Montevideo, 20 de agosto de 2018

Señores Miembros de la Subcomisión Académica de Posgrado
Instituto de Ingeniería Eléctrica
Facultad de Ingeniería, Universidad de la República

PRESENTE

Por medio de esta nota avalo la presentación de la versión definitiva de la Tesis de Maestría de Pablo Massaferro. El documento incorpora las sugerencias y correcciones del Tribunal de Tesis.

Sin otro particular los saluda atentamente,

Martín Rocamora
Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

ISSN 1688-2806

Universidad de la República
Facultad de Ingeniería

Identificación automática de cantante en música polifónica

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Pablo Massaferro Saquieres

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Eléctrica.

Directores de Tesis

Martín Rocamora ..... Universidad de la República
Pablo Cancela ..... Universidad de la República

Tribunal

Alvaro Gómez ..... Universidad de la República
Mauricio Delbracio ..... Universidad de la República
Luiz Biscainho ..... Universidad Federal de Río de Janeiro

Director Académico

Martín Rocamora ..... Universidad de la República

Montevideo
martes 31 julio, 2018

Identificación automática de cantante en música polifónica, Pablo Massaferro Saquieres.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 117 páginas.
Compilada el martes 31 julio, 2018.
http://iie.fing.edu.uy/

ACTA DE DEFENSA

TESIS DE MAESTRÍA

Fecha: Jueves 14 de junio de 2018 .-

Lugar: Montevideo, Facultad de Ingeniería – Universidad de la República.-

Plan de Estudio: Maestría en Ingeniería Eléctrica.-

Aspirante: **Pablo Massaferro Saquieres**.-

Documento de Identidad: 3,515,495-5

Director/es de Tesis: Dr. Martín Rocamora (DT); Dr. Pablo Cancela (coDT).-

Tribunal: **Dr. Luiz Wagner Pereira Biscainho (UFRJ, Brasil);**
Dr. Mauricio Delbracio (IIE, Fac. Ingeniería);
Mag. Álvaro Gómez (IIE, Fac. Ingeniería).-

Los miembros del Tribunal hacen constar que en el día de la fecha el **Sr. Ing. Pablo Massaferro** ha sido **APROBADO** en la defensa de su **Tesis de Maestría** titulada: **"Identificación automática de cantante en música polifónica"**

La resolución del Tribunal se fundamenta en los puntos detallados a continuación:

En la Tesis de Massaferro se exploran las técnicas existentes de identificación de cantantes en archivos de audio de música polifónica. En particular, se estudia el efecto del acompañamiento musical en el desempeño de la identificación. El trabajo es sumamente relevante en el contexto de búsqueda, clasificación y recomendación de música en grandes bases de datos, y es un tema activo de investigación.

Dentro de las contribuciones principales de esta Tesis se destacan:

- el diseño, implementación y evaluación de un sistema punta a punta, que a partir de un fragmento de audio identifica un cantante de una base de datos preestablecida,
- el diseño y creación de una base de datos de grabaciones de música cantada de calidad profesional, que permite abordar ciertos aspectos del problema de identificación de cantantes poco explorados en la literatura,

- el estudio exhaustivo de cómo el desempeño de la identificación es afectado por parámetros como la influencia del nivel de la voz respecto al acompañamiento y el largo del fragmento de audio usado para la identificación, entre otros.

Massaferro abordó el problema con una metodología rigurosa, generó un manuscrito claro y completo, e hizo una excelente presentación oral en la cual respondió con solvencia las preguntas del tribunal. Todo esto demuestra un conocimiento sólido en la temática abordada. Además, las contribuciones de su trabajo de Tesis abren interesantes perspectivas de investigación a futuro.

Para que conste,

Dr. Luiz Wagner Pereira Biscainho

Dr. Mauricio Delbracio

Mag. Álvaro Gómez

UNIVERSIDAD DE LA REPUBLICA – FACULTAD DE INGENIERIA – Julio Herrera y Reissig 565 CP. 11300
Tels: (598-2) 711 06 98– 711 07 98-711 08 98 – INTERNET: www.fing.edu.uy- Fax: 711.54.46- MONTEVIDEO-URUGUAY

Agradecimientos

Agradezco a los directores de esta tesis Martín Rocamora y Pablo Cancela por su tiempo, sus consejos, todos sus aportes y por su generosidad a la hora de compartir sus conocimientos.

Para realizar este trabajo fue fundamental la colaboración de Federico Graña quien aportó su conocimiento como productor musical y técnico de grabación, facilitando enormemente las tareas vinculadas a la generación de datos. Quiero agradecer especialmente a los músicos Nicolás Román, Diego Maturro, Lucía Ferreira, Sebastián Gavilanes, Javier Zubillaga, Diego Rosberg y Florencia Núñez por prestar su voz y su talento para que fuera registrado en este trabajo.

Por último un agradecimiento muy especial para Sofía García por acompañarme siempre y ayudarme en todas las etapas de esta tesis.

A Juani, Tina, Emi y Sofía.

Resumen

La aplicación de la tecnología digital a la producción y distribución de música ha dado lugar a una verdadera revolución, facilitando el acceso de los artistas a los estudios de grabación, y generando un crecimiento exponencial de la cantidad de registros fonográficos. Esto ha generado que los sistemas de clasificación y sugerencia, basados en herramientas de procesamiento de señales y aprendizaje automático, se hayan transformado en puntos clave en la gestión de la oferta musical. En este contexto, es de especial relevancia automatizar algunas tareas, como la identificación del cantante a partir de un archivo de audio.

La voz cantada es sin duda el instrumento musical más antiguo y familiar para nuestro sistema auditivo. Además, la voz suele transmitir mucha información en la música, porque generalmente interpreta la melodía principal, trasmite la letra y contiene características expresivas. Pero varios aspectos dificultan la tarea de reconocer automáticamente al cantante, en particular, a diferencia de la identificación del hablante, el acompañamiento musical es una señal de un nivel de energía similar al de la voz y no puede ser modelado como un ruido aleatorio independiente.

En este trabajo se exploran las técnicas existentes de identificación de cantantes en archivos de audio de música polifónica. Varios trabajos abordan el problema sin realizar separación de fuentes, debido a las dificultades que esto conlleva, lo que genera que los algoritmos de clasificación aprendan a reconocer al cantante junto con su acompañamiento musical.

La selección de la instrumentación, efectos de audio, mezcla y masterizado juegan un rol importante en el sonido final de las canciones que integran un álbum. En trabajos previos, los efectos vinculados a estos aspectos de la producción fonográfica han sido poco explorados. Para mostrar estos efectos y poder cuantificarlos, en este trabajo se crea la base de datos *VoicesUy*, en la cual canciones populares rioplatenses son cantadas por artistas profesionales y grabadas en multipista. Los cantantes interpretan las mismas canciones de forma de poder realizar identificación de voces entre archivos donde la única diferencia es la voz. Esta base de datos permite evaluar tanto algoritmos de separación de fuentes como de clasificación de voces. El hecho de que los cantantes que participan en la grabación de la base tengan su propia discografía, permite además evaluar la incidencia de los efectos de diferentes etapas de la producción musical en la identificación de cantante. *VoicesUy* es la primer base de datos de música popular en castellano para identificación de cantante y separación de fuentes.

Se presentan experimentos que muestran que, si bien el acompañamiento mu-

sical dificulta la identificación de cantante, un artista interpretando sus composiciones junto con su banda es más fácil de identificar que interpretando versiones. Denominamos a este comportamiento "efecto banda".

Se muestra cómo mejora la clasificación del intérprete al utilizar técnicas de separación de fuentes. Se prueba una técnica de enmascaramiento sobre una representación tiempo–frecuencia no tradicional y se comparan los resultados utilizado representaciones clásicas como el espectrograma. Para aplicar estas técnicas se utiliza la información de la frecuencia fundamental de la voz. Los resultados de identificación de cantante obtenidos son comparables con otros trabajos de referencia. La clasificación de voces sobre *VoicesUy*, aplicando separación de fuentes, alcanza un acierto del 95.1 %.

Tabla de contenidos

Glosario

ASR Automatic Speech Recognition. 15

BSS Blind Source Separation. 5, 13, 15, 58

CASA Computational Auditory Scene Analysis. 12, 34

CNN Convolutional Neural Network. 13, 16

CQT Constant Quality Transform. 57

DCT Discrete Cosine Transform. 36

DFT Discrete Fourier Transform. 58, 60

DTFT Discrete Time Fourier Transform. 36

EM Expectation Maximization. 16, 42, 43, 46

FChT Fan Chirp Transform. x, 13, 58, 63, 65, 66, 68, 70, 71, 73, 77, 80, 84

FFT Fast Fourier Tarnsform. 36, 58, 59

GFCC Gammatone Frequency Cepstral Coefficients. 11, 12, 39, 83

GlogS Gathered Log Spectrum. 52, 60

GMM Gaussian Mixture Model. x, 8, 9, 12, 42, 43, 45, 47, 49, 54, 83

HMM Hidden Markov Model. 9

ICA Independent Commponent Analysis. 13

ISMIR International Society of Music Information Retrival. 16

LPC Linear Predictive Coding. ix, 8, 9, 31, 38, 83

LPCC Linear Predictive derived Cepstral Coefficients. 9, 31, 38, 83

LPMCC Linear Predictive Mel-Frequency Cepstral Coefficients. 8, 9, 11, 38

Glosario

LSTM Long Short-Term Memory. 15

MFCC Mel-Frequency Cepstral Coefficients. IX, 8, 9, 11, 12, 31, 36, 38, 39, 44, 46, 54, 73, 83

MIR Music Information Retrieval. 3, 16, 17, 19

MIREX The Music Information Retrieval Evaluation eXchange. 17

MLP Multi Layer Perceptron. 12, 15, 16

NMF Non-negative Matrix Factorization. 12, 13, 15, 16

PLP Perceptual Linear Prediction. 8

RNN Recurrent Neural Network. 15

SAR Signal to Artefacts Ratio. 67, 69, 71, 73, 80, 84, 93, 96

SDR Signal to Distortion Ratio. 15, 67, 69, 71, 73, 80, 84, 94

SIR Signal to Interference Ratio. 67, 69, 71, 73, 80, 84, 92, 95

SNR Signal to Noise Ratio. 46, 67

SRC Sparse Representation Classifier. 11

STFT Short-Time Fourier Transform. 8, 12, 13, 16, 24, 52, 58, 62, 64, 68, 71, 73

SVM Support Vector Machine. 8, 9, 45

W-LPC Warped LPC. 8

Capítulo 1

Introducción

1.1. Motivación y contexto

La voz humana es definida por Platón como "un impacto del aire que llega por los oídos al alma" [1]. Lo cierto es que la voz humana ha sido objeto de estudio durante cientos de años, tanto desde el punto de vista fisiológico como desde el punto de vista musical. La voz no es solo la portadora del habla, contiene más información que el mensaje verbal. El habla es producto de la evolución de la especie humana por lo que el sonido de la voz, con sus modificaciones evolutivas, nos ha acompañado por miles de años. El ser humano es capaz de extraer información de identidad y de estado de ánimo en la percepción de la voz. Desde el punto de vista neurológico se puede trazar un correlato con la identificación de rostros [2].

La voz cantada es sin duda el instrumento musical más antiguo y el sonido musical más familiar para nuestro sistema auditivo. La versatilidad de la generación de sonidos vocales permite altísimos niveles de expresión donde pequeñas variaciones son fácilmente percibidas por los seres humanos [3, 4].

Los primeros registros de voces grabadas datan de finales del siglo XIX. En 1878 es patentado por Thomas Alba Edison el fonógrafo, basado en sistemas mecánicos, que podía registrar las perturbaciones de presión del aire generadas por el sonido. Primero fueron registrados en cilindros y luego en discos. Luego el registro fonográfico se expandiría acompañando las evoluciones tecnológicas, comenzando con las grabaciones eléctricas en los años 20 y las primeras experiencias de grabación magnética en los años 30. No es hasta finales de los '70 que se comienza a registrar música en formato digital [5]. El aumento en la capacidad de procesamiento de las computadoras y su reducción de costos en las últimas décadas ha generado un crecimiento exponencial en el registro fonográfico y en el acceso a éste de forma masiva a través de plataformas digitales como *spotify* [1] que ya cuenta con más de 30 millones de canciones. Este fenómeno ha motivado, en la últimas dos décadas, la investigación de extracción de información de audio para la realización automática de tareas tales como clasificación de música o transcripción de partituras (Music Information Retrieval,MIR).

[1] https://www.spotify.com

En este nuevo escenario, donde la oferta de música es personalizada y las opciones son inabarcables para una persona, los sistemas de clasificación y sugerencia, basados en procesamiento de señales y herramientas de aprendizaje automático, se han transformado en una herramienta indispensable para acceder a la enorme cantidad de música disponible. En este contexto, la identificación automática del cantante en un archivo de audio cobra relevancia.

1.2. Descripción del problema

La identificación automática de cantante se podría pensar como una extensión del problema de reconocimiento del hablante. Sin embargo, hay diversos aspectos que dificultan el uso de este abordaje. En particular, la voz cantada utiliza rangos dinámicos y variaciones tímbricas significativamente mayores [6]. Sumado a esto, el acompañamiento musical en el audio es una señal típicamente de un nivel de energía similar al de la voz y no se puede modelar como un ruido aleatorio independiente. Por tales motivos, los principales desafíos son lograr una correcta caracterización de la voz cantada que permita identificar al intérprete y minimizar los efectos adversos del acompañamiento musical en la caracterización e identificación del cantante.

Otro aspecto que dificulta la tarea de identificación automática del cantante, y que poco ha sido contemplado en los trabajos existentes, es el llamado "efecto álbum", introducido por Whitman et al. en 2001 [7]. En un álbum, los procesos de producción y pos-producción (selección de la instrumentación, efectos, grabación, mezcla y masterizado) juegan un rol importante en el sonido final de las canciones que lo integran. Resulta más difícil reconocer a un cantante en un cierto álbum si el sistema fue entrenado con otro álbum del mismo cantante. El problema fue mostrado por Mandel en 2005 [8] al entrenar y clasificar canciones de diferentes álbumes de los mismos artistas, mostrando una diferencia de desempeños considerable. En particular el proceso de masterización fue analizado por Kim et al. [9] al analizar diferencias entre audios de versiones de canciones originales y remasterizadas. Este último estudio no muestra su incidencia en el desempeño de sistemas de clasificación, pero si en la caracterización de las voces. Se podría diferenciar en este punto dentro del "efecto álbum" dos componentes : 1) el "efecto pos–producción" que tiene que ver con las modificaciones a los audios originales realizadas por los procesos de mezcla y masterización y 2) el "efecto banda" [2], que tiene que ver con las características propias del acompañamiento musical de un intérprete (composición, arreglos, selección de instrumentación, etc).

