	0.1966
VALOR MEDIO		**0.8371**	**0.1027**

Tabla 5.4: Resultados obtenidos por el segmentador automático (TFM) para la base de datos ADC2004.

A continuación, se presentan los resultados obtenidos por el sistema de segmentación implementado en mi TFG (basado en MFCC, ver Tabla 5.5) y los resultados obtenidos por el sistema de segmentación de Salomon y Emilia Gómez (ver Tabla 5.6).

Etiqueta	*Voicing Recall Rate*	*Voicing False Alarm Rate*
B01	0.7584	0.1546
B02	0.7245	0.1845
B03	0.6942	0.1988
B04	0.7147	0.2345
B05	0.5842	0.2911
B06	0.6412	0.2455
B07	0.5845	0.2874
B08	0.6485	0.2347
B09	0.5842	0.2745
B10	0.6411	0.2133
B11	0.6541	0.2745
B12	0.6875	0.1944
B13	0.8011	0.1784
B14	0.8410	0.1697
B15	0.8123	0.1999
B16	0.7842	0.1148
B17	0.7421	0.1488
B18	0.7514	0.0999
B19	0.7189	0.1221
B20	0.7346	0.0914
VALOR MEDIO	**0.7019**	**0.2099**

DAISY

JAZZ

MIDI

OPERA

POP

Tabla 5.5: Resultados obtenidos por el segmentador automático (TFG) para la base de datos ADC2004, tabla extraída del TFG.

Collection	Voicing Recall	Voicing False Alarm	Raw Pitch	Raw Chroma	Overall Accuracy
ADC2004	0.83	0.11	0.79	0.81	0.76
	0.84	0.05	0.84	0.84	0.84
MIREX05	0.88	0.23	0.83	0.84	0.78
	0.86	0.07	0.84	0.85	0.86
MIREX09	0.87	0.24	0.81	0.84	0.76
	0.86	0.14	0.85	0.85	0.83

Tabla 5.6: Resultados obtenidos por Justin Salomon y Emilia Gómez en el concurso ADC2004. Tabla extraída de [29].

Como se puede observar en la Tabla 5.4 el sistema de segmentación automática desarrollado en el TFM tiene unos resultados bastantes eficaces, por lo que consigue llevar a cabo una correcta segmentación. Por otro lado, los resultados que se obtuvieron para el TFG (Tabla 5.5) tienen peores expectativas y esto es debido a que el segmentador desarrollado en el TFG se optimizo para el género correspondiente al Flamenco. Además,

la técnica que utiliza (MFCC) se comporta muy bien para este estilo de música y no para estilos donde la complejidad musical aumenta. Por lo tanto, si realizamos una comparación entre los resultados obtenidos tanto en el TFG (Tabla 5.5), como por Justin y Emilia (Tabla 5.6), podemos observar que hemos conseguido llevar a cabo un sistema cuyos resultados hablan por sí solos. De esta forma afirmamos que nuestro segmentador va a conseguir llevar a cabo una segmentación entre zonas vocales e instrumentales efectiva para que en la consiguiente fase de separación los resultados sean mejores. Además, aquí se hace un matiz de porque surgió la idea de llevar a cabo este Trabajo fin de Master, la cual consistía en llevar a cabo un segmentador automático y un separador que no se especializase en ningún género musical como le ocurrió al sistema del TFG.

Por otro lado, y si observamos los resultados obtenidos por el TFM, llegamos a la conclusión de que es más sencillo determinar las zonas instrumentales para las canciones pertenecientes al género de la ópera, y esto se debe principalmente en que en este tipo de canciones cuando el cantante comienza a cantar lo hace con una intensidad mucho mayor en comparación a la intensidad presente en las zonas instrumentales. Es por ello que los mejores resultados en términos de “Voicing Recall Rate” y “Voicing False Alarm Rate” se consigan para las canciones pertenecientes al género de la ópera y etiquetadas en la base de datos como (B13, B14, B15 y B16).

Llegados a este punto nos podemos preguntar, el porqué de que nuestro sistema falle en las zonas en las que lo hace. Realizando un estudio exhaustivo en la que se realizaron pruebas con una gran variedad de bases de datos se llegó a la conclusión de que el sistema se comporta peor cuando existe la presencia de un instrumento tonal con mucha energía. Esto es lógico, ya que en este caso el sistema de segmentación en su primera fase (preprocesado NMF para obtener el espectrograma de la fuente vocal), puede determinar que parte de este instrumento tonal se considera fuente vocal, cuando realmente es un acompañamiento. Esto es lo que ocurre, por ejemplo, en el peor de los casos B07 (ver tabla 5.4), donde está presente un instrumento de viento.

5.1.4 Segmentador automático TFM vs Segmentadores automáticos investigadores

5.1.4.1 Base de datos

En este caso la base de datos que se va a utilizar corresponde con la base de datos "Jamendo database" que fue creada y publicada para el experimento descrito en [70]. Esta consiste en una colección de 93 canciones de aproximadamente una duración de 6 horas en total. Las canciones corresponden a diferentes géneros musicales y pertenecen a diferentes artistas. De esta base de datos se han seleccionado 16 canciones de 30 segundos de duración para obtener los resultados. Se han seleccionado las mismas que la utilizadas en el segmentador [31] definido por Vembu y Baumann, las utilizadas por el segmentador de Ramona [30] y las utilizadas por Bernhard Lehner [32] para poder realizar una comparación de mi sistema de segmentación con ellos.

NOMBRE DE LA CANCIÓN
Say me Good Bye
School
Si Dieu
Une charogne
castaway
Believe
Healing Luna
Inside
You are
LIrlandaise
16 ans
2003-Circons
A Poings Fermes
Crepuscule
Dance
Elles disent

Tabla 5.7: Selección de la base de datos Jamendo.

5.1.4.2 Métricas utilizas

La métrica que se ha tenido en consideración para obtener los resultados de esta base de datos es la tasa de exactitud (ACC) que representa el porcentaje de tramas de voz y de música detectado correctamente.

$$ACC = \frac{N_v^{corr} + N_u^{corr}}{N_v + N_u} \times 100 \tag{171}$$

donde N_v y N_u corresponden con el número total de tramas de voz y de tramas musicales, respectivamente. N_v^{corr} corresponden con el número de tramas de voz y de tramas musicales clasificadas correctamente. N_u^{corr} corresponde con el número de tramas de música clasificadas erróneamente.

5.1.4.3 Resultados

En la siguiente tabla podemos observar los resultados de las medidas objetivas obtenidas del segmentador automático implementado en el TFM para la base de datos "Jamendo".

Nombre de la canción	Accuracy %
Say me Good Bye	89.24
School	86.14
Si Dieu	90.47
Une charogne	87.14
castaway	80.74
Believe	87.09
Healing Luna	84.17
Inside	82.79
You are	82.74
LIrlandaise	83.94
16 ans	81.72
2003-Circons[. . .]	80.66
A Poings Fermes	90.47
Crepuscule	86.10
Dance	90.74
Elles disent	84.25
VALOR MEDIO	**85.5250**

Tabla 5.8: Resultados obtenidos por el segmentador automático (TFM) para la base de datos JAMENDO

En la siguiente tabla se muestran los resultados obtenidos por Ramona [30] ("Ramona et al"), los resultados obtenidos por Bernhard Lehner [32] ("PROP") y el sistema desarrollado por Vembu and Baumann [31] ("VB"). Únicamente se ha definido la medida del Accuracy, porque el f-measure (F%) se utiliza en mayor medida para evaluar los transcriptores.

	Ramona et al.		PROP		VB	
	Acc%	F%	Acc%	F%	Acc%	F%
03 - Say me Good Bye	80.1	85.8	91.4	83.6	90.6	82.7
03 - School	84.3	87.3	84.8	86.6	71.5	77.8
03 - Si Dieu	76.4	80.7	87.2	89.4	76.2	66.7
03 - Une charogne	85.3	91.7	89.8	93.5	78.8	85.9
03 - castaway	79.0	87.3	71.3	80.5	73.0	80.0
04 - Believe	80.0	88.5	94.1	95.6	83.0	87.2
04 - Healing Luna	85.5	81.6	87.8	84.4	72.7	70.8
04 - Inside	83.3	68.2	79.4	66.0	75.3	58.0
04 - You are	87.0	91.9	87.9	90.6	74.4	77.7
05 - 05 LIrlandaise	57.7	64.2	65.0	68.6	61.7	60.5
05 - 16 ans	91.5	84.8	87.3	79.8	70.8	60.3
05 - 2003-Circons[...]	87.6	88.2	75.5	77.7	79.8	79.6
05 - A Poings Fermes	93.7	92.2	89.7	83.0	86.9	81.2
05 - Crepuscule	85.2	88.8	80.1	83.6	76.8	80.0
05 - Dance	77.0	83.2	84.1	88.7	75.7	82.2
05 - Elles disent	71.8	78.7	84.4	87.4	69.6	77.0
ALL	82.2	84.3	84.8	84.6	77.4	76.9

Tabla 5.9: Resultados del modelo propuesto por Bernhard Lehner en [32], comparado con el definido por Ramona en [19] y el modelo de Vembu and Baumann [31].

Como se puede ver haciendo un análisis de los resultados mostrados anteriormente el sistema de segmentación propuesto en el Trabajo Fin de Master consigue mejorar los resultados de segmentación propuestos por los anteriores investigadores. Aunque en este análisis de los resultados únicamente nos basamos en la métrica del Accuracy, es medida necesaria para asegurar que la mayoría de las zonas detectadas como zonas instrumentales corresponden con las originales.

El hecho de tomar distintas bases de datos a la hora de comparar hace que el hecho de afirmar que el segmentador funciona de forma correcta quede más respaldado. Es por ello la utilización de tres bases de datos distintas.

5.2 Separador de señales musicales

En esta sección se llevarán a cabo la implementación de los distintos resultados objetivos para evaluar la etapa de separador basado en modelos NMF. Es importante tener en cuenta que en esta sección de la memoria se realizan dos tipos de análisis:

- **Análisis armónico/percusivo,** donde únicamente entran en juego las fuentes armónica y percusiva, el cual se lleva a cabo para evaluar el sistema de separación por sí solo. Ya que no depende del segmentador al no necesitar una segmentación entre zonas vocales e instrumentales (todo es zona instrumental). Se compararán resultados entre el sistema óptimo y el sistema previo a la implementación del mismo.
- **Análisis armónico/percusivo/vocal,** donde ya entran en juego las tres fuentes sonoras. Los resultados obtenidos los compararemos con los obtenidos en el TFG y los obtenidos por Durrieu [64].

Para no tener que incluir un apartado de "Setup" por cada una de los análisis llevados a cabo (ya que el sistema óptimo que se compara en todos los casos es el mismo) se introducirá un único punto que lo defina en este apartado. De igual forma para no tener que incluir un apartado de métrica para cada análisis se incluirá únicamente uno en este apartado, ya que todos los análisis realizados de separación van a utilizar la misma métrica.

5.2.1 Setup

Finalmente, para considerar que el sistema de separación funciona de forma óptima se han tenido que tomar las siguientes consideraciones. En este caso diferenciaremos entre los parámetros tomados para el primer modelo de separación NMF armónico/percusivo, que permite obtener las bases entrenadas de las zonas instrumentales y el segundo modelo de separación NMF armónico/percusivo/vocal que permite realizar una separación de las tres fuentes presentes.