Para mostrar estos efectos y poder cuantificarlos no se puede utilizar una base de datos de música comercial. En este trabajo se crea la base de datos *VoicesUy*, en la cual canciones populares rioplatenses son cantadas por artistas profesionales y grabadas en multipista. Los cantantes interpretan las mismas canciones de forma de poder realizar identificación de voces donde lo único que cambia es el cantante, manteniendo el acompañamiento musical y pudiendo alterar su nivel en la mezcla.

[2] Conceptos identificados y denominados así por el tesista.

Esta base de datos permite evaluar tanto algoritmos de separación de fuentes como de clasificación de voces. El hecho de que los cantantes que participan en la generación de la base tengan su propia discografía, permite además evaluar la incidencia de los efectos mencionados. Dado que las canciones que integran la base son de diferentes compositores y el genero y la instrumentación también cambia, la comparación entre resultados obtenidos sobre *VoicesUy* y resultados obtenidos sobre la discográfica propia de los cantantes permitiría mostrar el "efecto banda". *VoicesUy* es la primera base de datos de música popular en castellano para identificación de cantante y separación de fuentes.

Diferentes enfoques han sido utilizados para reducir el efecto del acompañamiento musical en la identificación de cantantes, incluyendo estrategias de separación de fuentes (Blind Source Separation, BSS), las cuales se mencionan en el Capítulo 2. El problema de separación de fuentes, en señales de audio, es un tema muy vigente y en constante evolución debido a su dificultad y a la oportunidad de mejora de los resultados actuales. En esta tesis se exploran algunas técnicas de separación utilizando como información la frecuencia fundamental de la voz principal. Existen varios algoritmos que son capaces de estimar la frecuencia fundamental con precisión, en esta tesis se realizan las estimaciones de la frecuencia fundamental sobre la señal de la voz sin acompañamiento musical y se incluye como parte de las anotaciones de la base de datos.

1.3. Objetivos

Objetivo general: El objetivo principal de esta tesis es explorar los límites de las técnicas existentes en el problema de reconocimiento automático de cantante. Se busca mostrar la incidencia del "efecto banda" en el desempeño de los algoritmos de clasificación, así como analizar el impacto del uso de técnicas de separación de fuentes.

Objetivos específicos:

- Implementar un *framework* de identificación de cantante utilizando técnicas del estado del arte.
- Diseñar y generar una base de datos que permita observar los fenómenos ligados al objetivo principal de la tesis. Lo que implica contar con voces grabadas sobre un mismo acompañamiento musical y las mismas voces en álbumes comerciales.
- Diseñar experimentos que permitan: evaluar el desempeño de diferentes enfoques, dimensionar los efectos del acompañamiento musical, analizar el efecto de la duración de los audios en el grado de acierto y mostrar el "efecto banda".
- Probar algunas técnicas de separación de fuentes basadas en el uso de la frecuencia fundamental de la voz principal y su impacto en la identificación de cantantes.

1.4. Organización del documento

A continuación se describe la organización del resto del documento. En el capítulo 2 se realiza una revisión del estado del arte, dividiendo el problema en tres: identificación de cantante, separación de voz cantada y bases de datos. En el capítulo 3 se presentan dos bases de datos creadas específicamente para este trabajo. Se presenta el contenido y las técnicas utilizadas en su generación, así como la motivación de su creación. En el capítulo 4 se muestran las principales características utilizadas en los sistemas de identificación automática de cantantes, incluyendo un resumen de los mecanismos de generación de la voz humana y el modelado de la percepción auditiva. En el capítulo 5 se presenta el sistema de clasificación seleccionado para este trabajo y se muestran con experimentos los siguientes efectos sobre la clasificación: nivel de energía del acompañamiento musical, duración de intervalos de audio a clasificar y "efecto banda", entre otros. En dicho capítulo se presenta también una variación al sistema de clasificación utilizado que permite alcanzar mejores resultados. En el capítulo 6 se presentan diferentes técnicas de separación de la voz cantada, incluyendo el uso de representaciones tiempo–frecuencia no tradicionales y de la frecuencia fundamental de la voz en los archivos analizados. Dichas técnicas son aplicadas sobre dos bases de datos y los resultados son reportados de forma sistemática con métricas específicas. Por último, se presenta un experimento de clasificación de voces en contexto de música polifónica utilizando técnicas de separación de fuentes. Los resultados alcanzados son comparados con los principales trabajos revisados en la temática. Las conclusiones y trabajos futuros se incluyen en el capítulo 7.

Capítulo 2

Estado del arte

2.1. Introducción

Si bien desde los años 70 se ha trabajado en el problema de identificación del hablante, no es hasta los primeros años de este siglo que se comienza a investigar el problema de reconocimiento automático de cantante. Este es un problema típicamente de aprendizaje supervisado y, como cualquier problema de clasificación automática, implica generar un conjunto de atributos que caractericen a la señal y un algoritmo que modele el espacio de características y permita discriminar entre las diferentes clases. Este problema, analizado desde el punto de vista del procesamiento de señales, tiene la particularidad de que la señal de interés (la voz) está interferida por otras fuentes (los instrumentos musicales), lo que dificulta la correcta caracterización de cada clase (voz). Diferentes enfoques se han utilizado para disminuir el efecto del acompañamiento musical en la caracterización de las voces: selección de fragmentos (*frames*) de presencia de voz con sistemas de detección automática, separación de fuentes por factorización de matrices, ajustes cepstrales según modelos de generación vocal, separación de fuentes con aprendizaje profundo, entre otros. Algunas variables que inciden directamente en el desempeño han sido poco estudiadas, como es el caso de la duración de los fragmentos de audio de evaluación, el "efecto álbum" [7] y el "efecto banda", propuesto en esta tesis.

Por otro lado, varios conjuntos de datos han sido utilizados por diferentes grupos de investigación, lo que dificulta la comparación de diferentes técnicas (ver tablas 2.1 y 2.2). En este capítulo se revisa el estado del arte dividiendo el contenido en tres partes: identificación de cantante, separación de fuentes y bases de datos utilizadas.

2.2. Identificación de cantante

Un sistema de identificación de cantante implica al menos las siguientes etapas: 1) identificación de fragmentos de audio con presencia de voz, 2) cálculo de características a partir de las señales de audio que permitan capturar información relevante sobre las voces y 3) el entrenamiento de algoritmos de aprendizaje

automático. Para la identificación de presencia de voz diferentes métodos de entrenamiento automáticos han sido utilizados mientras que varios de los trabajos que se describen a continuación realizan la selección de forma manual.

Las primeras aproximaciones al problema de identificación de cantante se concentran en el problema de clasificación automática, sin tener en cuenta el efecto del acompañamiento musical en la identificación de voces. En 2002, Kim et al. [3] desarrollan, en el MIT Media Lab, un sistema de clasificación basado en la extracción de características acústicas utilizadas previamente para modelado, análisis y síntesis del habla. Las características usadas fueron Linear Predictive Coding (LPC) y Warped LPC (W-LPC) que es la misma predicción lineal pero sobre un eje de frecuencias no lineal, lo que permite mejor resolución en bajas frecuencias. En dicho trabajo se presentan resultados del orden del 55 % de acierto en identificación de regiones de presencia de voz, con un método que evalúa un índice de armonicidad. Se utiliza con una base de datos de 20 voces y como clasificadores para la identificación de cantante proponen el uso de Gaussian Mixture Model (GMM) y Support Vector Machine (SVM), obteniendo un desempeño de 38.5 % y 41.5 % respectivamente sobre los *frames* de presencia de voz seleccionados manualmente. Un sistema similar es presentado en 2003 por Zhang [10] para Hewlett-Packard Laboratories, donde también se determina la presencia de voz de forma automática y la caracterización es hecha con LPC sobre los coeficientes cepstrales en bandas Mel (Mel-Frequency Cepstral Coefficients (MFCC)). En dicho estudio se identifica la voz con GMM sobre una base de ocho voces alcanzando un desempeño de 84.4 %. La principal diferencia entre estos dos trabajos es que en el primero utiliza para representar a cada *frame* la Short-Time Fourier Transform (STFT), mientras que en el segundo, el filtrado LPC es realizado sobre una representación de cada *frame* hecha con MFCC (de ahora en más LPMCC). La principal similitud es que en ambos trabajos no se realizan procesamientos para disminuir el efecto del acompañamiento musical.

Tsai et al. [11] introducen el problema del acompañamiento musical y presentan un método de reducción de ruido derivado de métodos de identificación de voz [12, 13]. Para utilizar estos métodos se basan en la suposición de que el acompañamiento y la voz son estadísticamente independientes, para lo cual modelan el acompañamiento con las secciones de audio donde no hay presencia de voz y la voz con los *frames* del audio mezclado donde sí hay voz. En este último punto, el modelo presenta una limitación dada por la interferencia del acompañamiento. El problema de optimización planteado consiste en maximizar la probabilidad de que la voz esté dada por un modelo GMM, dados los modelos realizados de GMM para la "voz" y para el acompañamiento. Las características utilizadas en este trabajo son MFCC y Perceptual Linear Prediction (PLP). Con este método de preprocesamiento y clasificando con GMM una base de 23 cantantes, los autores obtienen un 87.8 % de acierto (utilizando solo segmentos de presencia de voz etiquetados manualmente).

El primer *framework* que, además de realizar una selección automática de *frames* por presencia de voz, utiliza un método de separación de fuentes para disminuir el efecto del acompañamiento musical en la clasificación de cantantes, es

presentado por Fujihara et al. en 2005 [14]. En la figura 2.1 se pueden ver las etapas propuestas por los autores. En primer lugar, se realiza una estimación de la frecuencia fundamental de la fuente más predominante en cada *frame*. Con dicha estimación se realiza un filtrado por enmascaramiento de la estructura armónica sobre el espectrograma (STFT) y la señal es resintetizada obteniendo una nueva señal de audio correspondiente a la fuente más prominente. La señal de audio resultante es modelada con características cepstrales y luego son seleccionados los *frames* que contienen voz, maximizando la verosimilitud respecto a dos modelos GMM realizados con una base de entrenamiento etiquetada. Los vectores de características de los *frames* seleccionados son utilizados para clasificar los cantantes con un sistema de clasificación por mezcla de gaussianas GMM con 10 modelos, dado que se utilizan 10 cantantes en la base. La base utilizada es tomada de "RWC Music Database: Popular" [15] de donde se seleccionaron 4 canciones de 10 cantantes. Esto permite presentar los resultados de clasificación de canciones enteras en validación cruzada con cuatro particiones. Otros 16 artistas de la misma base son utilizados para entrenar el algoritmo de selección de *frames*. Diferentes conjuntos de características fueron probados, incluyendo MFCC, LPC, Linear Predictive derived Cepstral Coefficients (LPCC) y LPMCC, alcanzando con este último un desempeño de 95.0 %.

La identificación de presencia de voz es un punto importante del sistema, ya que permite seleccionar los *frames* más relevantes. El sistema presentado en la figura 2.1 dos modelos GMM son entrenados con fragmentos de presencia y ausencia de voz. Otro enfoque con buenos resultados se basa en la clasificación con SVM y un pos–procesamiento con cadenas de Markov (Hidden Markov Model (HMM)) [16]. Una comparación sobre diferentes características espectrales utilizadas para identificar presencia de voz sobre modelos SVM es presentada por Rocamora et al. [17]. En este trabajo se concluye que los MFCC son las mejores características para identificar la presencia voz en dicho contexto.

En 2007, Mesaros et al. [18] abordan el problema del error que genera el acompañamiento musical en la clasificación. Proponen un método de síntesis de la voz basado en la extracción de la melodía principal (F0) y generando un modelo de serie de sinusoidales sobre los múltiplos de F0, donde las amplitudes y fase de cada armónico son estimadas de forma de maximizar la correlación con el audio original. Son probados como métodos de clasificación GMM y vecinos más cercanos, utilizando como medida de distancia la divergencia de Kullback-Leibler entre los modelos GMM. Dos puntos importantes diferencian este trabajo de los anteriores: 1) una base de datos es desarrollada con 13 voces sobre las mismas 4 canciones (20-30 segundos por canción) con el objetivo de reducir la incidencia de la instrumentación de cada intérprete (esta base no está disponible) y 2) el nivel de mezcla (relación de energía entre voz y acompañamiento musical) es modificado en los experimentos para evaluar su incidencia tanto en la clasificación sobre los audios de mezcla como sobre la señal de audio estimada. El trabajo no mide la calidad de la separación pero muestra que la clasificación de voces mejora de 36.0 % a 75.0 % al realizar separación de fuentes sobre una mezcla con igual energía de la voz y el acompañamiento musical.

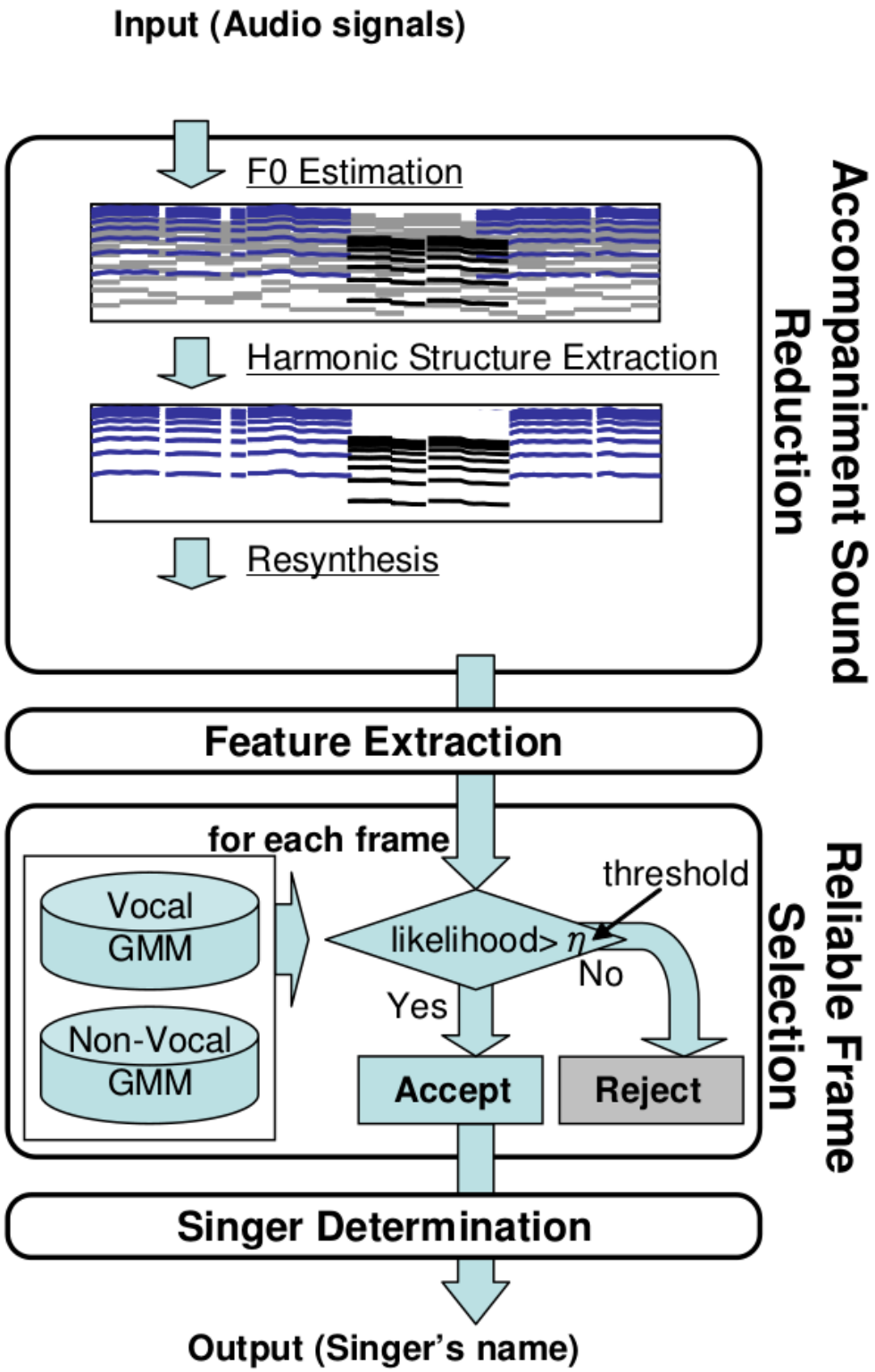


Figura 2.1: Método de reducción de efecto de acompañamiento musical en identificación de cantante. Imagen extraída de Fujihara et al. [14]

En 2010, Fujihara et al. [19] vuelven a presentar su método de clasificación pero agregando experimentos sobre una base de música comercial de 20 artistas. Realizan una comparación entre el uso de MFCC y LPMCC obteniendo 86.6 % y 95.3 %, respectivamente. Agregan, además, como característica de las voces, la pendiente de la curva F0 (frecuencia fundamental) en cada *frame* y presentan su *software* llamado *VocalFinder*.

Los dos métodos de separación presentados en [18,19] se basan en una correcta estimación de la F0, que en un contexto de música polifónica donde hay varias fuentes tonales, tiene sus dificultades. Tsai et al. [20] proponen utilizar un método de separación sin necesidad de calcular la F0. Utilizando una base de audios de karaoke estudian el efecto que genera sobre el cepstrum de la voz el acompañamiento musical. Se propone disminuir el impacto del acompañamiento ajustando parámetros de una transformación para pasar de la serie de MFCC de la mezcla a las MFCC de la voz sola. Se plantea como un problema de optimización ajustando los modelos de mezcla de gaussianas de ambos cepstrums. El algoritmo de reducción de acompañamiento es calibrado sobre una base 20 voces y probado sobre datos extraídos de la base RWC [15], los mismos datos que se usan en [14] y [19]. En el trabajo se menciona que las modificaciones realizadas en el cepstrum introducen importantes distorsiones en el audio si éste se quisiera recuperar, la calidad de la separación no es medida y este efecto es relativizado con el objetivo de clasificar las voces. Los resultados de clasificación alcanzados son de 92.5 % para las canciones completas. Este trabajo muestra, además, un punto interesante que otros trabajos no han analizado, que es la incidencia del tiempo de análisis sobre los resultados de identificación de cantante, mostrando una clasificación acierto del orden del 40 % cuando se analizan segmentos de 10 segundos. Este mismo enfoque es implementado y probado en otra base por Nithin et al. en 2014 [21] obteniendo 85.0 %. Trabajos previos [3,18] también utilizan fragmentos de canciones de corta duración, sin mencionar el hecho de que utilizar archivos de audio de pocos segundos de duración puede impactar negativamente en el desempeño del sistema de clasificación.

Un nuevo clasificador para identificar presencia de voz en música polifónica, Sparse Representation Classifier (SRC), es propuesto por Cai et al. [22] como primera etapa del sistema de identificación. El otro aspecto novedoso que presenta dicho trabajo es la inclusión de un nuevo set de características, Gammatone Frequency Cepstral Coefficients (GFCC), cuyo cálculo es similar al de MFCC pero sobre otro banco de filtros [23]. Muestra su mejor resultado al combinar MFCC, LPMCC y GFCC, obteniendo 90 % de acierto sobre una base de 10 voces. El grupo de datos de test contiene diez canciones por cantante y realiza el entrenamiento con de tres canciones por cantante.

En un estudio de Tsai del año 2012 [24] se muestra que utilizar archivos de audio de voz hablada para clasificar voces cantadas no genera buenos resultados. Sin embargo, se presenta un método para modificar el cepstrum de la voz hablada a voz cantada. Igualmente dicho método requiere tener datos de la voz cantada sola para poder realizar el entrenamiento.

Al realizar un pre–procesamiento de la señal, para realzar la voz se introducen

artefactos que se propagan a la extracción de características. Si bien esto había sido tratado de alguna forma al realizar selección de *frames* en trabajos previos, Lagrange et al. [25] muestran que las distorsiones afectan en diferentes instantes de tiempo a diferentes rangos de frecuencia y, por ende, la confiabilidad de diferentes características en cada *frame*. En dicho trabajo los autores adaptan una técnica presentada en 2005 [26] para reconocimiento del hablante, basada en incertidumbre gaussiana de las características. Basandose en trabajos anteriores, se utiliza como sistema de clasificación MFCC, modelado GMM y clasificación por máxima verosimilitud. En este trabajo se utiliza un método de separación para señales de audio estéreo [27]. Dicho método se basa en modelar la voz como una combinación de fuente y filtro, y el acompañamiento con Non-negative Matrix Factorization (NMF). Para cada *frame* de la STFT son estimados los parámetros de una representación de gaussiana multivariada. La matriz de covarianza es transformada en el cálculo de MFCC y para hacerlo se utiliza la aproximación de primer orden de la función, ya que el cálculo incluye la no linealidad de un logaritmo. La probabilidad de pertenecer a cada clase es ponderada por el valor de incertidumbre gaussiano obtenido. Los experimentos son realizados sobre la misma base de datos RWC que los trabajos de Fujihara [14,19] y Tsai [20]. Modelando con 32 gaussianas cada voz obtienen 94.0 % de acierto.

Un sistema basado en el modelado computacional de la percepción del sonido (Computational Auditory Scene Analysis, CASA) es presentado por Hu et al. en 2014 [28]. En primer lugar, el audio de mezcla es filtrado por un banco de filtros distribuidos en frecuencia por la escala de frecuencias de banda rectangular equivalente (ERB), que busca modelar la respuesta acústica de la cóclea, obteniendo una representación T-F alternativa llamada Cocleograma. Además, es estimada la frecuencia fundamental de cada frame F0. Con esta información se entrena una red neuronal (Multi Layer Perceptron, MLP) para determinar una máscara binaria que permite seleccionar en la representación T-F las regiones donde la información de la voz está menos distorsionada por el acompañamiento. Coeficientes GFCC con 64 filtros son utilizados como características para clasificar las voces sobre un modelo GMM previamente entrenado con los coeficientes de las voces solas. La calidad de la separación no es medida. Los experimentos son realizados sobre una base de 22 voces de música popular china extraída de una base de karaoke. El método propuesto alcanza valores de clasificación sobre segmentos de 4 segundos de 85.0 %.

Hasta este punto, todos los trabajos reseñados se basan en extraer características vinculadas al timbre de la voz para reconocer al cantante. Kroher et al. [29] usan información de más alto nivel para identificar voces: información sobre el vibrato y la interpretación. Para realizar los experimentos utiliza una base de música flamenca con 5 cantantes. El flamenco, a diferencia de la música pop, utiliza muchos ornamentos en la interpretación del cantante y arreglos de afinación no igualmente temperada. En el trabajo se presentan cuatro características derivadas del vibrato (media y desviación estándar del *rate*, duración promedio de los vibratos y cantidad total) y 13 de la interpretación (se destacan rango de F0, máximo y mínimo F0 y fluctuación de afinación). Las características son extraídas de la

curva F0, la cual es calculada utilizando MELODIA vamp plugin [30]. Utilizan SVM como clasificador, reportando una mejora al incluir las nuevas características sobre una base de voces a capella. Otro experimento es realizado sobre una base polifónica, reportando que los resultados no se ven alterados al incluir las nuevas características.

Recientemente, en 2018, se aborda el problema de clasificación de voz cantada utilizando *Deep Learning*. Wang et al. [31] proponen el uso de una Convolutional Neural Network (CNN) para derivar características de espectrogramas basados en la escala Mel [32]. Se experimenta utilizando una selección de datos de la base DAMP[1] con 46 voces y 10 canciones a capella. Se prueban tres arquitecturas alcanzando su mejor valor para una red CNN simple de tres capas y Softmax como clasificador lineal en la última capa de la red. Alcanza un desempeño de 74 %.

2.3. Separación de voz cantada

El problema de separación de fuentes (BSS) es un problema fundamental en el área de procesamiento de señales. Ha sido tratado con diferentes propósitos, tanto para señales biológicas, audio o imágenes durante muchos años y sigue siendo un tema muy vigente. En particular, en el tratamiento de señales de audio para realzar la voz humana, han habido fundamentalmente tres enfoques en los últimos años: filtrado por enmascaramiento sobre la frecuencia fundamental y sus armónicos [14, 18, 19], factorización de matrices [33–40] y, en los últimos tres años, aprendizaje profundo [41–46].

Hasta 2015, la factorización en matrices no negativas (NMF, por sus siglas en inglés) del espectrograma es el método predominante en la separación de voz grabada en un contexto de ruido. NMF se basa en la aparición repetida de ciertos patrones espectrales generando, básicamente, un diccionario de espectro y una matriz de ocurrencias. En 2007, Virtanen et al. [33] agregan al NMF básico un término para favorecer la continuidad entre *frames*, de forma de modelar mejor las evoluciones temporales de la voz. Comparan resultados con metodologías previas como el ICA (Independent Commponent Analysis). Centrados en las características melódicas de la voz cantada, ya se habían presentado algunas técnicas de enmascaramiento del espectrograma (STFT) sobre la frecuencia fundamental y sus armónicos [14] [18], pero una combinación de ambas metodologías (enmascaramiento armónico y NMF) es presentada en [34]. En dicho trabajo son ajustados los coeficientes de la máscara modelando el acompañamiento con NMF sobre las secciones de audio donde la voz no está presente. La gran mayoría de los trabajos se basan en la representación de tiempo y frecuencia clásica, la STFT. En 2011, Durrieu et al. [35] presentan una representación alternativa basada en un modelo fuente-filtro. La representación es generada al realizar una factorización de matrices con el modelo de fuente glotal KL-GLOTT88 [47]. Otras representaciones alternativas, como la FChT (Fan Chirp Transform) [48] presentada por Cancela et

[1] https://ccrma.stanford.edu/damp/

Tabla 2.1: Resumen de experimentos de los principales trabajos en identificación automática de cantantes. Los resultados son obtenidos sobre diferentes bases de datos, por más detalles referirse a los artículos citados.

Autor	Cantantes	Selección	Separación	Características	Clasificación	Resultado
Kim 2002 [3]	17	H	-	LPC/ W-LPC	SVM	41.5 %
Zhang 2003 [10]	8		-	12 LPMCC	GMM 10	84.4 %
Tsai 2003 [11]	23	GMM		20 MFCC	GMM 64	87.8 %
Fujihara 2005 [14]	10	GMM	FFT / máscara	15 LPMCC	GMM 64	95.0 %
Mesaros 2007 [18]	13	MFCC-0	F0 síntesis	12 MFCC	GMM 10	75.0 %
Fujihara 2010 [19]	20	GMM	FFT / máscara	15 LPMCC/ DF0	GMM 64	95.3 %
Tsai 2011 [20]	10	GMM	Cepstrum	20 MFCC	GMM 64	92.5 %
Cai 2011 [22]	10	SRC	-	13 MFCC / 15 LPMCC / 13 GTCC	GMM 5	90.0 %
Lagrange 2012 [25]	10	manual	Source-Filter / NMF	MFCC	GMM 32	94.0 %
Hu 2014 [28]	22	manual	Cocleagrama / MPL	64 GTCC	GMM 512	85.0 %
Kroher 2014 [29]	5	-	-	13 MFCC / 4 Vibrato / 13 Interp	SVM	88.0 %
Wang 2018 [31]	46	-	solo voz	CNN (3 capas)	Softmax	74.8 %

al., pueden ser utilizadas también como alternativas al espectrograma para realizar separación de voz cantada [49]. Con el objetivo de tener una definición más precisa de la matriz de diccionario de NMF un método para el control de la "esparcidad" es presentado por Gao et al. [50] y es probado en separación de instrumentos musicales como piano y trompeta en tríos de jazz. Otro método de factorización diferente a NMF es presentado por Sprechmann et al. [36] donde se presenta un *framework* para representar señales cuasiarmónicas, independizando la frecuencia fundamental de su envolvente espectral. Suponiendo conocida la frecuencia fundamental, una de las matrices de la descomposición contiene las variaciones de frecuencia y su estructura armónica, mientras la otra contiene modelos de mezcla de gaussianas de las envolventes espectrales. Se utiliza música de cuarteto de cuerdas con anotaciones y muestran mejoras comparando con el método de NMF de Hennequin [37].