- **Modelo de separación NMF armónico/percusivo:** para este modelo se ha tenido en cuenta que la matriz de bases armónicas W_H se inicialice con un matiz de bases armónicas entrenadas pertenecientes al piano. El resto de matrices se inicializan aleatoriamente. Por otro lado, se ha dado mayor peso a la constante de restricción de la matriz de activaciones armónicas (smothness=0.6), mientras que los demás pesos se han mantenido con un peso de 0.1. Por otro lado, se ha utilizado un valor de β=1.3, el número de bases armónicas es 200 y el número de bases percusivas es 100, debido a que de esta forma se mejoran los resultados de separación (existe más posibilidad de tocar distintas notas armónicas que percusivas).

- **Modelo de separación NMF armónico/percusivo/vocal:** para este modelo se ha tenido en cuenta que la matriz de bases armónicas W_H se inicialice con un matiz de bases armónicas entrenadas pertenecientes al piano. Por otro lado, la matriz correspondiente a los filtros de fonema se inicializará con otra previamente entrenada. El resto de matrices se inicializan como se indica en el desarrollo del segundo modelo de la separación del TFM. Por otro lado, se ha dado mayor peso a la constante de restricción de las matrices de activaciones vocales (single activation=0.3), mientras que los demás pesos se han mantenido con un peso de 0.1. Por otro lado, se ha utilizado un valor de β=1.3, el número de bases armónicas es 200, el número de bases percusivas es 100 y el número de bases correspondientes a la parte vocal es 63 debido a que de esta forma se mejoran los resultados de separación.

5.2.2 *Métricas utilizadas*

Las medidas que se han tomado en consideración para evaluar el rendimiento del algoritmo desarrollado son las siguientes:

- **SDR (Source-to-distorsion ratio):** Proporciona información de la calidad de separación del algoritmo.

$$SDR = 10\log_{10}\frac{||s_{true}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{172}$$

- **SIR (Source-to-interferences ratio):** Proporciona una medida de la presencia de sonidos armónicos en la señal percusiva y viceversa.

$$SIR = 10\log_{10}\frac{||s_{true}||^2}{||e_{interf}||^2} \tag{173}$$

- **SAR (Source-to-artifacts ratio):** Proporciona información de los artefactos en la señal separada debido al proceso de separación y de reconstrucción.

$$SAR = 10\log_{10}\frac{||e_{interf} + e_{noise} + e_{artif}||^2}{||e_{artif}||^2} \tag{174}$$

donde, s_{true} es una matriz que contiene la imagen de la fuente original, e_{noise} es una matriz que contiene la componente espacial de distorsión, e_{interf} es una matriz que contiene la componente de interferencia y e_{artif} es una matriz que contiene la componente de artefactos.

5.2.3 Separador armónico/percusivo óptimo vs Separador armónico/percusivo sin optimizar

5.2.3.1 Base de datos

Para el análisis de estos resultados se utilizará la misma base de datos utilizada en la comparación del sistema automático de segmentación del TFM y el sistema manual. Más específicamente se utilizará la base de datos perteneciente al ISMIR 2015, cuyo contenido se puede ver en la tabla 5.1. Es necesario tener en cuenta que esta base de datos está formada por canciones que únicamente tienen presente la fuente armónica y la fuente percusiva. Esta es la única diferencia que existe con la base de datos mencionada anteriormente y que también incluye la fuente vocal

5.2.3.2 Resultados

En primer lugar, se van a mostrar los resultados obtenidos por el sistema de separación armónico/percusivo sin haber optimizado el sistema. En la siguiente tabla se podrán observar los resultados obtenidos para cada canción.

Nombre de la canción	SDR	SIR	SAR
Hollywood_Nights_ph	[0.14; -0.51]	[1.77; -1.47]	[-1.44; -10.55
Hotel_California_ph	[1.47; 2.11]	[6.11; 5.22]	[-6.14; -18.41]
Hurts_So_Good_ph	[0.81; -1.64]	[5.24; 1.11]	[-5.52; -1.51]
La_Bamba_ph	[0.97; -0.55]	[-1.15; 6.21]	[9.77; 3.99]
Make_It_Wit_Chu_ph	[1.88; 2.41]	[6.14; -1.22]	[2.69; 3.17]
Ring_Of_Fire_ph	[2.11; 0.99]	[10.56; 9.52]	[4.31; 8.98]
Rooftops_ph	[1.80; 2.10]	[-2.14; 6.14]	[7.11; 2.50]
Sultans_Of_Swing_ph	[-2.14; -1.72]	[6.88; 13.74]	[-2.45; -20.82]
Under_Pressure_ph	[1.47; 2.05]	[2.34; 20.51]	[10.59; 1.75]
VALOR MEDIO	[0.9456; 0.5822]	[3.9722; 6.6400]	[2.1022; -3.4333]

Tabla 5.10: Resultados obtenidos por el separador NMF armónico/percusivo sin optimizar.

Una vez que se han mostrado los resultados obtenidos por el separador armónico/percusivo sin optimizar en la siguiente tabla podemos observar los resultados obtenidos para la misma base de datos con el sistema NMF armónico/percusivo optimizado.

Nombre de la canción	SDR	SIR	SAR
Hollywood_Nights_ph	[2.33; 0.41]	[3.14; 12.77]	[7.12; 6.56]
Hotel_California_ph	[4.84; 11.43]	[14.9; 16.93]	[10.72; 8.60]
Hurts_So_Good_ph	[7.06; 4.07]	[12.22; 11.73]	[9.21; 8.52]
La_Bamba_ph	[9.14; 4.62]	[16.24; 10.99]	[10.74; 7.18]
Make_It_Wit_Chu_ph	[7.39; 3.96]	[13.23; 13.87]	[8.23; 9.84]
Ring_Of_Fire_ph	[5.57; 4.29]	[15.58; 13.14]	[10.42; 6.78]
Rooftops_ph	[3.52; 8.72]	[8.79; 10.34]	[8.04; 6.29]
Sultans_Of_Swing_ph	[5.47; 0.44]	[13.43; 6.55]	[6.27; 7.49]
Under_Pressure_ph	[6.60; 8.78]	[13.18; 14.65]	[11.04; 5.23]
VALOR MEDIO	[5.7689; 5.1911]	[12.3011; 12.3300]	[9.0878; 7.3878]

Tabla 5.11: Resultados obtenidos por el separador NMF armónico/percusivo optimizado.

Si comparamos los resultados obtenidos por el sistema sin optimizar y los resultados obtenidos por el sistema óptimo nos damos cuenta que conseguimos una notable mejora en los resultados. Esto se debe a que en el modelo de separación armónico/percusivo inicializamos las bases armónicas con una matriz de bases armónicas previamente entrenada que corresponden al piano y por lo tanto estamos añadiendo información adicional al sistema que facilita la obtención de mejores resultados.

Como la introducción a este tipo de medidas puede ser un poco confusa, hasta tal punto que no se entienda que resultados de SDR, SIR y SAR son mejores que otros, se ha realizado un experimento comparando dos canciones que son exactamente la original, es decir, hemos realizado estas medidas suponiendo que el separador nos ha devuelto las fuentes sonoras exactamente iguales a la original. Como se puede ver en la siguiente tabla, las medidas de SDR, SIR y SAR son mejores conforme aumenta su valor, ya que si suponemos que las fuentes que obtiene el separador son las mismas que con las que se comparan, los resultados son máximos.

Canción	SDR	SIR	SAR
Hollywood_Nights_ph	[Inf; Inf]	[218.19; 196.95]	[218.82; 196.94]

Tabla 5.12: Valores óptimos de las medidas objetivas SDR, SIR y SAR comparando la misma canción y suponiendo de esta forma que la canción obtenida por el separador es exactamente la misma que la que se grabaría con un micrófono independiente.

5.2.4 Separador armónico/percusivo/vocal TFM vs Separador armónico/percusivo/vocal TFG vs Separador armónico/percusivo/vocal de Durrieu

Es importante destacar que en este análisis vamos a compararnos tanto con los resultados obtenidos en mi TFG, como con los resultados obtenidos por Durrieu, partiendo de la misma base de datos, para conseguir una comparación exacta.

5.2.4.1 Base de datos

La base de datos utilizada para la evaluación de los resultados en la etapa de separación de señales musicales se obtiene del trabajo realizado por Jean-Louis Durrieu [64], ya que al fin y al cabo los resultados obtenidos por nuestro sistema de separación basado en NMF se van a comparar con los resultados obtenidos por Durrieu y por los obtenidos en mi TFG (donde se realizó igualmente una comparación con Durrieu).

La base de datos utilizada se compone de un total de 5 canciones de unos 20 segundos de duración, donde puede encontrar música perteneciente al pop, al jazz y al rap.

En la Tabla siguiente se puede observar la base de datos que hemos utilizado para medir los resultados de nuestro sistema separador.

NOMBRE DE LA CANCIÓN
BEARLIN
TAMY
ANOTHER DREAMER
4 FORT MINOR
ULTIMATE NZ TOUR

Tabla 5.13: Base de datos utilizada para el análisis de los resultados del sistema de separación NMF armónico/percusivo/vocal.

5.2.4.2 Resultados

En este apartado se lleva a cabo un análisis de los resultados obtenidos por el separador NMF armónico/percusivo/vocal diseñado en el TFM para la base de datos descrita anteriormente y así poder realizar una comparación entre los resultados que se obtuvieron en la realización del TFG y los resultados obtenidos por Durrieu. De esta forma mediremos el rendimiento del sistema de separación NMF implementado en función de resultados obtenidos por un sistema propio anterior y los obtenidos por uno de los grandes investigadores en el campo de la separación de fuentes sonoras.

Así pues, las medidas obtenidas se presentan en la siguiente tabla. La unidad de estos datos es el decibelio dB. Por un lado, se ha comparado la canción solo vocal original y la canción solo vocal estimada, columna izquierda. Por otro lado, se ha comparado la

canción solo instrumental original y la canción solo instrumental estimada, columna derecha.

En la siguiente tabla se muestran los resultados obtenidos por el sistema propuesto en este Trabajo Fin de Master.

Nombre de la canción	SDR	SIR	SAR
BEARLIN	[7.195, 6.451]	[12.82, 23.78]	[10.52, 13.55]
TAMY	[6.482, 5.880]	[15.74, 19.47]	[13.77, 8.50]
ANOTHER DREAMER	[5.952, 4.612]	[10.70, 25.64]	[10.23, 6.43]
4 FORT MINOR	[5.990, 4.420]	[15.11, 10.88]	[5.47, 3.69]
ULTIMATE NZ TOUR	[7.463, 3.781]	[18.66, 9.72]	[15.44, 7.91]
VALOR MEDIO	**[6.1820, 5.4632]**	**[14.5060, 17.9980]**	**[9.6760, 9.4260]**

Tabla 5.14: Resultados que se obtuvieron por el separador NMF armónico/percusivo/vocal implementado en el TFM.

En la siguiente tabla se muestran los resultados que se obtuvieron en el TFG, en el que se realizó un sistema NMF de separación distinto al descrito en este TFM.