Basados en la idea de que la música pop tiene estructuras que se repiten periódicamente en el tiempo, Rafii et al. [51] proponen utilizar técnicas de tratamiento de imágenes para extraer el fondo de las imágenes (parte estática), para luego mejorar el método agregando matrices de similitud donde el período de repetición no tiene por qué ser estable [38]. La autocorrelación del espectrograma es usada para calcular el período de repetición y un filtro de mediana de segmentos de igual período es realizado para estimar el acompañamiento. Un resumen de los trabajos sobre esta línea de repetición y la presentación del *framework* REPET se encuentra en [52]. Una comparación detallada de los diferentes modelos utilizados para separación hasta 2014 fue realizada por Vincent et al. en [39] y el *framework* FASST para implementarlo se presenta en [53]. Una comparación de metodologías y de las implementaciones REPET, FASST, KAM [54] y RPCA [41] es realizada en 2015 por Lehner et al. [55]. Por otra parte, un resumen de los métodos de separación basados en factorización y su extensión a modelos temporales es presentado por Smaragadi et al. [40] en 2014.

El gran desarrollo del aprendizaje profundo a partir de 2015 genera un nuevo enfoque hacia el problema BSS, básicamente como una etapa de pre-procesamiento enfocado al reconocimiento de palabras y frases en archivos de audio de voz hablada. Los modelos basados en aprendizaje profundo (*Deep Learning*) han irrumpido recientemente como una alternativa poderosa a los métodos tradicionales.

Con el objetivo de mejorar la inteligibilidad de la señal de voz en [42] se prueba el uso de redes neuronales recurrentes (Recurrent Neural Network (RNN)) Long Short-Term Memory (LSTM) para realzar la voz y poder aplicarse al reconocimiento del hablante robusto al ruido (Automatic Speech Recognition (ASR)). El enmascaramiento se hace sobre el espectro complejo logrando preservar mejor la fase alcanzando así mejores resultados en la síntesis. Sobre la base de datos de "The 2nd 'CHiME' Speech Separation and Recognition Challenge" [56] muestran una mejora significativa respecto al uso de NMF de 5db a 14db de SDR (Signal to Distortion Ratio). Siendo éste uno de los pocos trabajos en estimar las modificaciones en la fase del espectrograma es importante notar que reportan por tal motivo una mejora de 0.5db. RNN también es probada para estimación de máscara suave en [41]. Una arquitectura de MLP de tres capas ocultas es utilizada en [43]

para ajustar un filtro de Wienner, donde la STFT de la mezcla es representada por una sumatoria de gaussianas multivariadas de valores complejos, dadas por la densidad de potencia espectral de cada fuente y su matriz de correlación. Se supone independencia entre las fuentes. Los pesos de la red son ajustados iterando por EM (Expectation Maximization). Utilizando los datos del concurso de separación CHiME-3 [57], se compara el método presentado con el método de separación basado en ajuste de NMF por EM ([53]), mostrando una mejora muy significativa en SDR de 7.7 a 13.2 dB. Diferentes funciones de costo son mostradas obteniendo la divergencia Kullback-Leibler los mejores resultados.

La arquitectura de redes convolucionales (Convolutional Neural Network (CNN)) es probada en [44] con el objetivo de tratamiento en tiempo real, disminuyendo la latencia respecto a sistemas basados en NMF. El *framework* propuesto modifica la magnitud de la STFT, mientras que la fase se mantiene según el audio original. Las redes convolucionales utilizan muchos menos parámetros que otros algoritmos de *Deep Learning*, lo que permite disminuir el tiempo a la cuarta parte respecto al uso de MLP (Multi Layer Perceptron). CNNs también son probadas en separación de voces recientemente en [46].

Con la motivación de tratar la restauración de un espectrograma como un problema de imágenes, utilizando *Deep Learning*, en [45] proponen el uso de la arquitectura U-Net desarrollada para el tratamiento de imágenes médicas. Utilizan una arquitectura de seis capas de convolución de dos dimensiones. La red trabaja con el módulo del espectrograma generando una máscara a la salida y utilizando la fase original para la reconstrucción. Dada la complejidad computacional del algoritmo de entrenamiento, los audios son submuestreados a 8192Hz. El *framework* es probado sobre las bases MedleydB [58] e iKala [59].

2.4. Bases de datos

Uno de los puntos más importantes en cualquier investigación de ciencia aplicada es la calidad y cantidad de datos y su accesibilidad. Las investigaciones que se basan en la extracción de información de audios musicales (MIR) se enfrentan al problema de uso de materiales fonográficos comerciales con derechos de autor, esto motiva la creación de bases de datos específicas orientadas a la investigación[2]. Por ejemplo en 2002 fue creada la base RWC [15] que contiene cuatro bases de datos originales: the Popular Music Database (100 canciones), Royalty-Free Music Database (15 canciones), Classical Music Database (50 piezas), y Jazz Music Database (50 canciones). Para la base de datos de música popular se crearon 100 composiciones originales, para lo cual trabajaron 25 compositores, 30 letristas, 23 arregladores y 34 cantantes entre otros. Cada una de las piezas fue transcrita en MIDI de forma de tener información que sirva para diferentes problemas de MIR.

[2]Una variada cantidad de bases para MIR puede ser encontrada en el sitio oficial de ISMIR (International Society of Music Information Retrival), http://www.ismir.net/resources.html o en la página web del libro "An Introduction to Audio Content Analysis" de Alexander Lerch, http://www.audiocontentanalysis.org/data-sets/.

Específicamente para el problema de separación de voz cantada es necesario contar con al menos dos pistas de audio por canción, una para la voz y otra para el acompañamiento musical. Hasta el año 2010, la mayoría de los trabajos utilizaban CDs comerciales y alguna base es generada en formato karaoke sin hacerla pública [18]. Ese año se presenta la base MIR-1k [60] específicamente diseñada para separación de voz cantada, contando con 110 canciones interpretadas por 19 cantantes amateurs en formato karaoke. Los audios están segmentados en 1000 fragmentos de 4 a 13 segundos y sólo contienen audios con presencia de voz. Esta base incluye información de melodía principal y etiquetado de fonemas.

Competencias internacionales como CHiME o MIREX (The Music Information Retrieval Evaluation eXchange) han desarrollado sus propias bases de datos para comparar el desempeño de algoritmos en diferentes temáticas de MIR. En particular MIREX 2009 incluye una base de 374 canciones de pop chino para evaluación de algoritmos de estimación de la frecuencia fundamental más predominante.

En los últimos años, varias plataformas web han posibilitado compartir música de forma libre. Un caso que se aplica al problema de separación es el de la comunidad CCMIXTER donde más de 45 mil músicos suben sus arreglos musicales y otros miembros cantan para que DJs profesionales y aficionados generen sus propias mezclas. La herramienta de separación KAM [54] utiliza audios de esta plataforma en su evaluación.

El CCRMA (Center for Computer Research in Music and Acoustics) de la universidad de Stanford crea una base (DAMP-balanced) de karaoke con 5429 voces a capella interpretando algunas canciones de una lista de 14. La base contiene un total de 24874 interpretaciones de canto, la cual podría ser utilizada para identificación de cantante. La base no cuenta con metadata más que la identidad de los cantantes (amateurs), su edad, sexo y nacionalidad.

Para realizar algunas investigaciones es necesario generar anotaciones a los audios y en este punto muchos de los trabajos realizaron grandes esfuerzos de etiquetado manual con personas idóneas. En particular, el etiquetado de la melodía de varios instrumentos, como la voz y su activación, es fácilmente automatizable y da resultados confiables si se cuenta con las pistas de cada fuente por separado. Una base muy completa multi-pista es creada en el año 2014 en la NYU (New York University) con el nombre MedleyDB [58]. Esta base contiene 128 canciones de diferentes géneros, de las cuales 62 contienen voz cantada con anotaciones de melodía y activación de los diferentes instrumentos presentes en cada canción.

La base de datos iKala [59] es diseñada específicamente para separación de voz cantada, contando con un total de 352 fragmentos de canciones de 30 segundos de duración, de los cuales 100 no se hicieron públicos reservándolos para la evaluación de algoritmos de separación en MIREX 2016 (el último realizado a la fecha sobre esta temática). Cada canción contiene anotaciones realizadas de forma manual de frecuencia fundamental y tiempo de ocurrencia de las letras de cada canción. La base no está diseñada para identificación de cantante.

En la tabla 2.2 se resumen los principales parámetros de las bases de datos públicas mencionadas. Es importante notar que solo una de las bases está en formato multi-pista y que básicamente manejan los idiomas japonés, mandarín e

inglés. Si bien varias de estas bases han sido utilizadas para trabajos dentro de la separación de fuentes, su uso en el problema de identificación de cantante básicamente se resume a una selección de datos de la base RWC de 10 artistas y 4 canciones de cada uno ([14, 20, 22, 25]).
Dadas las limitaciones de estas bases de datos para un análisis completo del problema de identificación de cantante en música polifónica, en el Capítulo 3 se presenta el diseño y la creación de una base de datos apropiada para los objetivos de esta tesis.

Tabla 2.2: Comparación de bases públicas de MIR enfocadas que pueden ser utilizadas para separación o identificación de cantante.

Base	Audio	Pistas	Canciones	Dur. med. (s)	Cantantes	Anotaciones	Idioma
RWC-pop	16kHz,16bits	No	100	180	34	MIDI	Jap-Ing
MIR-1k	16kHz,16bits	Karaoke	110	60	19	F0/fonemas	Mandarín
MIREX 09	22.05kHz,16bits	Karaoke	374	–	–	No	Mandarín
MedleyDB	44.1 kHz,16 bits	Multitrack	64	180	42	mF0/V_act/INS_act	Ing
iKala	44.1 kHz,16 bits	Karaoke	352	30	–	F0/letra	Mandarín
DAMP	mp3 –	Karaoke	(24874) 14	180	5429	No	Ing

Capítulo 3

Descripción de bases de datos

3.1. Introducción

Al realizar una revisión de las bases de datos disponibles y de los trabajos realizados en el área de clasificación de voz cantada, se ve que, si bien existen varias bases de datos para evaluar algoritmos de separación de voz, pocas son las opciones para evaluar algoritmos de identificación de voz cantada. En particular, no existe ninguna base de cantantes profesionales que permita evaluar la incidencia de la instrumentación en la clasificación y el efecto de la masterización. Una base de datos donde solo las voces sean la variable, del estilo karaoke, donde todos los cantantes interpretan las mismas canciones, permitiría estudiar los efectos del acompañamiento. Por otra parte, si de cada una de las voces de la base se contara con un álbum editado con todos los procesos de producción que esto implica, se podría evaluar el impacto del "efecto álbum" [7].

Dentro de las etapas de producción de un álbum podemos identificar cinco etapas como se muestra en el esquema de la figura : 1) Composición, 2) Selección de instrumentación y arreglos musicales, 3) Grabación, 4) Mezcla y 5) Masterizado.

En este trabajo definiremos entonces el efecto de los ítems 1 y 2 como "efecto banda" y el de los ítems 4 y 5 como "efecto pos–producción", ambos efectos forman parte del "efecto álbum".

Con el objetivo de que estos puntos puedan ser estudiados, se crea una base de datos convocando a grabar a artistas uruguayos que hayan editado al menos un álbum de forma profesional, donde su voz sea la voz líder.

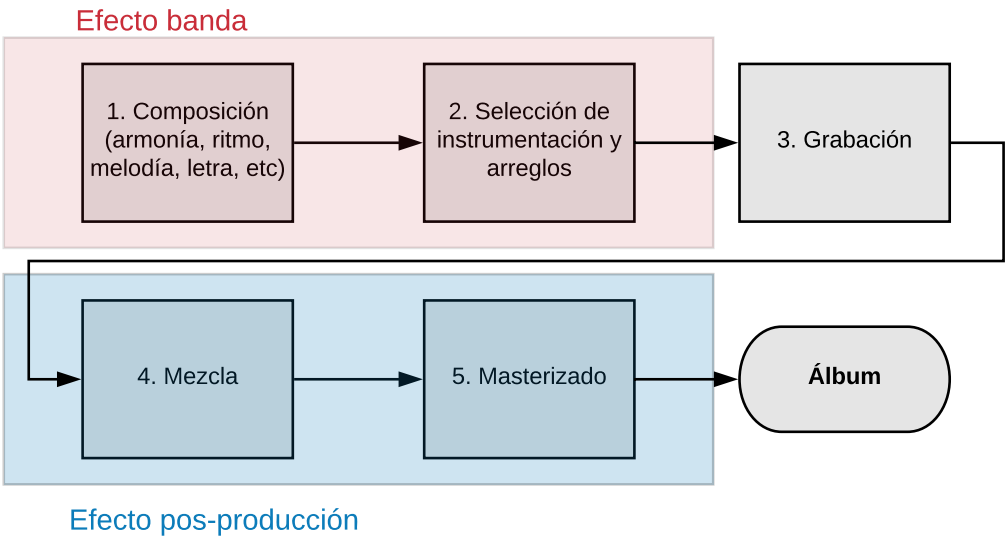


Figura 3.1: Esquema de procesos de producción musical.

3.2. Base de datos *VoicesUy*

3.2.1. Descripción

La base de datos creada cuenta con ocho voces de las cuales seis son masculinas y dos femeninas. Cada cantante interpreta cinco canciones. Lo que da un total de 40 canciones en idioma español con un total de 83 minutos de audio en fomato wav con cuantización de 16 bits y una frecuencia de muestreo de 44.100 Hz.

Se seleccionan cinco canciones de música popular, cuatro de ellas de origen uruguayo y una argentina. El acompañamiento musical de cada canción es generado específicamente para esta base y si bien difiere de sus versiones originales respeta la estructura armónica de la composición original[1]. Todos los instrumentos son grabados de forma independiente generando pistas de audio separadas. En la tabla 3.1 se muestran los temas seleccionados, su compositor, la duración total y la instrumentación utilizada en la grabación. Todos los instrumentos son grabados en *San Fruta Records* por los músicos Federico Graña, Nicolás Román y Pablo Massaferro. Llamaremos a esta base *VoicesUy*.