Nombre de la canción	SDR	SIR	SAR
BEARLIN	[4.415, 8.124]	[6.78, 25.47]	[4.83, 9.33]
TAMY	[3.122, 5.142]	[13.19, 32.14]	[4.02, 4.27]
ANOTHER DREAMER	[3.524, 7.613]	[6.31, 28.88]	[7.06, 8.33]
4 FORT MINOR	[2.144, 0.144]	[3.03, -1.79]	[-0.24, -21.74]
ULTIMATE NZ TOUR	[2.499, 2.147]	[19.45, 12.11]	[4.42, 6.47]
VALOR MEDIO	**[3.14, 4.63]**	**[9.75, 19.362]**	**[4.01, 1.332]**

Tabla 5.15: Resultados obtenidos por el separador NMF armónico/percusivo/vocal propuesto en el TFG.

. En la siguiente tabla se pueden ver los resultados que ha obtenido Durrieu con sus diferentes sistemas, hasta haber llegado al óptimo.

Song	#1	#2	#3	#4	#5
Original SIR	-5.3	0.2	-3.0	-7.2	-7.5
Wiener	10.9	13.7	11.5	11.4	10.1
VIMM - Oracle	8.5	12.4	8.4	7.4	6.4
VUIMM - Oracle	8.6	12.4	8.8	6.8	6.2
VIMM - Melody	6.9	11.3	5.7	4.2	4.9
VUIMM - Melody	7.5	11.8	6.6	4.5	5.5
VIMM - Pres.	5.9	10.5	5.3	2.6	4.4
VUIMM - Pres.	6.7	10.7	6.1	2.7	4.8
VIMM	6.0	11.3	5.3	3.3	4.3
VUIMM	6.5	11.6	5.8	3.3	4.9
Best SiSEC2010	x	x	3.1	3.9	2.6

Tabla 5.16: Resultados obtenidos por Durrieu dependiendo del sistema utilizado.

En primer lugar, se puede observar en la Tabla 5.14 (donde se muestran los resultados obtenidos por el sistema de separación desarrolla en el TFM), como la separación, tanto vocal como instrumental es positiva en términos de la calidad de la separación, ya que al fin y al cabo los resultados obtenidos del SDR son positivos. Dependiendo de la canción el sistema funciona con una calidad mejor o una calidad peor, pero en este caso el sistema se comporta eficazmente para todas las canciones.

Para la estimación instrumental y vocal, podemos comprobar como la mayoría de las pistas vienen determinadas por interferencia, es decir, se ha realizado una buena separación. Sin embargo, también se ha dado algún caso en el que la separación viene determinada por artefacto, es decir, se ha realizado una mala separación debido, principalmente, a que no se ha podido obtener una cantidad suficiente de información instrumental o vocal que ayude a realizar el proceso de separación correctamente.

Si realizamos una comparación entre los resultados medios obtenidos entre el modelo desarrollado en el TFM (Tabla 5.14), el desarrollado en el TFG (Tabla 5.15), y el mejor de los algoritmos diseñados por Durrieu para esta base de datos (Tabla 5.16), podemos observar que los resultados obtenidos en este caso por el modelo del TFM superan con creces a los obtenidos por el modelo desarrollado en el TFG, sin embargo, no conseguimos alcanzar los resultados que obtiene en este caso Durrieu (Tabla 5.17). Es necesario tener en cuenta que Durrieu para llegar a su algoritmo óptimo ha tenido que llevar a cabo otra serie de algoritmos como se muestra en la tabla 26, hasta alcanzar sus mejores resultados.

Visto de esta forma en nuestro caso hemos conseguido mejorar los resultados obtenidos en el TFG de forma considerable. Por lo que siguiendo el camino que siguió Durrieu conseguiremos alcanzarlo con futuros modelos de separación armónicos/percusivos/vocales.

Nombre de la canción	Sistema TFG	Sistema TFM	Durrieu
BEARLIN	4.415	8.195	10.9
TAMY	3.122	7.482	13.7
ANOTHER DREAMER	3.524	6.952	11.5
4 FORT MINOR	2.144	6.990	11.4
ULTIMATE NZ TOUR	2.49	8.463	10.1

Tabla 5.17: Comparación entre el separador del sistema implementado en el TFG, el sistema implementado en el TFM y el sistema implementado por Durrieu. Los resultados son el valor de SDR parte vocal en dB.

Dando una valoración subjetiva de los resultados que se han obtenido en términos de la calidad del sonido separado, podemos decir en primer lugar que la fuente vocal que es capaz de extraer el algoritmo de Durrieu es limpia, en el sentido de que no se cuela ningún instrumento paralelo a la señal vocal. Sin embargo, en los resultados subjetivos obtenidos por nuestro sistema en el TFM no se llega a conseguir una señal vocal tan pura como en el caso del sistema a de Durrieu, ya que en algunos casos se cuela parte de la señal armónica (correspondiente al bajo).

6 CONCLUSIONES

Una vez que se ha llevado a cabo la explicación de los distintos sistemas que componen el Trabajo Fin de Master y se ha llevado a cabo una valoración de los distintos resultados objetivos se procederá a la elaboración de la sección de conclusiones, en la que se dejaran claras las conclusiones más importantes que se han adquirido con este Trabajo Fin de Master.

En primer lugar, es importante enfatizar que se han conseguido los objetivos con los que se introducían este TFM. Ya que se ha conseguido implementar un sistema totalmente automático que nos permite por un lado realizar una segmentación entre las zonas instrumentales y las zonas vocales para cualquier tipo de canción (independientemente del género musical) y por otro lado realizar una separación de las distintas fuentes sonoras presentes en la señal musical (armónica, percusiva y vocal).

El atractivo de este TFM corresponde a la implementación de dos bloques de investigación claramente diferenciados (segmentación y separación) en un único sistema general. E incluso merece la pena destacar el hecho de que el sistema implementado sea totalmente automático, sin la necesidad de indicar los fragmentos de las señales musicales correspondientes a las zonas instrumentales y los correspondientes a las zonas vocales. Gracias a ello se ha conseguido implementar un sistema inteligente capaz de actuar por sus propios medios. Este sistema permite a su vez ahorrar tiempo a la hora de tratar con la separación de fuentes sonoras, por el hecho de haber implementado un bloque automático que estudie cada una de las señales musicales y sea capaz de determinar que partes corresponden con zonas instrumentales y cuales otras con zonas vocales, evitando de esta forma que este trabajo sea realizado por una persona.

Por otro lado, es importante destacar que para la realización de los distintos sistemas que componen el TFM se ha llevado a cabo una investigación exhaustiva sobre las distintas técnicas presentes hasta el momento.

A lo largo de la ejecución del TFM se han explicado tantos los sistemas óptimos implementados finalmente para la segmentación y la separación, como las distintas técnicas que se han demostrado que su funcionamiento no es efectivo para esta aplicación. Esto se debe a que no es únicamente importante tener claro cuáles son las técnicas óptimas para la realización de los distintos sistemas, sino además tener claro cuáles son

las técnicas que no se consideran óptimas para que el día de mañana otros investigadores no cometan el mismo error.

Para realizar una valoración completa del TFM implementado y de esta forma poder medir su rendimiento se han tenido que seleccionar distintas bases de datos, como se puede ver en la sección de resultados. El motivo de haber escogido diferentes bases de datos no es otro que el de conseguir comparar los resultados de las medidas objetivas obtenidas por el sistema implementado en el TFM, con los de otros investigadores dedicados a la realización de los bloques (segmentación y separación) por separado. Además, si comparamos nuestro sistema teniendo en cuenta varias bases de datos y con ello varios géneros musicales, podemos demostrar que el sistema funciona de forma efectiva para cualquier género musical. Por otro lado, las bases de datos se han escogido de los propios investigadores, por lo tanto, es importante destacar que los resultados que se presentan en los artículos publicados se especializan para esa base de datos.

A continuación, se muestran una serie de imágenes en las que claramente se pueden comparar los resultados obtenidos (tanto en segmentación, como en separación) por el sistema desarrollado en el TFM y por los sistemas desarrollados por los investigadores con los que se han comparado en la sección de resultados.

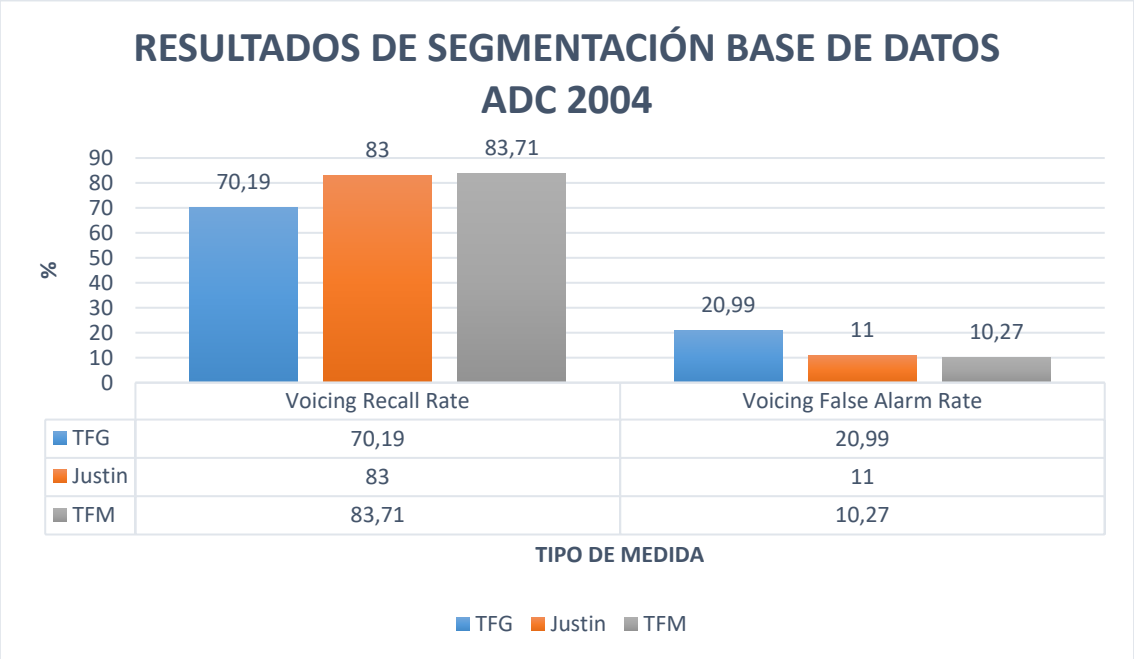

Figura 6.1: Resultados de segmentación, base de datos ADC 2004, comparación entre los resultados obtenidos en el TFG, en el TFM y por Justin y Emilia.

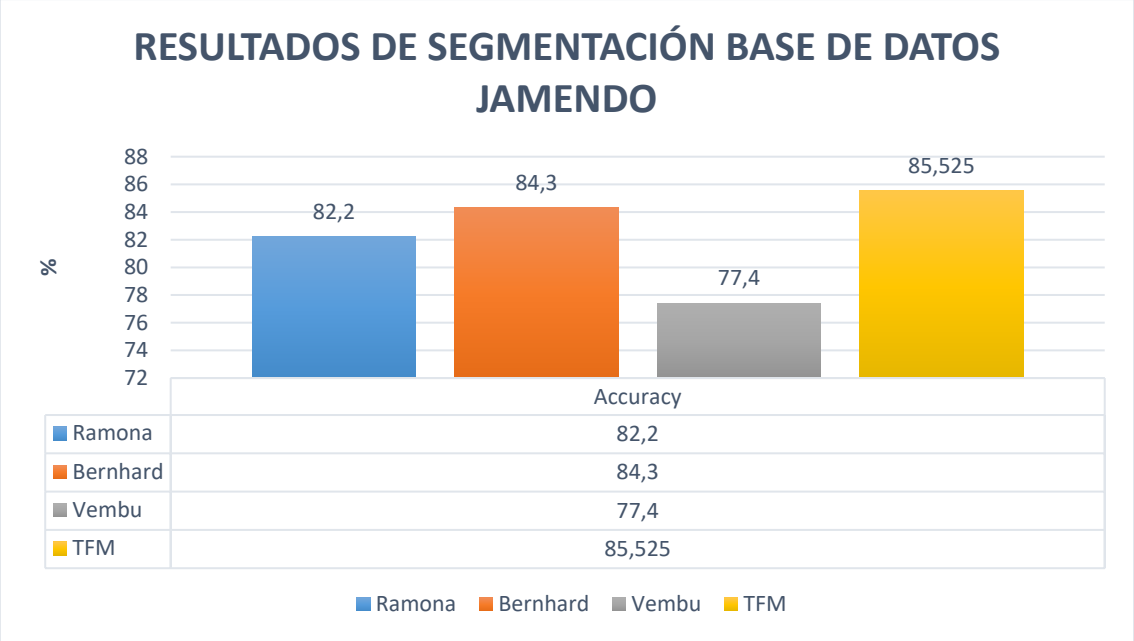

Figura 6.2: Resultados de segmentación, base de datos JAMENDO, comparación entre los resultados obtenidos por el TFM, por Ramona, Bernhand y por Vembu y Baumann.

Como se puede observar en los gráficos anteriores los resultados obtenidos por nuestro sistema en segmentación supera con creces a los que se obtuvieron en la ejecución del Trabajo Fin de Grado. Por lo tanto, se ha implementado un sistema de segmentación mejorado con respecto al del TFG. Por otro lado, los resultados de segmentación que han obtenido otros investigadores han sido superados por el segmentador que se ha realizado en este TFM, por lo tanto, la etapa de segmentación del TFM ha sido implementada de forma efectiva.

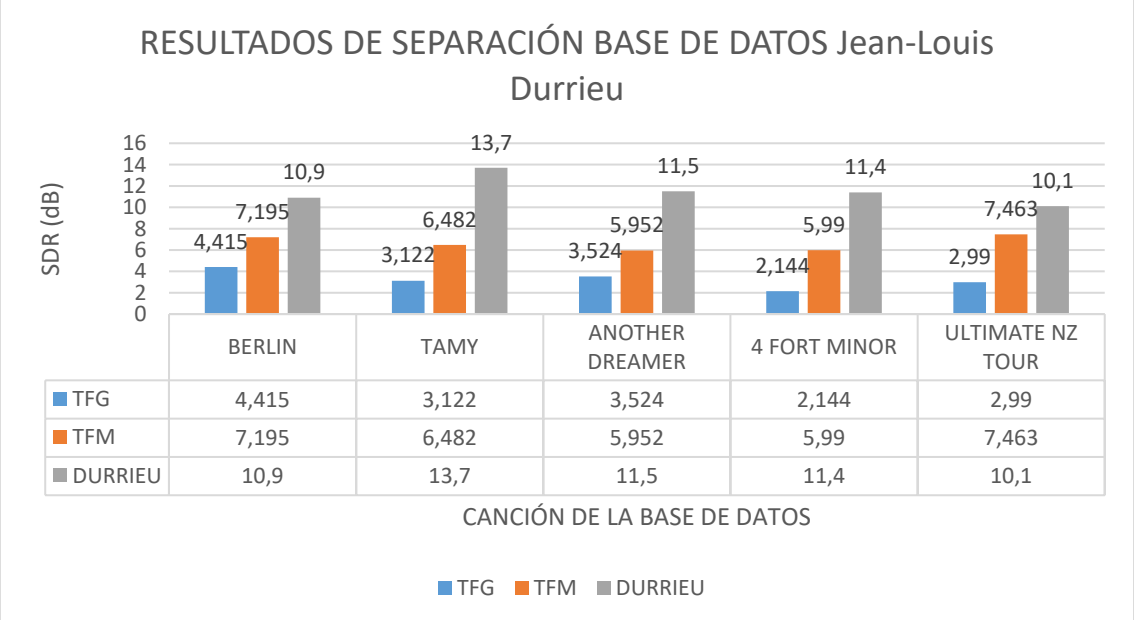

Figura 6.3: Resultados de separación, base de datos Jean-Louis Durrieu, comparación entre los resultados obtenidos por el TFM, el TFG y por el propio Durrieu.

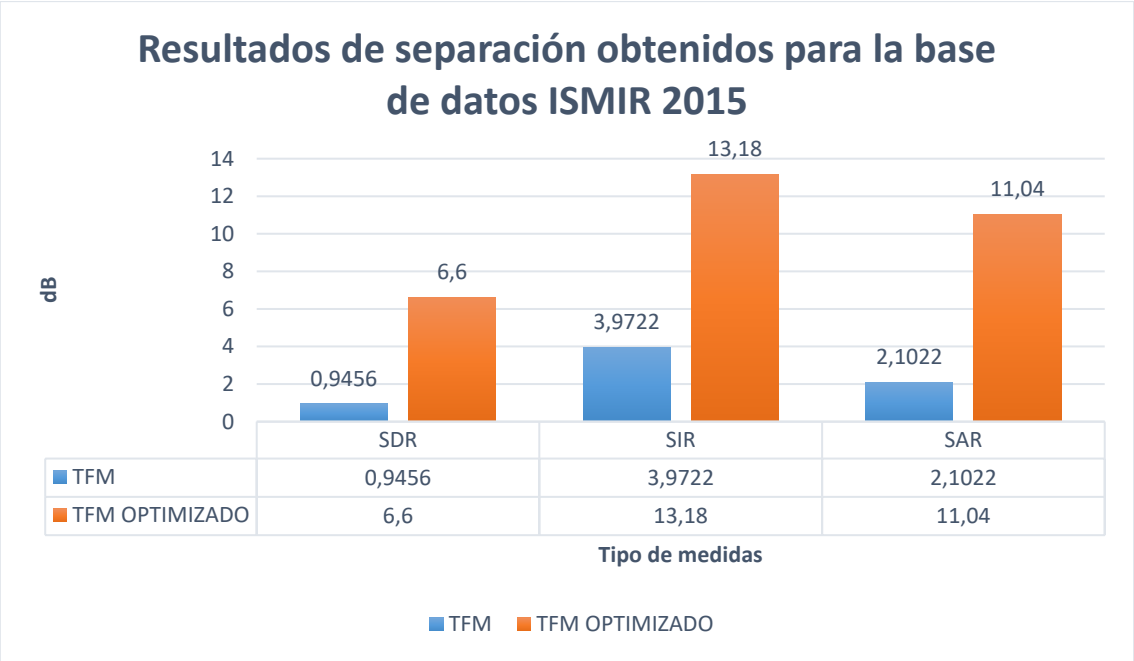

Figura 6.4: Resultados de separación, base de datos ISMIR2015, comparación entre los resultados obtenidos por el TFM sin optimizar y el optimizado.

Por otro lado, se puede observar que los resultados obtenidos por nuestro sistema de separación supera con creces los que se obtuvieron en el TFG. Por lo tanto, se ha implementado un sistema de separación más efectivo que el que se llevó a cabo en el TFG. Además, si se comparan los resultados con el de otros investigadores (Durrieu) podemos observar que no conseguimos alcanzar sus resultados, pero esto se debe a que este investigador ya ha tenido que implementar variar versiones de su algoritmo para conseguir estos resultados. Como se dijo en la sección de resultados el hecho de haber mejorado considerablemente los resultados de separación del TFM al TFG, hace que haya motivos suficientes para seguir investigando y llegar a conseguir resultados que superen a investigadores como Durrieu.

Finalmente, en la última imagen se muestra un análisis de los resultados objetivos obtenidos por el sistema de separación del TFM sin optimizar y por el optimizado, consiguiendo unos resultados que mejoran con creces al modelo inicial.

7 LINEAS DE FUTURO

Dentro de este Trabajo Fin de Master como se ha mencionado anteriormente el gran atractivo lo ocupa la parte del segmentador automático, ya que muchos investigadores se han centrado en la separación de fuentes musicales, ya sea en el ámbito de la música (para obtener las señales musicales separadas de la señal vocal), como en el ámbito de las telecomunicaciones (más específicamente en el ámbito de las llamadas de teléfono para poder separar la voz del hablante del ruido de fondo), etc. Sin embargo, la tarea de realizar un segmentador que sea capaz de determinar de forma automática cuales son los tramos temporales en los que únicamente hay base instrumental no ha sido tarea principal para los investigadores. Si bien cabe decir que de esta forma estaríamos innovando en una tecnología en la que cada vez avanzamos a un mundo en el que se plasma la inteligencia del hombre en los pc, para que de esta forma podamos ahorrar tiempo en la ejecución de las tareas. Hay quien piensa que de esta forma nos estamos dirigiendo a un mundo en el que la tecnología sustituya al ser humano, pero podríamos pensar que la tecnología motiva a la evolución.

Una de las principales líneas futuras podría consistir en realizar un karaoke de forma automática, ya que con este Trabajo Fin de Master disponemos de la teoría necesaria para quedarnos con la parte instrumental de la señal de entrada. Más que como línea de futuro es algo que se ha implementado de forma paralela al TFM. Esto se debe a que durante la realización del TFM he estado realizando una beca Talentum en la cual se buscaba implementar una idea de negocio. La idea que se llevó a cabo para esta beca fue la realización de una plataforma web que permitiese crear una base de datos elegida directamente por el usuario y obtenerla en el formato propio del karaoke, de esta forma rompemos con el mito de las canciones populares y convertimos al usuario en el dueño de lo que quiere cantar.

Por lo tanto, queda demostrado que la realización de estos algoritmos y su optimización, aunque a nivel de investigación no se vean claras las aplicaciones que se podrían llevar a cabo con ellos en la vida real, son importantes para que las aplicaciones futuras que lo utilicen funcionen de forma óptima.

8 ANEXOS

8.1 ANEXO I: MANUAL DE USUARIO

Se ha desarrolla una interfaz destinada a la ejecución de los sistemas de segmentación y separación llevados a cabo en el TFM de una forma sencilla y bastante intuitiva. La razón por la cual se ha visto necesaria la elaboración de esta interfaz no es otra que intentar plasmar mediante una sencilla aplicación la idea fundamental que buscar este TFM. Para ello se ha usado la herramienta GUIDE de Matlab, permitiendo poner al alcance de cualquier usuario el análisis de las señales musicales a tratar tanto en el sentido de la segmentación, como en el sentido de la separación.

A continuación, se muestra un manual de usuario a seguir para poder entender el funcionamiento del interfaz diseñado para la segmentación automática y la separación de fuentes sonoras. En primer lugar, se va a mostrar el entorno de trabajo usado, tal y como se aprecia en la Figura 8.1. Donde se puede observar el aspecto que tendrá la interfaz, y la identificación de las distintas secciones que la forman.