Todas las tomas de voces son realizadas con los mismos equipos. Se utiliza el micrófono *Rode NT2000*, que es un micrófono de calidad de estudio de condensador

[1]La producción musical estuvo a cargo del músico Federico Graña.

Tabla 3.1: Canciones que componen la base de datos *VoicesUy* y su instrumentación

Canción	Compositor	Duración	Instrumentación
Biromes y servilletas	Leo Masliah	1:26	batería, bajo, piano, guitarra eléctrica
La edad del cielo	Jorge Drexler	1:30	batería, samples de percusión, bajo, guitarra eléctrica
Pa' los músicos	Federico Graña	3:40	batería, bajo, hammond, guitarra acústica, guitarra eléctrica, piano
Príncipe azul	Eduardo Mateo	2:11	bajo, guitarra acústica, banjo, mandolina, cascabeles, guitarra eléctrica, dobro
Promesas sobre el bidet	Charly García	1:34	batería, bajo, guitarra eléctrica, sintetizadores, hammond

de diafragma grande. Éste es un equipo de muy bajo ruido interno, 7dBA (dB ajustados por la curva de modelado de la audición A) al soportar una presión sonora máxima de 147dB permite captar un amplio rango dinámico. El micrófono va conectado con un cable balanceado (que evita los ruidos de modo común) a un pre–amplificador de estudio de la línea *Universal Audio* modelo 4-710d; este amplificador combina la tecnología valvular con estado sólido logrando grabaciones de calidad profesional. El equipo tiene un filtro pasa altos en 75 Hz y la conversión a digital es hecha en 24bits con una frecuencia de muestreo de 192 kHz. Todos los archivos de audio son submuestreados para generar los archivos en formato wav de 44.100 muestras por segundo.

3.2.2. Contenido base *VoicesUy*

La base *VoicesUy* está compuesta por cuatro carpetas conteniendo los archivos de audio de voces e instrumentos, anotaciones con información sobre la pista de voz, audios de mezcla de voz y acompañamiento musical y archivos de audio de los resultados de separación de fuentes (ver Capítulo 6).

Audio

Dentro de la carpeta `/Audio` se encuentra el material separado según las canciones como se muestra en la siguiente estructura de carpetas: Dentro de cada canción se dividen los archivos de audio entre voces y pistas de instrumentos (`VOICES` y `STEMS`). Contiene además un archivo de audio con la mezcla de todos los instrumentos, generando el acompañamiento musical con el que se realizan los experimentos (`SongName_MIX.wav`). Un archivo en formato YAML es también incluido con información relevante sobre cada canción: compositor, productor y la identidad de cada cantante.

Anotaciones

La carpeta `/Pitch` contiene un total de 40 archivos (8 voces x 5 canciones) de texto separado por comas (CSV). Cada archivo consta de dos columnas: una con el instante de tiempo en segundos y otra con una estimación de la frecuencia

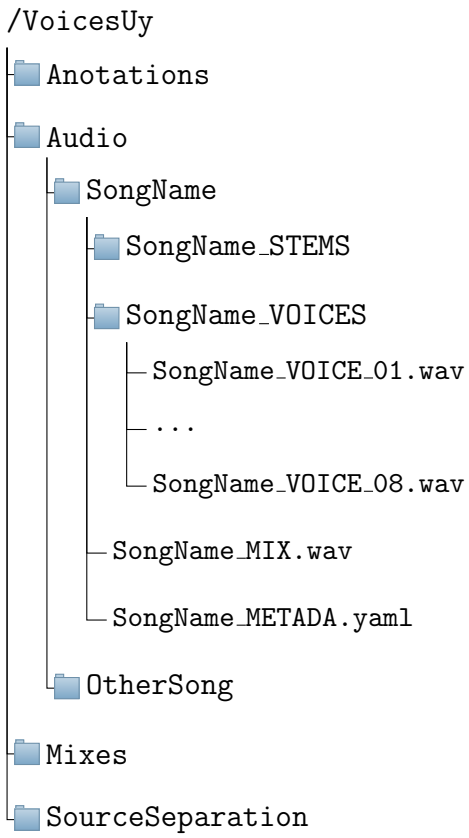


Figura 3.2: Estructura de archivos de audio en la base de datos *VoicesUy*

fundamental en Hz, la estimación es realizada cada 2.9 ms. Para generar estos archivos se utiliza el PlugIn MELODIA [30] desde el *software Sonic Visualiser* [2]. El algoritmo es aplicado sobre la pista de audio de la voz sola, de forma de disminuir errores en la estimación (ver ejemplo en figura 3.4). El parámetro de filtrado de ruido para audios monofónicos es ajustado manualmente observando la salida sobre los segmentos donde la voz está efectivamente activa.

A continuación se describen brevemente las etapas del algoritmo de estimación de la frecuencia fundamental (F0). El detalle completo está en [30].

1. Filtrado con filtro de sonoridad según curva de percepción de intensidad.
2. Cálculo de STFT con un bloque de 2048 muestras y un paso de 128. Se estiman los máximos locales de cada *frame*.
3. La frecuencia de cada máximo y su amplitud son corregidas utilizando la diferencia de fase de *frames* consecutivos.
4. Para todos los máximos locales se calcula la suma de energía de los armónicos con una suma ponderada por la diferencia entre la frecuencia exacta y los

[2] www.sonicvisualiser.org

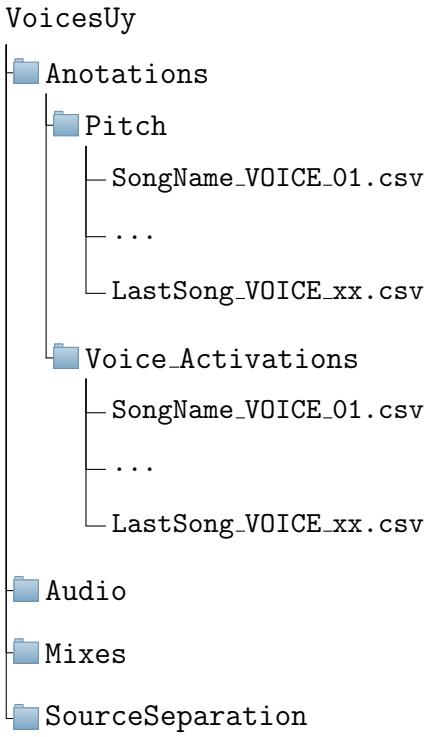


Figura 3.3: Estructura de archivos de anotaciones en la base de datos *VoicesUy*

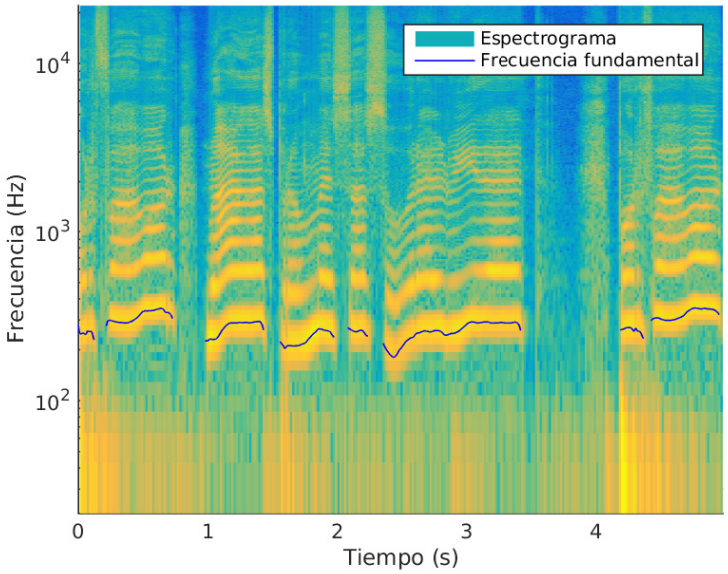


Figura 3.4: Espectrograma y F0 de un fragmento de uno de los archivos de audio de voces de *VoicesUy*

bins que disten menos de un semitono.

5. Filtrado de picos de poca predominancia.

6. Construcción de posibles contornos de F0 contemplando "continuidad" de la curva (menos de 80 cents de diferencia entre *frames* consecutivos).

7. Caracterización de las curvas de F0 utilizando media de F0, varianza de F0, predominancia, presencia de vibrato, entre otras.

8. Selección de la curva F0 teniendo en cuenta errores de octava y suavidad en el contorno de la curva F0.

La carpeta `/Voice_Activations` contiene un total de 40 archivos de texto separados por comas (CSV) indicando en qué momentos del archivo de audio hay presencia de voz. Cada archivo consta de tres columnas: la primera contiene un indicador numérico del segmento, la segunda el tiempo de inicio y la tercera el tiempo de finalización de cada activación de la voz. Para calcular los intervalos de tiempo se computa el valor cuadrático medio de la energía de cada archivo de audio de voz en ventanas de 2048 muestras (N) con un paso P de 512 muestras (ecuación 3.1), asignando un valor binario a cada *frame* según se supere un umbral de energía λ ,ecuación (3.2). El umbral es fijado empíricamente de forma de evitar los falsos positivos generados por el ruido ambiente de la sala de grabación (λ=0.006).

$$rmse[i] = \sqrt{\frac{\sum_{k=1+\cdot P}^{N+i\cdot P} s[k]^2}{N}} \tag{3.1}$$

$$Act[i] = \begin{cases} 1, & rmse[i] > \lambda \\ 0, & rmse[i] \leq \lambda \end{cases} \tag{3.2}$$

Otros archivos de utilidad

En la carpeta `/Mixes` se incluyen como referencia los audios de mezcla de voces y acompañamiento musical en 0dB de relación de energía. Contiene entonces 40 archivos *wav* con la siguiente nomenclatura `SongName_VOICE_xx.wav`.

La carpeta `/SourceSeparation` contiene todos los audios producto de realizar una separación de fuentes por un método de enmascaramiento basado en la transformación FChT que se describe en detalle en el capítulo 5. La separación es realizada sobre los audios contenidos en `/Mixes` y la estructura de carpetas es análoga a `Audio` teniendo como archivos de audio de voces el producto de la separación y como archivos de audio de acompañamiento musical el residuo de la separación de la voz.

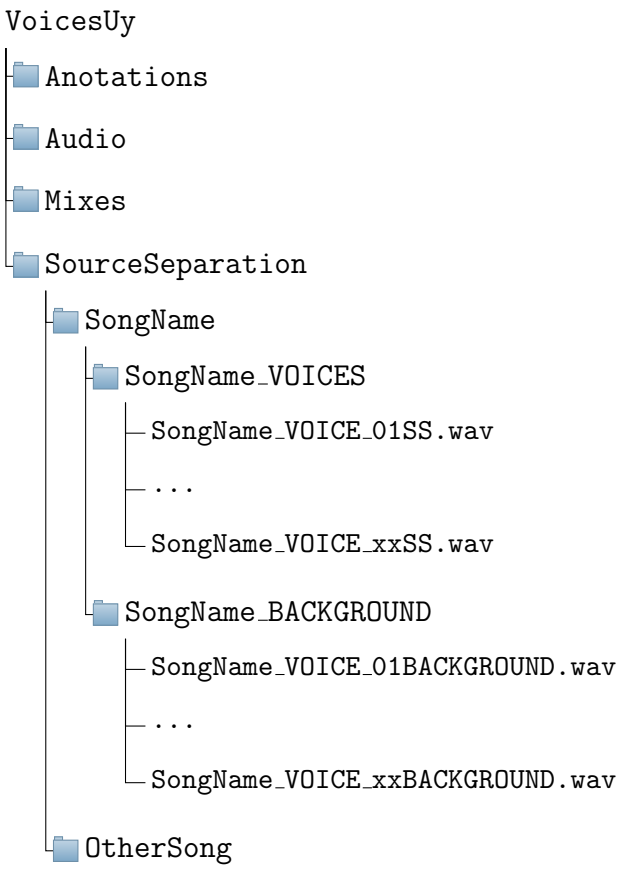


Figura 3.5: Estructura de archivos de audios producto de la separación de fuentes utilizada en este trabajo sobre la base de datos *VoicesUy*. Ver Capítulo 6.

3.3. Base de datos *AlbumsUy*

3.3.1. Descripción

Para conformar una base de datos que represente el problema real de identificación de cantantes en música comercial, se selecciona un álbum de cada uno de los cantantes de la base *VoicesUy*. En la tabla 3.2 se presentan los artistas y el álbum de referencia. Cinco canciones de cada uno de estos álbumes conforman una base a la que llamaremos *AlbumsUy*. Cada uno de los ocho álbumes cuenta con todas las etapas de producción y pos–producción. Una de las principales diferencias de esta base con la anterior es que en *AlbumsUy* cada cantante está acompañado con su banda, interpretando canciones que representan el estilo musical de la propuesta artística. La base contiene 40 canciones, cinco de cada álbum, con un total de 152 minutos y anotaciones de actividad de voz.

Tabla 3.2: Cantantes que integran las bases *VoicesUy* y contenido de *AlbumsUy*

Tag	Cantante	Banda	Álbum
VOICE_01	Nicolás Román	Los Prolijos	Rústico
VOICE_02	Federico Graña	Los Prolijos	Feria
VOICE_03	Diego Maturro	Sirilo	Mirando pa la costa
VOICE_04	Lucía Ferreira	La Tabaré	Blues de los esclavos de ahora
VOICE_05	Sebastián Gavilanes	El Gavilán	Debut
VOICE_06	Javier Zubillaga	Javier Zubillaga	Mentira
VOICE_07	Diego Rosberg	4 pesos de propina	Surcando
VOICE_08	Florencia Núñez	Florencia Núñez	Palabra clásica

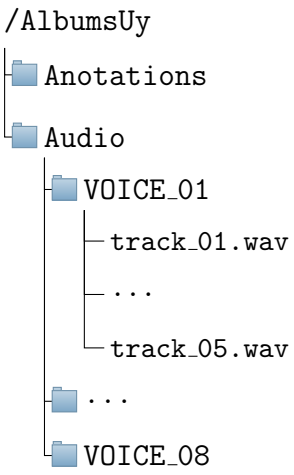


Figura 3.6: Estructura de archivos de audio de la base de datos *AlbumsUy*

3.3.2. Contenido

Audio

Dentro de la carpeta `/Audio` se encuentra el material separado según los artistas como se muestra en la siguiente estructura de carpetas. Cada carpeta de artista contiene cinco archivos de audio con la canción original del álbum de referencia de la tabla 3.2.