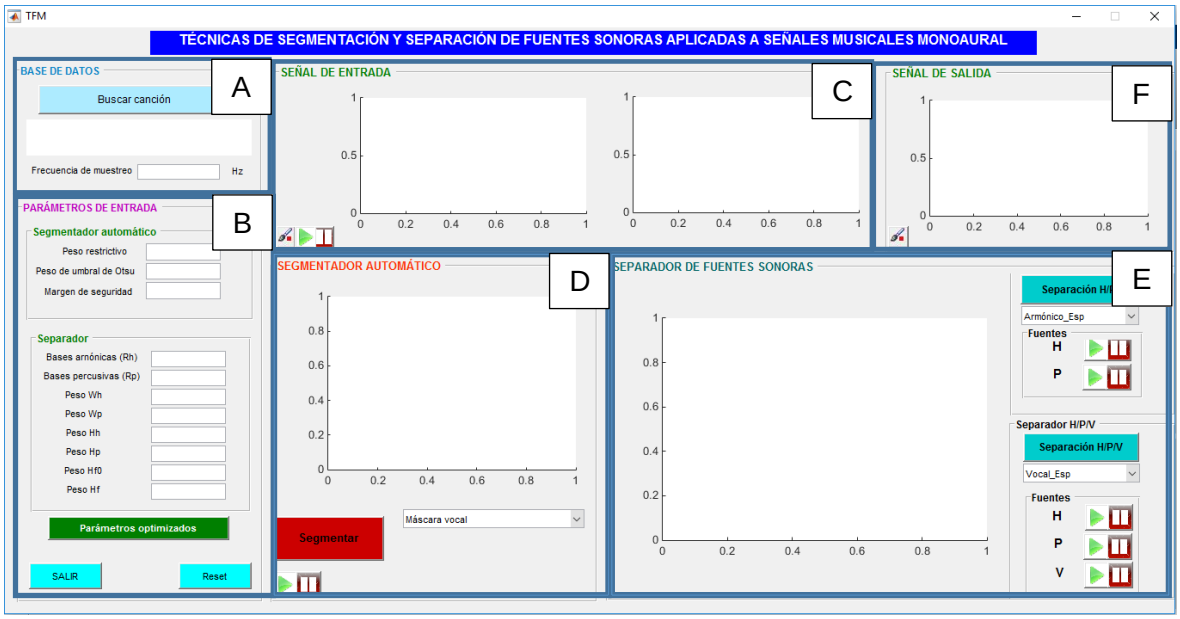

Figura 8.1: Interfaz de usuario para el segmentador automático y el separador de fuentes armónico/percusivo/vocal.

Como se puede apreciar en la figura anterior el entorno de trabajo es muy intuitivo y está formado por 6 secciones, identificadas desde la letra A hasta la letra E. Esto nos permitirá llevar a cabo un análisis completo en el tema que nos aborda (segmentación y separación de señales musicales). A continuación, se detallará el funcionamiento de la interfaz, explicando cada una de las secciones por separado.

En primer lugar, se detallar el funcionamiento de los distintos tipos de botones que entran en juego en esta aplicación, los cuales son:

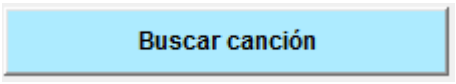

Figura 8.2: Botón buscar canción

Este botón nos permite seleccionar la canción con la que queremos trabajar, y a la cual le vamos a aplicar la segmentación y la separación.

Figura 8.3: Botón parámetros optimizados

Este botón nos permite introducir los parámetros, de forma directa y visual para el usuario, con los que el sistema de segmentación y separación funciona de forma óptima para este TFM.

Figura 8.4: Botón salir

Este botón nos permite cerrar la interfaz de forma directa.

Figura 8.5: Botón reset

Este botón nos permite resetear la interfaz por completo, limpiando todos los parámetros introducidos por el usuario, las canciones que se han cargado previamente y todas las gráficas de representación.

Figura 8.6: Botón segmentar

Este botón nos permite iniciar la etapa de segmentación para obtener los resultados, tanto auditivos, como visuales correspondientes a esta etapa.

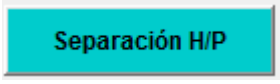

Figura 8.7: Botón separador H/P

Este botón nos permite iniciar la etapa de separación H/P para obtener los resultados, tanto auditivos, como visuales correspondientes a esta etapa.

Figura 8.8: Botón separador H/P/V

Este botón nos permite iniciar la etapa de separación H/P/V para obtener los resultados, tanto auditivos, como visuales correspondientes a esta etapa.

Figura 8.9: Botón de representación

Este botón nos permite iniciar las representaciones de las señales de entrada y de salida del sistema completo.

Figura 8.10: Botón play

Este botón nos permite iniciar la reproducción del audio correspondiente a cada sección.

Figura 8.11: Botón stop

Este botón nos permite parar la reproducción del audio correspondiente a cada sección.

Una vez que se ha detallado el funcionamiento de los distintos tipos de botones presentes en la interfaz de usuario, vamos a explicar cada una de las secciones que forman la interfaz. Para ello nos basaremos de un ejemplo correspondiente a una de las canciones de la base de datos ISMIR 2015 (la cual se utilizó para obtener algunos resultados de segmentación y separación).

- **Sección A: BASE DE DATOS**

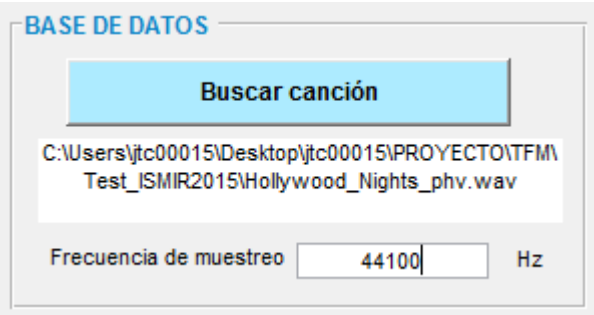

Figura 8.12: Sección correspondiente a la base de datos

Esta sección nos permitirá seleccionar la canción o se la musical con la que vamos a trabajar a lo largo de la utilización de la interfaz de usuario. En esta sección se puede introducir la frecuencia de muestreo con la que se va a digitalizar la canción. Se considera óptimo utilizar una frecuencia de muestro de 44100 Hz para señales musicales.

- **Sección B: PARÁMETROS DE ENTRADA**

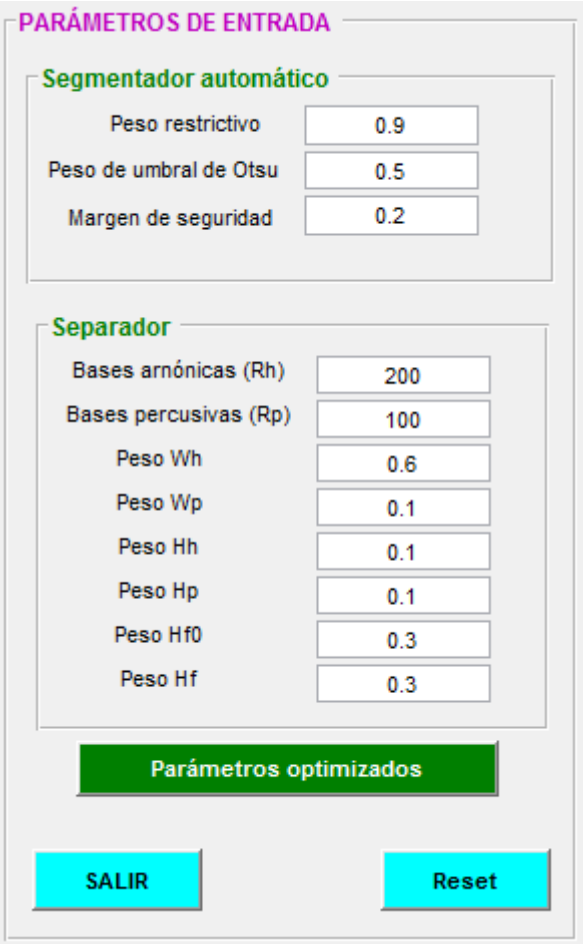

Figura 8.13: Sección correspondiente a los parámetros de entrada.

Esta sección nos permitirá introducir los distintos parámetros de entrada necesarios tanto para la etapa de segmentación, como para la etapa de separación. De esta forma se le dará libertad al usuario para probar las diferencias de utilizar unos parámetros u otros, y analizar las mejorar o las degradaciones que se produzcan. Además, se le da la posibilidad al usuario de introducir directamente los parámetros optimizados para este TFM. Por último, esta sección permite resetear todos los datos introducidos, así como las gráficas de los resultados y salir del programa.

- **Sección C: SEÑAL DE ENTRADA**

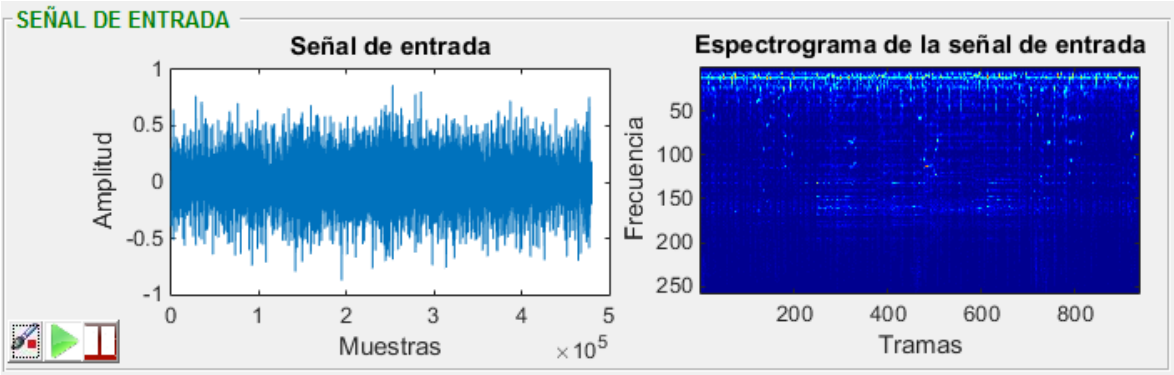

Figura 8.14: Sección correspondiente a la señal de entrada.

Esta sección nos permitirá analizar de una forma visual (mediante las gráficas correspondientes a la señal de entrada, y al espectrograma de entrada) y auditiva (reproducción de la señal de entrada con los botos de play y stop) la señal de entrada con la que se va a trabajar.

- **Sección D: SEGMNETADOR AUTOMÁTICO**

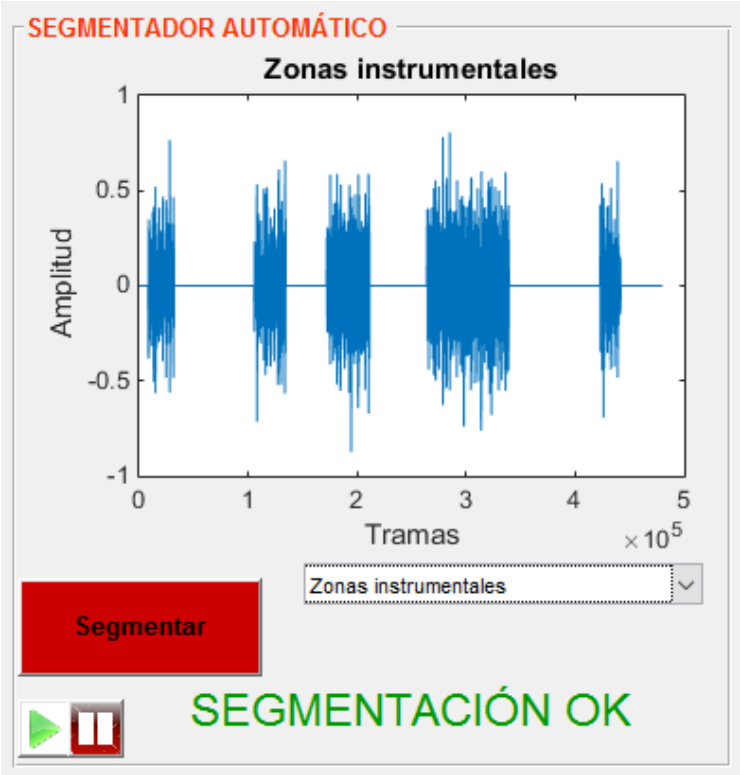

Figura 8.15: Sección correspondiente a la etapa del segmentador automático.

Esta sección nos permite iniciar la etapa de segmentación mediante el botón “Segmentar”. Una vez que ha finalizado nos parecerá el mensaje “SEGMENTACIÓN OK”. Seguidamente podremos hacer un análisis auditivo de las zonas instrumentales detectadas mediante los botones de play y stop. Además, permite llevar a cabo un análisis visual de las distintas fases que componen la etapa de segmentación hasta llegar a la solución final

de las zonas instrumentales detectadas (como se puede ver en la Figura 8.15). Para ello desplegaremos el menú que se muestra en la Figura 8.16 para ver las distintas fases, las cuales se pueden ver en la Figura 8.17.

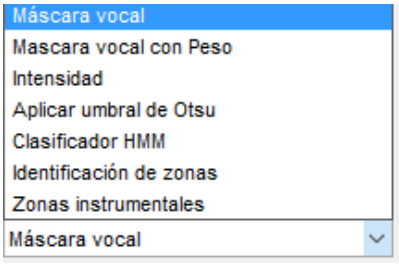

Figura 8.16: Opciones de visualización de la etapa de segmentación

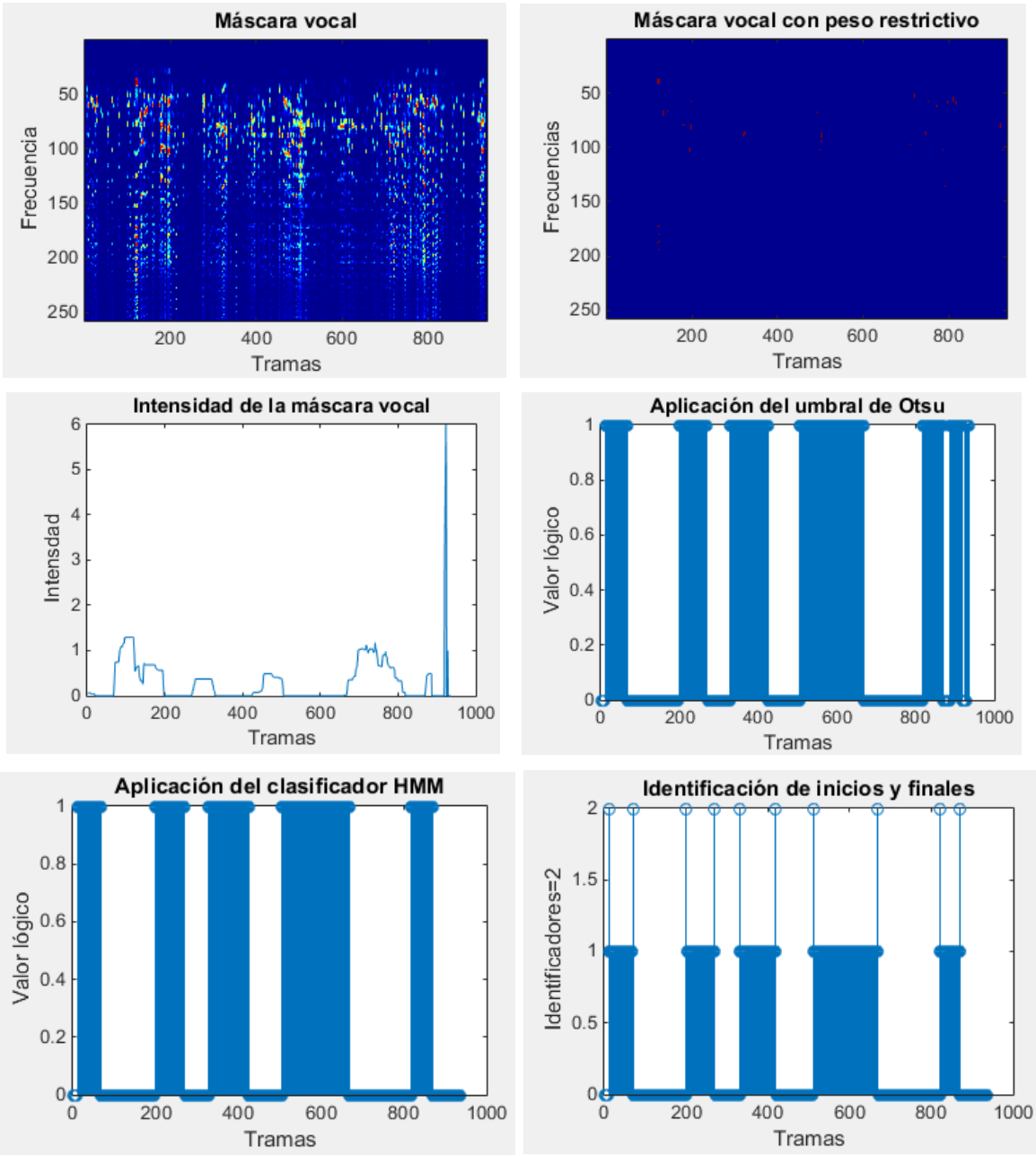

Figura 8.17: Distintas posibilidades del menú, que muestra las distintas fases de segmentación, hasta llegar a la solución final que se muestra en la imagen 83.

- **SECCIÓN E: SEPARACIÓN DE FUENTES SONORAS**

  Esta sección es la que nos va a permitir llevar a cabo la fase de separación de fuentes sonoras. Esta sección ofrece dos posibilidades, como se muestra a continuación:

- **Separación H/P:** en primer lugar, permite llevar a cabo una separación H/P donde únicamente estén presentes las fuentes armónicas y percusivas. Esto es útil para probar el sistema de separación independientemente de los resultados obtenidos en la etapa de segmentación.

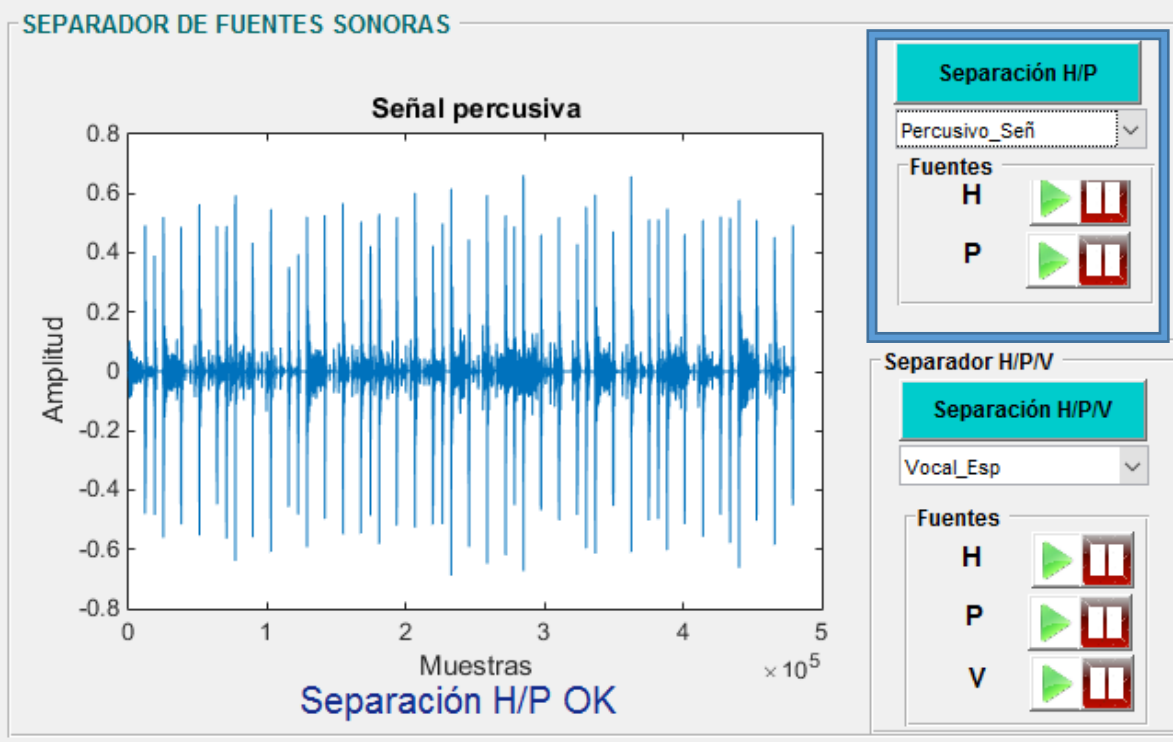

Figura 8.18: Sección correspondiente a la etapa del separador H/P.

Como se puede ver en la imagen anterior, una vez que se ha iniciado la etapa de separación, para fuentes armónico y percusivas, el final viene indicado por un mensaje de texto “Separación H/P OK”. Esta sección da la posibilidad de escuchar la fuente armónica y percusiva por separado. Además, dispone de un menú que permite mostrar la señal armónica, la señal percusiva, el espectrograma de la señal armónica y el espectrograma de la señal percusiva.

Es importante destacar de nuevo que esta sección es independiente de la etapa del segmentador ya que únicamente estamos separando la parte armónica y percusiva de una señal música (cargada previamente al clicar en el botón Separación H/P) que únicamente tiene presentes esas dos fuentes. Como se ha mencionado anteriormente esto es útil para probar el sistema independientemente de los resultados de las zonas instrumentales obtenidos del segmentador.

- **Separación H/P/V:** por otro lado, permite llevar a cabo una separación de las fuentes armónica, percusiva y vocal de la señal musical cargada previamente en la sección correspondiente a la base de datos. Es importante tener en cuenta que para llevar a cabo esta utilidad es necesario que anteriormente se hayan extraído las zonas instrumentales mediante la etapa de segmentación, ya que como se ha podido ver a lo largo del TFM, las zonas instrumentales extraídas del segmentador automático son utilizadas posteriormente para realizar un entrenamiento de las bases armónico y percusivas. Para que así posteriormente se puedan utilizar en las zonas vocales y realizar una separación de las tres fuentes de forma efectiva.