La selección de temas de cada álbum es realizada contemplando que el archivo contenga al menos 40 segundos de presencia de la voz del artista. Temas con invitados o instrumentales no son seleccionados para conformar la base. En la tabla 3.3 se presentan las 40 canciones que conforman la base de datos. Todos los álbumes están disponibles en plataformas digitales para ser adquiridos por cualquier interesado. Los artistas seleccionados, en su mayoría, tienen otros álbumes editados, por lo que la base puede ser ampliada para otros estudios.

Tabla 3.3: Archivos de audio de base de datos *AlbumsUy*

Álbum	Archivo	Canción	Duración
Rústico	track1.wav	Corriste	03:26
	track2.wav	Hay gente	02:52
	track3.wav	Que pase algo	03:17
	track4.wav	Zamba vieja	02:47
	track8.wav	Cuando bajen	02:40
Feria	track1.wav	Hoy	02:57
	track2.wav	Azul día gris	03:09
	track3.wav	Sesentas que no sirven para nada	03:57
	track4.wav	Cosa más linda	02:17
	track5.wav	Juan	02:38
Mirando pa' la costa	track1.wav	Vago y bohemio	04:19
	track2.wav	Pensando quién eras	03:15
	track5.wav	Villa Luján	05:57
	track6.wav	Barro	02:27
	track9.wav	Pateque	03:56
Blues de los esclavos de ahora	track2.wav	Boogie naturista	02:23
	track3.wav	Es ese misterio	05:12
	track4.wav	Distopía del blues	04:19
	track7.wav	Rasga corazón	03:33
	track9.wav	Rapsodia melodramática	12:19
Debut	track1.wav	El club de los poetas malditos	03:53
	track3.wav	Yo no quiero	04:21
	track5.wav	Algo viene volando	02:36
	track6.wav	Hombre para vos	04:49
	track7.wav	Yo me voy y no te veo	02:27
Mentira	track1.wav	Andá cortando el mazo	03:20
	track2.wav	Declaración de amor	03:50
	track3.wav	Halloween no	02:31
	track4.wav	La cosa menos horrible que me ha pasado	05:26
	track5.wav	Las galletas de arroz	04:08
Surcando	track1.wav	Caigo y me levanto	03:52
	track2.wav	Náufrago	03:39
	track4.wav	Mi revolución	04:41
	track5.wav	La fruta permitida	04:07
	track8.wav	Lara lara lara	03:30
Palabra clásica	track1.wav	Tengo un imán contigo	03:37
	track2.wav	Todo indica que caí	03:33
	track3.wav	Palabra clásica	02:12
	track4.wav	Pacto	05:47
	track5.wav	Bailo en la silla	02:34

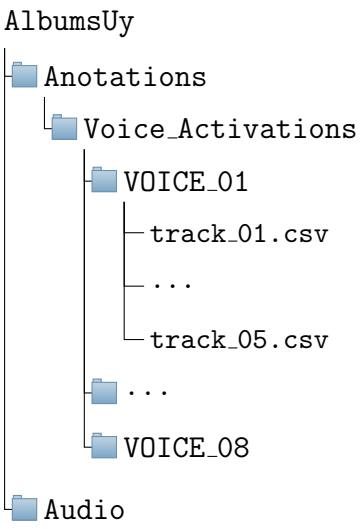


Figura 3.7: Estructura de archivos de anotaciones de la base de datos *AlbumsUy*.

Anotaciones

Cada uno de los archivos de audio de la base *AlbumsUy* es etiquetado manualmente para identificar los intervalos de tiempo donde la voz de interés está presente. Para el etiquetado se utiliza el *software Sonic Visualiser*. El archivo de salida es re-formateado para tener coherencia con la base *VoicesUy*. El formato elegido es el mismo que utiliza la base MedleyDB [58]. Cada activación es indicada por un tiempo de inicio y uno de fin (t_start,t_end). La estructura de archivos de anotaciones se presenta a continuación:

3.4. Resumen

En este capítulo se presentó el la motivación, el diseño y el contenido de las bases de datos *VoicesUy* y *AlbumsUy*. Bases de datos creadas específicamente para para estudiar la incidencia del acompañamiento musical en la identificación automática de cantantes. Las bases descritas son utilizadas en los capítulos 5 y 6 de esta tesis tanto para identificación de cantante como para evaluar técnicas de separación de fuentes. *VoicesUy* aún no está disponible mientras que las canciones que componen *AlbumsUy* pueden ser adquiridas en formato físico en Uruguay o en plataformas digitales desde los sitios oficiales de cada artista.

Capítulo 4

Modelado de la señal de voz

4.1. Introducción

Para poder automatizar con algoritmos computacionales tareas vinculadas a la percepción sonora, como la identificación de voces, es necesario extraer información relevante de las señales de audio.

El objetivo del capítulo es presentar el tipo de modelado que se hace de la señal de voz de forma de extraer características a partir de la señal de audio que sean útiles en el problema de identificación del cantante. Esto involucra aspectos de la generación de la señal de voz, así como de la percepción sonora. En suma, el modelado se implementa como una serie de coeficientes, que logran capturar gran parte de la información relevante de cada *frame* de audio.

En los años 60 se realizan varias investigaciones con base en el estudio de sistemas lineales homomórficos a la convolución. El objetivo es poder estudiar señales combinadas por convolución como señales combinadas por adición. A la representación de la señal que cumple dicha propiedad se le llama *cepstrum*. En 1967, Noll et al. utilizan el *cepstrum* para estimación de frecuencia fundamental [61] y es utilizado para análisis y síntesis de voz humana por Oppenheim en 1969 [62]. Como se verá más adelante, este tipo de representaciones permite obtener información sobre la respuesta al impulso del tracto vocal. Dentro del área de análisis del habla, se han propuesto diferentes representaciones intermedias de la señal de audio correspondiente a la voz humana, apropiadas para su análisis automático [32, 63]. Furui es el primero en notar la utilidad del *cepstrum* para verificación automática del hablante en los años 80 [64–66]. Desde entonces representaciones como MFCC, LPC y LPCC han sido ampliamente utilizadas en problemas de reconocimiento automático del habla. Estas representaciones basan su cálculo en el modelado matemático del sistema auditivo humano.

En este capítulo se presentan los mecanismos fisiológicos de generación de sonidos vocales y percepción auditiva, junto con algunos modelos matemáticos que los representan. Asimismo, se muestran las representaciones más utilizadas para la caracterización de las señales de voz.

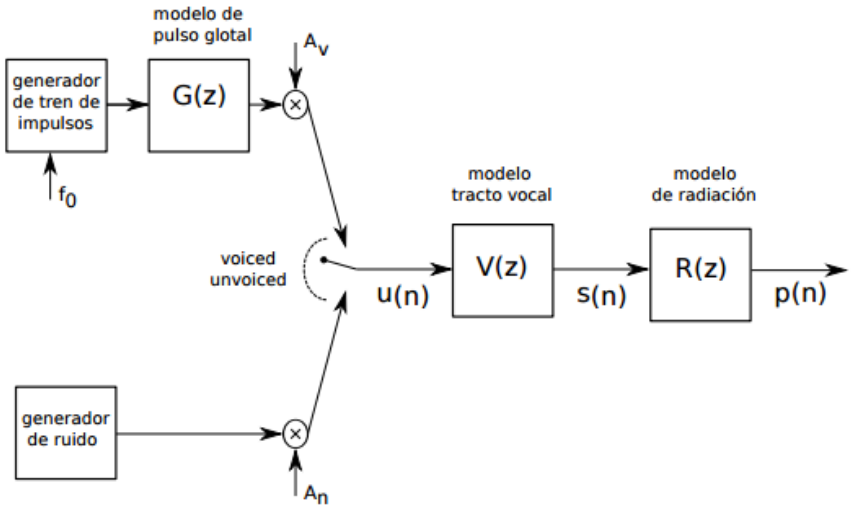


Figura 4.1: Modelo de Generación de Voz. Imagen extraída de transparencias del Curso "Procesamiento Digital de Señales de Audio"Facultad de Ingeniería, UdelaR

4.2. Modelo de generación de sonidos

En el estudio de los mecanismos de producción de voz se distingue entre sonidos sonoros y sonidos sordos. La distinción está relacionada con las características del origen de la producción de cada sonido. En el caso tonal, la excitación es generada por una vibración periódica de las membranas llamadas cuerdas vocales, mientras que en los sonidos sordos es generada por la expulsión de aire a través de una restricción. Dicha restricción se configura por la distancia entre los dientes superiores e inferiores y la posición de la lengua.
Cada fonema (unidad mínima del sonido de un idioma) se puede analizar como la salida de un sistema, donde la fuente de excitación es un tren de pulsos (en el caso tonal) o ruido de banda ancha (en el caso fricativo) convolucionado con la respuesta al impulso del tracto vocal (en la configuración de dicho fonema) como se presenta en la figura 4.1 [66].

En el caso de un fonema tonal (como lo son las vocales) donde la excitación es una señal periódica, la energía se distribuye en el espectro principalmente en la frecuencia fundamental y sus múltiplos. Mientras que en los sonidos fricativos, donde la turbulencia de aire en el tracto vocal se modela como un ruido aleatorio, la energía en el espectro se distribuye más uniformemente en un amplio rango de frecuencias. En mayor o en menor medida todos los fonemas tienen una componente de sonido sordo, dado por las restricciones de pasaje del aire propias de la configuración del tracto vocal. Es por tal motivo que el espectro de la voz en sonidos sonoros no solo tiene energía en los múltiplos de la frecuencia fundamental.

En la imagen 4.2 se pueden ver los elementos fisiológicos que intervienen en la

generación del sonido vocal. El tracto vocal, compuesto por la cavidad nasal, faringe, laringe y cavidad vocal, funciona como un conjunto de resonadores del sonido producido por la vibración de las cuerdas vocales. El sonido, al pasar por dichos conductos y cavidades modifica sus componentes espectrales. Las articulaciones del sistema (lengua, maxilar inferior, labios y paladar blando) modifican las características de los resonadores generando variaciones en el sonido producido. Las frecuencias de resonancia del sistema son conocidas como formantes y se utilizan para caracterizar los fonemas, en particular las vocales.

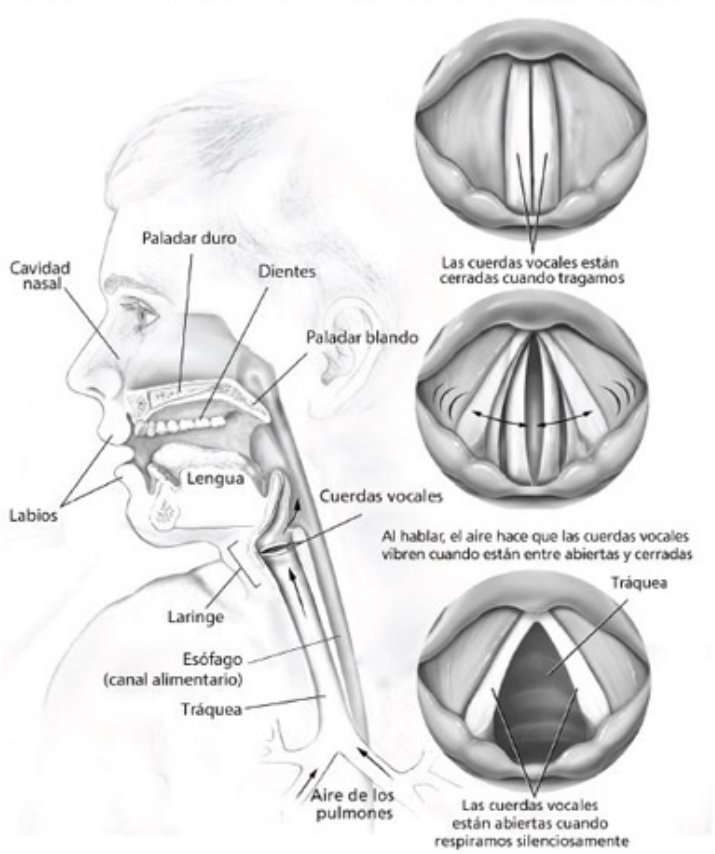


Figura 4.2: Fisiología de la generación de sonido de la voz humana. Imagen extraída del sitio web de *National Institute of Deafness and Other Communications Disorders* del Departamento de Salud de EEUU.

4.3. Modelo de percepción del sonido

Las variaciones de presión del aire generadas por fuentes sonoras son captadas por el oído humano a través del oído externo traduciéndose en oscilaciones de la membrana timpánica (tímpano). En el oído medio, una cadena de huesecillos, llamados osículos, transforma el movimiento del tímpano en variaciones de presión

del líquido (linfa) dentro del oído interno (ver figura 4.3). En el oído interno la cóclea se encarga de traducir las vibraciones generadas por el sonido en impulsos eléctricos mediante células especializadas del sistema nervioso (neuronas), denominadas células ciliadas. La cóclea está compuesta por tres canales enrollados en forma de caracol, separados por dos membranas: la membrana de Reissner y la membrana basilar. Sobre esta última se encuentra el órgano de Corti que contiene las células ciliadas. El espesor y la rigidez de la membrana basilar varían desde la base hasta el ápex (extremo final), generando que se activen diferentes neuronas (por el movimiento de las células ciliadas) según sea la frecuencia de la excitación. Las frecuencias más altas generan vibraciones sobre la parte inicial de la membrana mientras que las bajas generan mayores oscilaciones sobre el final de la membrana. Las frecuencias de resonancia de la membrana basilar se distribuyen de forma logarítmica a lo largo de su extensión (ver figura 4.4). El sistema auditivo cuenta con unas 16 mil células ciliadas sobre el órgano de Corti en contacto con el octavo nervio auditivo. La distribución de estas células junto con las características de la membrana basilar explican, en parte, una mayor resolución auditiva en bajas frecuencias.