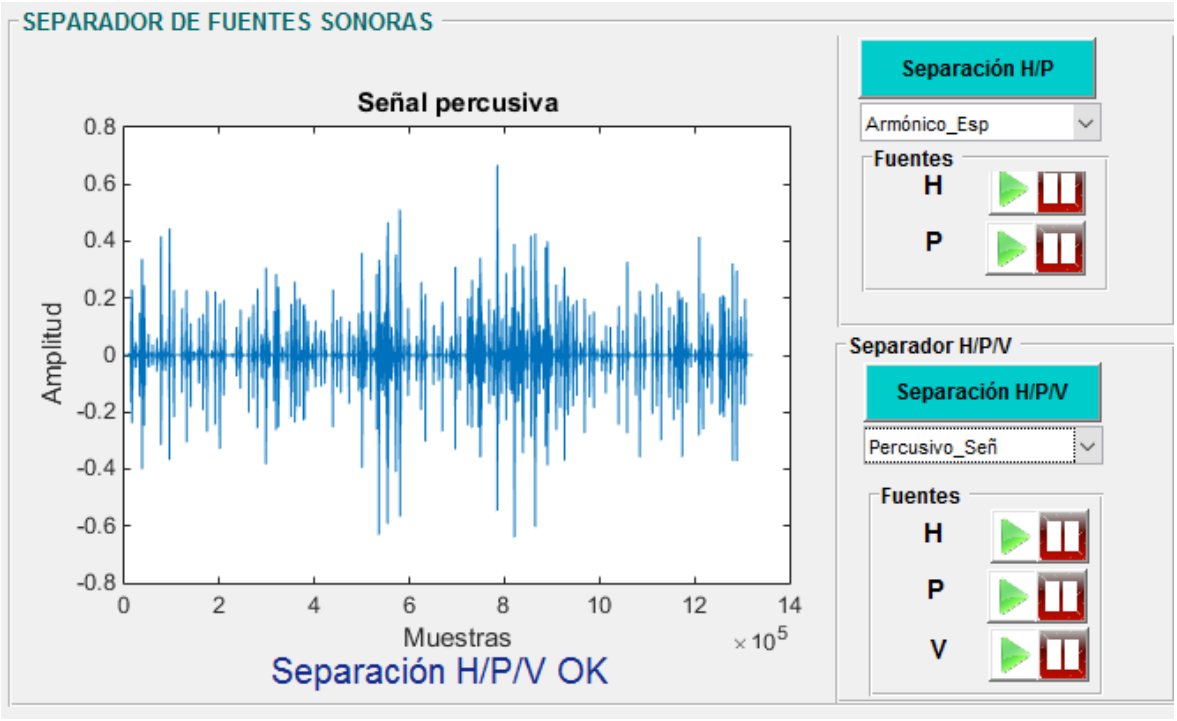

Figura 8.19: Sección correspondiente a la etapa del separador H/P/V.

Como se puede ver en la imagen anterior, una vez que se ha iniciado la etapa de separación, para fuentes armónico, percusivas y vocales, el final viene indicado por un mensaje de texto “Separación H/P/V OK”. Esta sección da la posibilidad de escuchar la fuente armónica, percusiva y vocal por separado. Además, dispone de un menú que permite mostrar la señal armónica, la señal percusiva, la señal, el espectrograma de la señal armónica, el espectrograma de la señal percusiva y el espectrograma de la señal vocal.

- **SECCIÓN F: REPRESENTACIÓN DE LA SEÑAL DE SALIDA**

  Esta sección nos permite representar el espectrograma de la señal de salida, una vez que se ha llevado a cabo el proceso de separación, y de esta forma poder ver la calidad con la que se ha reconstruido la señal. Es necesario tener en cuenta que se pueden ver dos posibles soluciones, una correspondiente a la opción separación H/P (en la que únicamente estarán presentes las cuentes armónicas y percusivas) y otra correspondiente a la opción separación H/P/V (en la que están presentes las fuentes sonoras, armónico, percusiva y vocal). En la siguiente imagen se muestra la señal reconstruida, teniendo en cuenta que se ha seleccionado la opción de que estén presentes las tres fuentes.

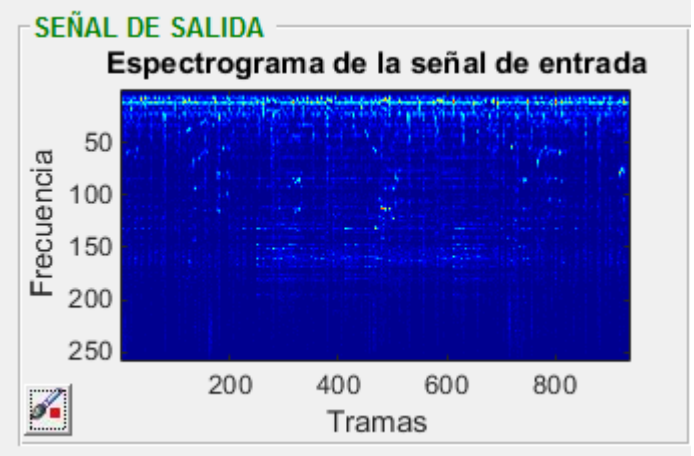

Figura 8.20: Sección correspondiente a la señal de salida del modelo NMF.

En la siguiente imagen se puede ver el resultado final de la interfaz tras haber realizado el ejemplo anterior paso a paso.

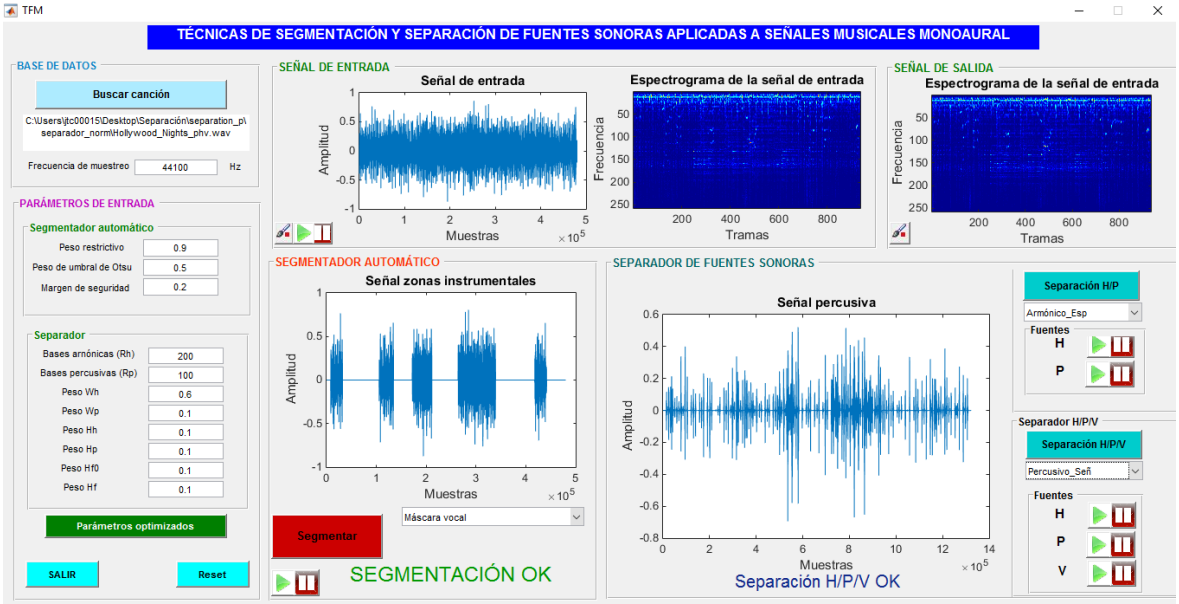

Figura 8.21: Interfaz de usuario para el segmentador automático y el separador de fuentes armónico/percusivo/vocal tras la ejecución de un ejemplo.

8.2 ANEXO II: NOMENCLATURAS

8.2.1 NOMENCLATURAS RELACIONADAS CON LA TEORÍA

ACC	Tasa de exactitud
AMT	Automatic Music Transcription
ASA	Auditory Scene Analysis
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DFT	Discrete Fourier Transform
DTW	Dynamic time warping
EUC	Euclidean (distancia)
F	Número de bins de frecuencia del espectro
FA	Tasa de falsas alarmas
FFT	Fast Fourier Transform
FN	Falso negativo
FP	Falso Positivo
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HR	Tasa de éxito
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
IS	Itakura Saito (divergencia)
ISS	Informed Source Separation
KL	Kullback-Leibler (divergencia)
MCA	Análisis de componentes morfológicos
MFCC	Mel Frequency Cepstral Coefficients
MFE	Estimación Multipitch
MIDI	Musical Instrument Digital Interface
MIR	Music information retrieval
NMF	Non- Negative Matrix Factorization
PLCA	Probabilistic Latent Component Analysis
REPET	Repeating Pattern Extraction Technique
SAR	Source-to-artifacts ratio
SC	Sparse Coding
SCA	Análisis de componentes dispersos
SDR	Source-to-distorsion ratio
SIR	Source-to-interferences ratio

SNR	Relación señal a ruido en decibelios
SSS	Sound Source Separation
STFT	Short Time Fourier Transform
T	Número de frames del espectro
TFG	Trabajo Fin De Grado
TFM	Trabajo Fin de Master
TP	Positivo cierto
VR	Voicing Recall
VFA	Voicing False Alarm

8.2.2 NOMENCLATURAS RELACIONADAS CON LAS ECUACIONES

x: seña de entrada

X: Espectrograma de la señal de entrada

$\hat{X}_n$: espectrograma estimado y normalizado de la función mezcla.

$\hat{X}_p$: espectrograma estimado y normalizado de la fuente percusiva.

$\hat{X}_h$: espectrograma estimado y normalizado de la fuente armónica.

$\hat{X}_v$: espectrograma estimado y normalizado de la fuente vocal.

$W_{P_{F,Rp}}$: matriz de bases percusivas.

$H_{P_{Rp,T}}$: matriz de activaciones percusivas.

$W_{H_{F,Rh}}$: matriz de bases armónicas.

$H_{H_{Rh,T}}$: matriz de activaciones armónicas.

$W_{F0_{F,Re}}$: matriz de bases de excitaciones vocales.

$H_{F0_{Re,T}}$: matriz de activaciones de excitaciones vocales.

$W_{F_{F,Rf}}$: matriz de bases de filtros vocales.

$H_{F_{Rf,T}}$: matriz de activaciones de filtros vocales.

$diag(ep_{Rp})$: diagonal de la norma de la parte percusiva.

$diag(eh_{Rh})$: diagonal de la norma de la parte armónica.

$diag(eFO_{Re})$: diagonal de la norma de las excitaciones vocales.

$diag(eF_{Rf})$: diagonal de la norma de los filtros vocales.

Rp: número de componentes percusivas utilizadas en la factorización.

Rh: número de componentes armónicas utilizadas en la factorización.

Re: número de componentes de la matriz de excitaciones vocales.

Rf: número de componentes de la matriz de filtros vocales.

K_{SSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las bases percusivas.

K_{TSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las activaciones percusivas.

K_{SSP}: es el peso que se le asignara a la restricción de dispersión para la matriz de las bases armónicas.

K_{TSM}: es el peso que se le asignara a la restricción de suavidad para la matriz de las activaciones armónicas.

K_{SAHFO}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de las excitaciones vocales.

K_{SAHF}: es el peso que se le asignara a la restricción de monofonía para la matriz de las activaciones de los filtros vocales.

M_p: máscara de la señal percusiva.

M_h: máscara de la señal armónica.

M_v: máscara de la señal vocal.

$x_p(t)$: señal percusiva temporal.

$x_h(t)$: señal armónica temporal.

$x_v(t)$: señal vocal temporal

8.3 ANEXO III: ÍNDICE DE FIGURAS Y DE TABLAS

8.3.1 ÍNDICE DE FIGURAS

8.3.2 ÍNDICE DE TABLAS

9 REFERENCIAS BIBLIOGRÁFICAS

[1] Zafar Rafii, Zhiyao Duan, and Bryan Pardo. “Combining Rhythm-based and Pitch-based Methods for Background and Melody Separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, December 2014.

[2] L. Le Magoarou, A. Ozerov and N. Q. K. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," Journal of Signal Processing Systems, July, 2014.

[3] J.-L. Durrieu and J.-Ph. Thiran, Source/Filter Factorial Hidden Markov Model, with Application to Pitch and Formant Tracking, IEEE Transactions on Audio, Speech and Language Processing, August 2013.

[4] Zhu, B., Li, W., Li, R., & Xue, X. (2013). Multi-stage non-negative matrix factorization for monaural singing voice separation. IEEE Transactions on audio, speech, and language processing, 21(10), 2096-2107.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. on Audio, Speech and Lang. Proc. special issue on Signal Models and Representations of Musical and Environmental Sounds, vol. 18, no. 3, pp. 550-563, March 2010.

[6] Perfecto Herrera-Boyer, Anssi Klapuri, and Manuel Davy. Signal Processing Methodsfor Music Transcription. Springer US, Boston, MA, 2006.

[7] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. G´omez, S. Steich, and B. Ong, “Melody transcription from music audio: Approaches and evaluation,” IEEE Trans. on Audio, Speech and Language Process., vol. 15, no. 4, pp. 1247–1256, 2007.

[8] R. Typke, “Music retrieval based on melodic similarity,” Ph.D. dissertation, Utrecht University, Netherlands, 2007.

[9] M. Ryyn¨anen and A. Klapuri, “Automatic transcription of melody, bass line, and chords in polyphonic music,” Computer Music J., vol. 32, no. 3, pp. 72–86, 2008.

[10] M. Ryyn¨anen and A. Klapuri, “Automatic transcription of melody, bass line, and chords in polyphonic music,” Computer Music J., vol. 32, no. 3, pp. 72–86, 2008.

[11] M. Goto, “A real-time music-scene-description system: predominantf0 estimation for detecting melody and bass lines in real-world audio signals,” Speech Communication, vol. 43, pp. 311–329, 2004.

[12] R. P. Paiva, T. Mendes, and A. Cardoso, “Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness,” Computer Music J., vol. 30, pp. 80–98, Dec. 2006.

[13] K. Dressler, “Audio melody extraction for mirex 2009,” in 5th Music Inform. Retrieval Evaluation eXchange (MIREX), 2009.

[14] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, “Adaptation of bayesian models for singlechannel source separation and its application to voice/music separation in popular songs,” IEEE Trans. on Audio, Speech, and Language Process., vol. 15, no. 5, pp. 1564 1578, Jul. 2007.

[15] J.-L. Durrieu, “Automatic transcription and separation of the main melody in polyphonic music signals,” Ph.D. dissertation, T´el´ecom ParisTech, 2010.

[16] J.-L. Durrieu, A. Ozerov, C. F´evotte, G. Richard, and B. D. David, “Main instrument separation from stereophonic audio signals using a source/filter model,” in Proc. 17th European Signal Process. Conf. (EUSIPCO), Glasgow, Aug. 2009, pp. 15–19.

[17] G. Poliner and D. Ellis, “A classification approach to melody transcription,” in Proc. 6th Int. Conf. on Music Inform. Retrieval, London, Sep. 2005, pp. 161–166.

[18] M. Rocamora and P. Herrera, “Comparing audio descriptors for singing voice detection in music audio files,” SBCM - Brasilian Sypmosium on Computer Music 07, 2007.

[19] W.H. Tsai, H.M. Wang, and D. Rodgers, “Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal,” in Eighth European Conference on Speech Communication and Technology. ISCA, 2003.

[20] H. Lukashevich, M. Gruhne, and C. Dittmar, “Effective singing voice detection in popular music using arma filtering,” Workshop on Digital Audio Effects (DAFx’07), 2007.

[21] AL Berenzweig and DPW Ellis, “Locating singing voice segments within music signals,” Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp. 119–122, 2001.

[22] A. Berenzweig, D.P.W. Ellis, and S. Lawrence, “Using voice segments to improve artist classification of music,” AES 22nd International Conference, 2002.

[23] Youngmoo E. Kim and Brian Whitman, “Singer identification in popular music recordings using voice coding features,” in ISMIR, 2002.

[24] T. Zhang, “Automatic singer identification,” Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on, vol. 1, 2003.

[25] W. Chou and L. Gu, “Robust singing detection in speech/music discriminator design,” Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on, vol. 2, 2001.

[26] Tin Lay Nwe and Haizhou Li, “Singing voice detection using perceptually-motivated features.,” in ACM Multimedia, Rainer Lienhart, Anand R. Prasad, Alan Hanjalic, Sunghyun Choi, Brian P. Bailey, and Nicu Sebe, Eds. 2007, pp. 309–312, ACM.

[27] L. R. Rabiner, B. H. Juang, "An introduction to hidden Markov models", IEEE ASSP Mag., vol. 3, no. 1, pp. 4-16, 1986.

[28] L. R. Rabiner, B. H. Juang, "An introduction to hidden Markov models", IEEE ASSP Mag., vol. 3, no. 1, pp. 4-16, 1986.

[29] J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans. on Audio, Speech, and Language Processing, 20 (6):1759–1770, 2012.

[30] M. Ramona, G. Richard, and B. David. ”Vocal detection in music with support vector machines”. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, pages 1885–1888. IEEE, 2008.

[31] S. Vembu and S. Baumann. ”Separation of vocals from polyphonic audio recordings”. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), volume 5, pages 337–344, 2005.

[32] B. Lehner, R. Sonnleitner, and G. Widmer, "Towards Lightweight, Real-time-capable Singing Voice Detection, " in Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR 2013), 2013.

[33] E Zwicker and H Fastl. Psychoacoustics, Facts and Models. Springer, 1990.

[34] A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," IEEE Trans. on Audio, Speech and Lang. Proc., vol. 20, no. 4, pp. 1118-1133, May 2012.

[35] A. Klapuri, T. Virtanen, and T. Heittola. Sound source separation in monaural music signals using excitation-filter model and EM algorithm, in proc. of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, 2010.

[36] J.J. Carabias-Orti, F.J. Rodriguez-Serrano, P. Vera- Candeas, F.J. Canadas-Quesada, and N. Ruiz-Reyes. Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription. Engineering Applications of Artificial Intelligence, 2013.

[37] Z. Duan and B. Pardo, "Soundprism: an online system for score-informed source separation of music audio," IEEE Journal of Selected Topics in Signal Processing, Volume: 5, Issue: 6, pp. 1205- 1215, October 2011.

[38] J.J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes and F.J. Canadas-Quesada. “Musical instrument sound multi- excitation model for nonnegative spectrogram factorization”. IEEE Journal of Selected Topics in Signal Processing, Volume: 5, Issue: 6, pp. 1144-1158, October 2011.

[39] F.J. Cañadas-Quesada, PP. Vera-Candeas, N. Ruiz-Reye, J.J. Carabias-Orti and P. Cabañas- Molero. “Percussive/Harmonic Sound Separation by Non-negative Matrix Approximation with Smoothness/Sparseness Constraints”, EURASIP Journal on Audio, Speech, and Music Processing, accepted for publication.

[40] S. Robila and L. Maciak. “Considerations on Parallelizing Nonnegative Matrix Factorization for Hyperspectral Data Unmixing." Geoscience and Remote Sensing Letters, IEEE, Volume 6, Issue 1, pp. 57-61, 2009.

[41] M.R. Every and J.E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. IEEE Transactions on Audio, Speech, and Language Processing, 14(5):1845–1856, 2006.

[42] Li Yipeng. Monaural Musical Sound Separation. PhD thesis, The Ohio State University, 2008.

[43] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. EURASIP Journal on Audio, 2010.

[44] A. Mesaros, T. Virtanen, and A. Klapuri. Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods. ISMIR, 2007.

[45] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. Audio, 2007.

[46] C K Lee and D G Childers. Cochannel speech separation. Journal of the Acoustical Society of America, March 1988.

[47] T. Parsons. Separation of speech from interfering speech by means of harmonic selection. Journal of the Acoustical Society of America, March 1976.

[48] M Weintraub. The GRASP sound separation system. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’84., pages 69–72, March 1984.

[49] M.R. Every and J.E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. IEEE Transactions on Audio, Speech, and Language Processing, 14(5):1845–1856, 2006.

[50] Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal processing, 24(1):1–10, 1991.

[51] J Barry. Polyphonic music transcription using independent component analysis. MSc, 2003.

[52] M Plumbley. Adaptive lateral inhibition for non-negative ICA. Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA), 2001.

[53] Canadas Quesada, F.J, Vera Candeas, P., Ruiz Reyes, N., et al. PERCUSSIVE/HARMONIC SOUND SEPARATION BY NON-NEGATIVE MATRIX APPROXIMATION WITH SMOOTHNESS/SPARNESS CONTRAINTS.Journal of New MusicResearch, 2012.

[54] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. Audio, 2007.

[55] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, 52(1):155–173, 2007.

[56] Dongmin Kim, Suvrit Sra, and Inderjit S Dhillon. Fast newton-type methods for the least squares nonnegative matrix approximation problem. Statistical Analysis and Data Mining, pages 38–51, 2008.

[57] P. Alonso, V. M. Garcia, F.J. Martinez-Zaldivar, A. Salazar, L. Vergara, and A.M. Vidal. Parallel approach to NNMF on multicore architecture. The Journal of Supercomputing, 2014.

[58] Jingu Kim and Haesun Park. Fast Nonnegative Matrix Factorization: An Active-Set-Like Method and Comparisons. SIAM J. Sci. Comput., 33(6):3261–3281, 2011.

[59] Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal processing, 24(1):1–10, 1991.

[60] M P Cooke. Modeling auditory processing and organisation. Cambridge University Press, Cambridge, March 1993.

[61] D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.

[62] D Lee and H Seung. Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 2001.

[63] F Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In Proc 6th International Congress on Acoustics, pages C–17 – C–20, 1968.

[64] Jean-Louis Durrieu, Associate Member, IEEE, Bertrand David, Member, IEEE, and Gaël Richard, Senior Member, IEEE "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation" IEEE Journal of Selected Topics In Signal Processing ,Vol.5,No.6,October 2011.

[65] Zhong Qu andLi Hang”Research on Iimage Segmentation Based on the Improved Otsu Algorithm.”,2010 [66] Z Rafii, B Pardo. A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure.

[67] Daniel P.W. Ellis. Beat Tracking by Dynamic Programming.

[68] D. Ellis. Beat tracking by dynamic programming. J. New Music Research, Special Issue on Beat and Tempo Extraction, 36:51–60, March 2007.

[69] D. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. Proc. Int. Conf. on Acous., Speech, Sig. Proc. ICASSP-07, 4:1429–1432,

[70] M. Ramona, G. Richard, and B. David, "Vocal detection in music with Support Vector Machines," in Proc. ICASSP '08, 2008, pp. 1885-1888.