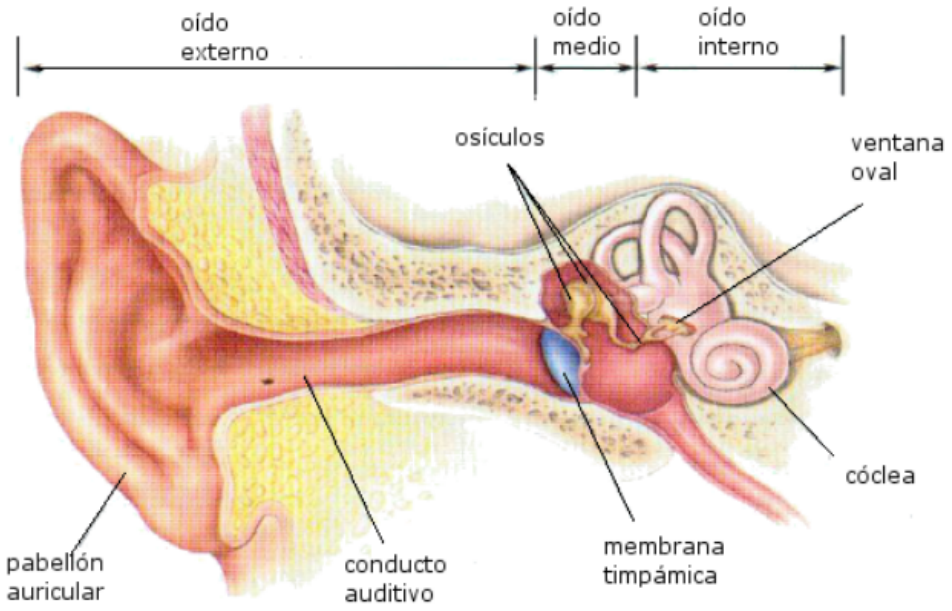


Figura 4.3: Esquema del oído humano. Imagen extraída de transparencias del Curso "Procesamiento Digital de Señales de Audio", Facultad de Ingeniería, UdelaR

Para contar con sistemas computacionales que modelen la percepción auditiva (Computational Auditory Scene Analysis (CASA)) [67] es necesario tener en cuenta la decodificación logarítmica de las frecuencias que realiza la coclea. El modelo más utilizado consiste en aplicar un banco de filtros con frecuencias centrales distribuidas logarítmicamente y respuesta al impulso de forma de simular

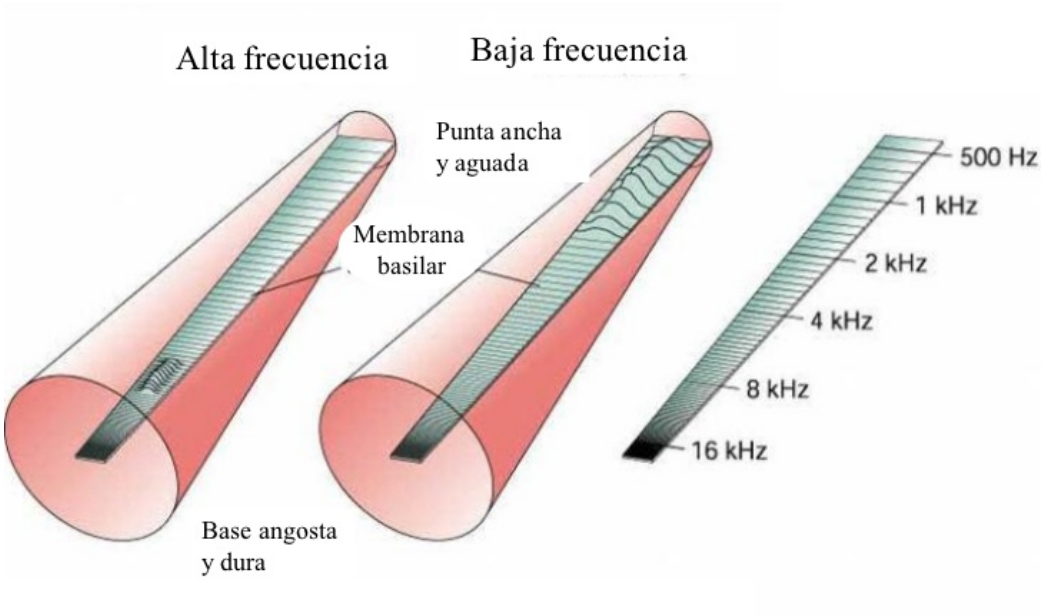


Figura 4.4: Tonotopía de la membrana basilar. Imagen extraída de transparencias del Curso "Procesamiento Digital de Señales de Audio", Facultad de Ingeniería, UdelaR.

el movimiento de la membrana basilar [68]. A continuación se presentan algunas representaciones intermedias de las señales de audio que tienen en cuenta estos principios.

4.4. Coeficientes Cepstrales

Teniendo en cuenta el modelo de generación de sonido vocal, en la identificación del cantante se puede intuir que las características biométricas más relevantes se deberían encontrar en las propiedades del filtro que modela el tracto vocal ($V(z)$ en la figura 4.1) . El filtrado es una operación de convolución en el tiempo y una multiplicación en el dominio de las frecuencias. El *cepstrum* es una operación matemática que permite ver el producto en frecuencia dado por el filtrado como una suma, a través de considerar el logaritmo de la transformada de Fourier. En la ecuación 4.1 se presenta la definición de *cepstrum* real para una señal discreta $y[n]$, es la transformada inversa de Fourier del logaritmo del espectro de la señal.

$$Cep_y(n) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log(|Y(e^{j\omega})|)e^{j\omega n}d\omega \tag{4.1}$$

$$\begin{aligned} Cep_y(n) &= \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log(|Y(e^{j\omega})|)e^{j\omega n}d\omega = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log(|X(e^{j\omega})H(e^{j\omega})|)e^{j\omega n}d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{+\pi} (\log(|X(e^{j\omega}|) + log(|H(e^{j\omega})|))e^{j\omega n}d\omega = Cep_x(n) + Cep_h(n) \end{aligned} \tag{4.2}$$

En la ecuación (4.2) se muestra el *cepstrum* de una señal $x[n]$ filtrada con un filtro con respuesta al impulso $h[n]$. En el dominio de frecuencias se puede ver la señal filtrada $y[n]$ como $Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega})$ y el *cepstrum* como una suma, como se deduce en la ecuación 4.2. El nuevo dominio si bien n es en unidades temporales se denomina cuefrencias. En el caso de señales discretas, a la representación del *cepstrum* le llamamos Coeficientes Cepstrales [69].

4.4.1. Mel-Frequency Cepstral Coefficients (MFCC)

Las características más utilizadas en el problema de reconocimiento de cantante son los Coeficientes Cepstrales en Frecuencias Mel o MFCC por su sigla en inglés. También son ampliamente utilizados en el problema de reconocimiento del habla. Desde su introducción en los años 80 por Davis y Mermelstein han sido el estado del arte en el área de clasificación de voces [32].

El diseño de los coeficientes logra codificar de forma compacta y con baja correlación información sobre el espectro de la señal. El proceso de cálculo, que se presenta en el esquema de la figura 4.6, comienza fraccionando la señal en *frames* típicamente de 20 a 40 ms utilizando una ventana de Hamming $w[n]$. Para cada *frame* se calcula la transformada discreta de Fourier, DTFT como se muestra en la siguiente ecuación

$$X_n[k] = \sum_{m=-\infty}^{+\infty} x[m]w[n-m]e^{-j\frac{2\pi k}{N}m}.$$

La ventana es de soporte acotado en el tiempo y a los efectos de optimizar el tiempo de máquina se utiliza el algoritmo FFT con largos múltiplos de potencias de 2. Los cálculos se repiten para todos los *frames* a analizar con un solapamiento temporal de 1/2 o 3/4 típicamente. A la distancia temporal entre dos *frames* consecutivos se le llama paso o salto.

El espectro resultante es entonces filtrado utilizando un banco de filtros triangulares con frecuencias centrales distribuidas logarítmicamente según la escala Mel,

$$F_{Mel} = 2595 log_{10}(1 + F_{Hz}/700)$$

. De esta forma se busca aproximar la relación entre frecuencias de la percepción de alturas del oído humano (ver figura 4.5).

Se puede ver este proceso de filtrado como computar la energía $E[l]$ en cada filtro V_l de la siguiente forma,

$$E[l] = \sum_k |V_l[k]X[k]|^2.$$

Luego se aplica la función logaritmo de forma de realzar la energía presente en altas frecuencias y poder obtener un proceso homomórfico a la convolución. Como último paso se calcula el *cepstrum* utilizando la Transformada Discreta de Coseno (Discrete Cosine Transform (DCT)) como se muestra en la siguiente ecuación,

$$MFCC[i] = \sum_{l=0}^{L-1} log(E[l]) cos\left(\frac{2\pi i}{L} l\right)$$

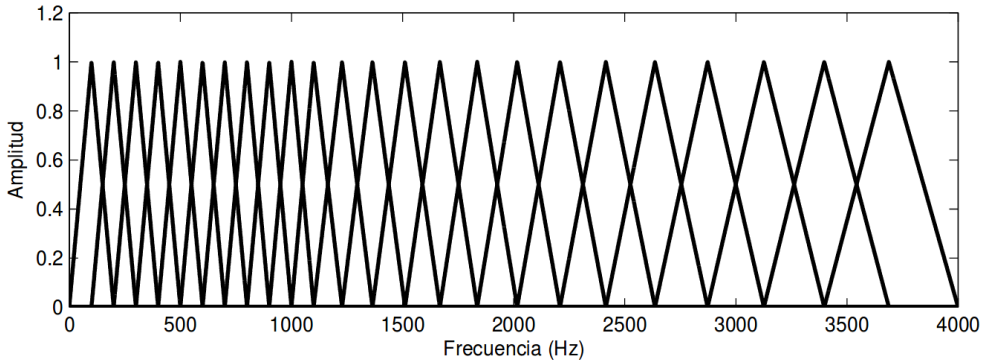


Figura 4.5: Banco de filtros sobre escala Mel

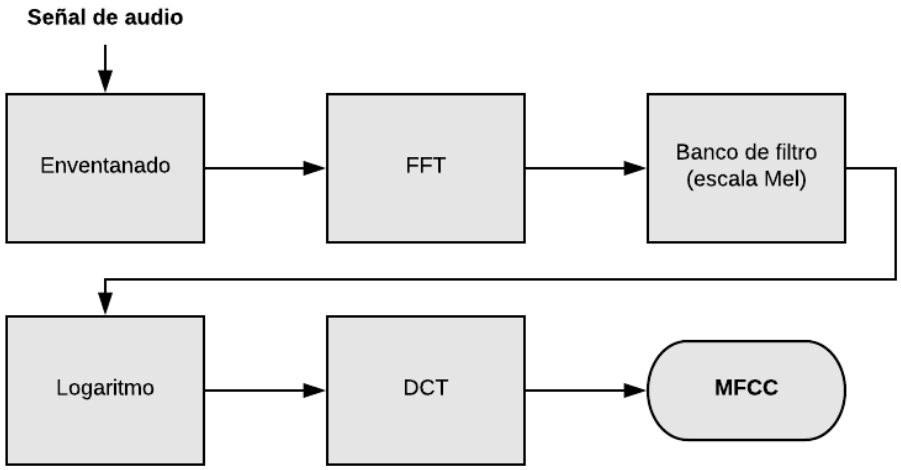


Figura 4.6: Diagrama de bloques para el cálculo de los coeficientes MFCC.

donde L es la cantidad de filtros triangulares.

La implementación utilizada en este trabajo corresponde a la presentada por Malcolm Slaney en [70] donde la escala utilizada para ubicar el banco de filtros es una aproximación de la escala Mel. Los primeros 13 filtros se centran cada 133 Hz con un ancho de banda de 133 Hz. De este modo se aproxima de forma lineal la escala Mel hasta 1 kHz y luego se distribuyen de forma logarítmica con un solapamiento de 50 % hasta completar el rango de frecuencias de interés. Para los cálculos experimentales de esta tesis se utiliza la librería de *Python librosa*[1] que implementa el mismo banco de filtros que Slaney [70].

[1]https://librosa.github.io/librosa/

4.4.2. Linear Predictive Coding (LPC)

Predicción lineal es una técnica de identificación de sistemas proveniente de la ingeniería de control. Se impone la forma del sistema H(z) como una transferencia todo polos, tal como se indica en la siguiente ecuación, y se estiman los parámetros α_k.

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}$$

Para determinar los parámetros que, dado el modelo, mejor aproximan la transferencia $H(z)$, hay que minimizar el error cuadrático medio entre la señal y su estimación. Este tipo de procesos son también conocidos como procesos auto–regresivos donde cada muestra de la señal $s[n]$ puede ser aproximada por una combinación lineal de las p muestras anteriores,

$$s[n] = \sum_{k=1}^{p} \alpha_k s[n-k] + Ge[n].$$

En el problema específico de análisis automático del habla es una de las técnicas más poderosas y de uso más extendido [66]. El objetivo de aplicar esta técnica radica en modelar la transferencia del tracto vocal $V(z)$ según el modelo ya visto en la figura 4.1. En el problema de reconocimiento de cantante esta técnica ha sido utilizada con algunas variantes que se presentan a continuación.

4.4.3. Variantes de MFCC y LPC

Linear Predictive derived Cepstral Coefficients (LPCC)

Si luego de estimar los coeficientes de un predictor lineal se realiza un análisis cepstral de dichos coeficientes, se obtienen los LPCCs [14]. En este caso, el cálculo difiere del presentado en la ecuación 4.1 siendo implementado por un proceso recursivo como se muestra en la siguiente ecuación

$$c(n) = \begin{cases} log\sigma^2, & (n = 0). \\ \alpha_n + \sum_{k=0}^{n-1}(1 - k/n)c(k)\alpha_{n-k}, & (1 \leq n \leq p). \\ \sum_{k=0}^{n-1}(1 - k/n)c(k)\alpha_{n-k}, & (n > p). \end{cases}$$

donde σ^2 representa la energía de la señal, α_n los coeficientes LPC y p el orden de la predicción lineal.

Linear Predictive Mel-Frequency Cepstral Coefficients (LPMCC)

LPMCC son los coeficientes cepstrales sobre la escala de frecuencias Mel del espectro de LPC. La forma simple de implementarlo es calcular los coeficientes MFCC del espectro de los coeficientes LPC. Se ha demostrado su eficacia en el problema de reconocimiento de cantantes cuando se tiene la voz sola o la voz separada [14, 19] ya que LPC modela el tracto vocal y el banco de filtros de los

MFCC intenta representar la distribución de la resolución en frecuencia dada por la cóclea.

Gammatone Frequency Cepstral Coefficients (GFCC)

La forma de cálculo y la motivación de estos coeficientes es similar a los MFCC, la diferencia radica en el banco de filtros. Un filtro gammatono es un filtro lineal cuya respuesta al impulso es el producto de la función gamma por una sinusoidal,

$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi f_c t + \Phi)$$

, donde f_c es la frecuencia central, Φ la fase de la portadora, a la amplitud, n el orden del filtro y b el ancho de banda.

Se ha estudiado la respuesta al impulso de la membrana basilar y su similitud con la respuesta del filtro gammatono [68] tanto a tonos puros como a la voz humana. Al igual que los filtros triangulares de las MFCC las frecuencias centrales son distribuidas logarítmicamente en el rango de frecuencias audibles. Siendo estos coeficientes, desde el punto de vista del modelado del sistema auditivo, más precisos pero no necesariamente mejores para identificar voces cantadas en música polifónica. Cai et al. [22] presentan estas características y muestran que al combinarlas con MFCC en un sistema de clasificación el desempeño mejora.

4.4.4. Resumen

En el presente capítulo se presentó el uso de coeficientes cepstrales para el modelado del tracto vocal con una motivación fisiológica. Se presentaron diferentes coeficientes cepstrales que han sido utilizados en sistemas de identificación de cantantes. En el siguiente capítulo se ve como utilizar los estos coeficientes para caracterizar una voz cantada de forma de poder predecir la identidad del cantante en un archivo de audio de música polifónica.

Capítulo 5

Identificación de cantante

5.1. Introducción

Un sistema de reconocimiento de cantante es típicamente un sistema de clasificación automática entrenado de forma supervisada. Es análogo a otros sistemas basados en biometría como lo son el reconocimiento de caras, del hablante o de huellas dactilares. Es importante notar que en la identificación biométrica se puede querer autenticar una identidad o identificar a una persona, en el primer caso es 1:1 mientras que en el segundo es 1:N. En el presente trabajo, así como en los reseñados en el capítulo 2, se trata del segundo caso, identificar en un archivo de audio nuevo, quién es la persona que está cantando de entre una cantidad finita y conocida de cantantes.

Podemos dividir a un sistema clásico de aprendizaje supervisado de forma general y resumida en las siguientes etapas:

1. Pre-procesamiento: Incluye todos los procesos necesarios a aplicar a las señales para una correcta caracterización de las clases. En este caso particular podría incluir procesos de filtrado para disminuir la incidencia del acompañamiento musical. Este punto se explora en el Capítulo 6.

2. Caracterización: Reducir la dimensionalidad de los datos generando características que permitan definir un nuevo sub-espacio donde las clases sean más distinguibles. En el problema de reconocimiento de cantante es necesario pasar de 44.100 muestras por segundo a un conjunto de datos que representen el timbre de cada voz. Varias opciones han sido estudiadas en este trabajo y fueron descritas en el capítulo 4.

3. Entrenamiento: Modelado de clases utilizando algún algoritmo de aprendizaje de máquina de forma de minimizar el error de clasificación. En el caso concreto de identificación de voces, el modelado de las características espectrales con un modelo de mezcla de Gaussianas (GMM) y la clasificación por Máxima Verosimilitud ha sido el sistema más extendido y con mejores resultados. Este se presenta en la sección 5.2.

Vale aclarar que cuando se utilizan técnicas de *Deep Learning* las etapas 1 y 2, de caracterización y entrenamiento se pueden realizar conjuntamente. Se pueden pensar las diferentes capas de una red como etapas de generación de características intermedias que representan la señal, en general reduciendo la dimensionalidad. Parte del entrenamiento de la red consiste en ajustar los coeficientes de los filtros que generan dichas características. En 2018, Wang et al. [31] presentan un trabajo donde son utilizadas redes convolucionales para extraer características de espectrogramas en escala Mel.

En este capítulo se describe el sistema de clasificación adoptado en este trabajo, fundamentos de los algoritmos y características utilizadas. Asimismo, se presentan los experimentos necesarios para mostrar la potencialidad del sistema y la dificultad que representa el acompañamiento musical en el problema de identificación automática de cantante.

5.2. Clasificación basada en Modelos de Mezclas de Gaussianas

5.2.1. Gaussian Mixture Model (GMM)

Los Modelos de Mezcla de Gaussianas (GMMs por su sigla en inglés) han tenido gran popularidad por la capacidad de aproximar distribuciones de probabilidad de formas arbitrarias. Como cualquier algoritmo de modelado, depende fuertemente de la cantidad y calidad de los datos. En particular para el problema de identificación de cantante han sido el estado del arte durante los últimos quince años. Un modelo GMM para una función de probabilidad de una variable aleatoria x se define como la suma ponderada de distribuciones normales multi-variadas como,

$$p(x) = \sum_{n=1}^{N} \omega_n \mathcal{N}(x; \mu_n, \Sigma_n), \tag{5.1}$$

donde N es la cantidad de gaussianas o componentes, ω_n es el peso del componente n, μ_n el vector de valores medios de la componente n de la Normal multi-variada y Σ_n su matriz de covarianza.

Para ajustar los parámetros (μ_i, Σ_i) se maximiza la verosimilitud usando el método Expectation Maximization (EM). EM es un método iterativo para encontrar un máximo local de la verosimilitud de una distribución estadística, dado un conjunto de datos. En primer lugar, se inicializan los parámetros GMM utilizando un algoritmo de *clustering* como *k-means* [71]. Luego se comienza a iterar entre los pasos E (Expectation) y M (Maximization). En la etapa E, se calcula la esperanza de la log-verosimilitud dada la estimación de parámetros del paso i, se supone un conjunto de características faltantes Z y se calcula $Q(\theta; \theta^i)$ con θ^i fijo,

$$Q(\theta; \theta^i) = E_Z[log\, p(X, Z; \theta)|X; \theta^i]$$

. En el paso M se busca el máximo de $Q(\theta; \theta^i)$, el conjunto de parámetros encontrado será la entrada del siguiente paso E, como se muestra en el pseudo-código

```
1 begin initialize θ^0, T, i = 0
2         do i ← i + 1
3             E step : compute Q(θ; θ^i)
5             M step : θ^{i+1} ← arg max_θ Q(θ; θ^i)
6         until Q(θ^{i+1}; θ^i) − Q(θ^i; θ^{i−1}) ≤ T
7   return θ̂ ← θ^{i+1}
8 end
```

Figura 5.1: Pseudocódigo de algoritmo EM. Extraído del libro Pattern Classification [73]

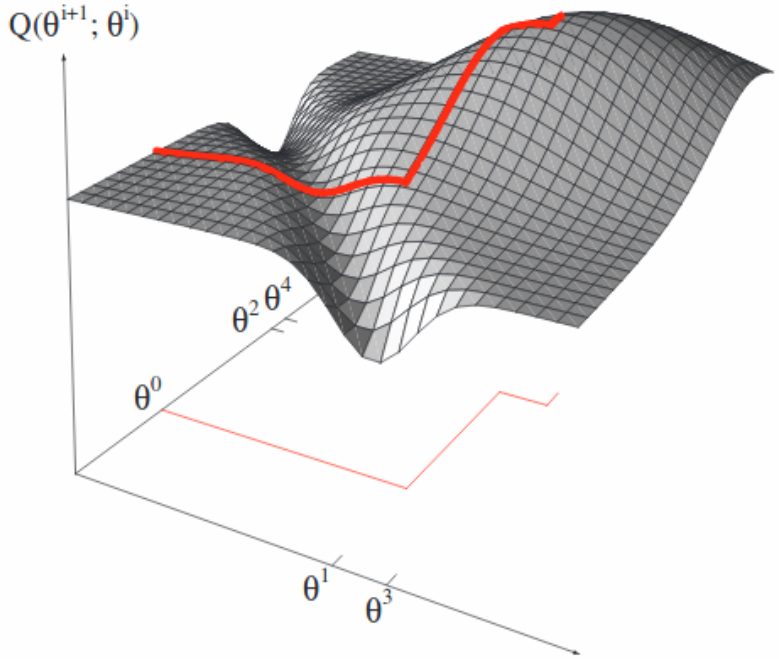


Figura 5.2: Algoritmo EM. Imagen extraída del libro Pattern Classification [73]

de la figura 5.1. En la figura 5.2 se ve un ejemplo de búsqueda de máximo local en un problema de dos parámetros con EM. Una explicación detallada del uso de EM como método de estimación de parametros aplicado a GMM se puede encontrar en el libro *Machine Learning: A Probabilistic Perspective*, de Kevin P. Murphy [72].

5.2.2. Sistema de clasificación de cantante

En la figura 5.3 se presentan de forma esquemática los procesos involucrados en el sistema de clasificación de cantantes. A continuación se describen las etapas de entrenamiento y clasificación quedando el módulo de prepocesamiento basado en separación de fuentes para ser analizado en el Capítulo 6.

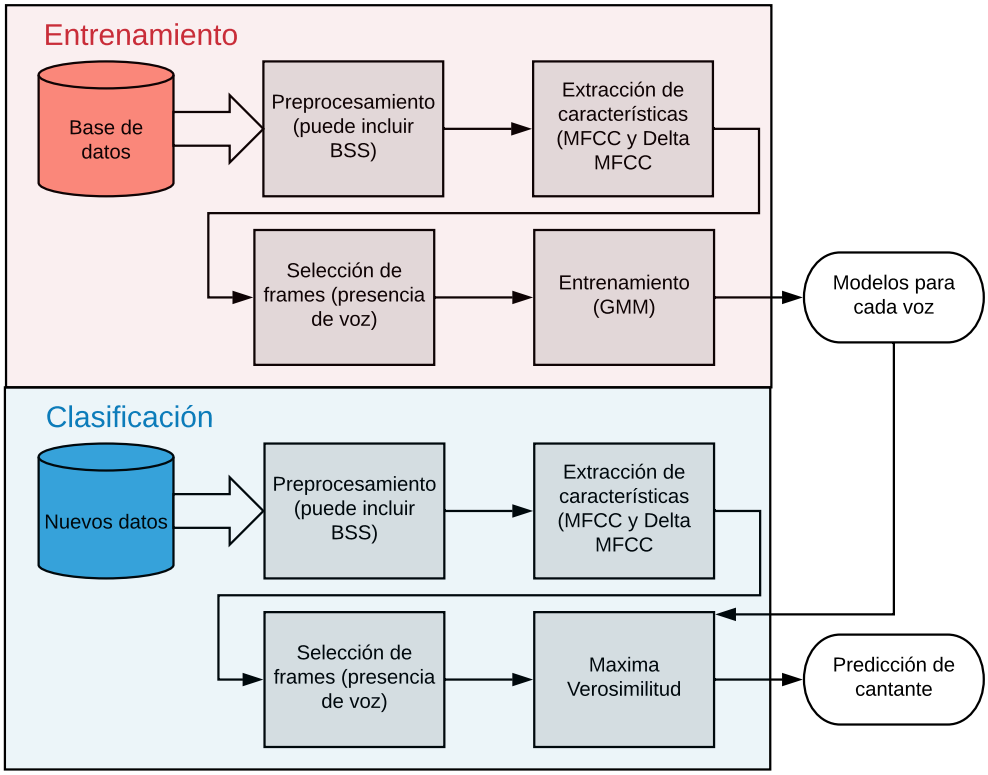


Figura 5.3: Esquema de sistema de clasificación de cantante basado en GMM.

Extracción de características

Diferentes combinaciones de las características presentadas en el Capítulo 4 han sido utilizadas para clasificar voces cantadas [3,14,18–20,22,28]. En este trabajo se utilizaran los coeficientes MFCC. El primer coeficiente es descartado por ser una medida de la energía de la señal y se utilizan los siguientes 19 coeficientes con las derivadas de primer orden de su evolución temporal, para capturar información sobre la variación de los coeficientes en el tiempo.

Selección de *frames*

Uno de los pre-procesamientos más importantes consiste en seleccionar solo aquellos *frames* de los archivos de audio en donde la voz está presente. Con tal objetivo, varios trabajos presentan una clasificación automática en un problema de dos clases entrenando un modelo GMM [14, 19, 20], otros lo hacen de forma manual utilizando una base etiquetada [25, 28]. También la selección de *frames* se puede dar dentro de un proceso de separación de fuentes [74]. En algunos trabajos de clasificación de cantantes este punto es trivial ya que se trabaja con archivos de audio de voces sin acompañamiento musical [29, 31].

En este trabajo se realiza la selección de *frames* utilizando las anotaciones presencia de voz de la base de datos *VoicesUy*, como se presenta en el capítulo 3.

Clasificador GMM

Un sistema de clasificación basado en GMM implica el ajuste de un modelo para cada clase (cada voz en este caso). Otros algoritmos de clasificación clásicos como, SVM o árboles de decisión, en su etapa de entrenamiento determinan fronteras en un sub-espacio dado por las características (*features*) de forma de minimizar el error de clasificación de un conjunto de datos de validación. En el caso de GMM los modelos para cada clase se ajustan independientemente utilizando solo datos de entrenamiento de la clase que se está entrenando. Para clasificar nuevas muestras se selecciona aquella clase que, dado su modelo, maximiza la verosimilitud de los datos. El uso del logaritmo de la verosimilitud permite sumar el aporte de cada muestra y por ser una función monótona creciente mantiene la ubicación de los máximos locales inalterada, es decir:

$$voz_{pred} = \underset{i}{\operatorname{argmax}} \frac{1}{N} \sum_{n=1}^{N} \log\, p(x_n|\lambda_i) \tag{5.2}$$

donde λ_i son los parámetros de la distribución de probabilidad de la clase i y N es la cantidad de frames dentro del intervalo de clasificación T.

5.3. Experimentos de clasificación con GMM y máxima verosimilitud

En esta sección se presentan experimentos sobre la base de datos *VoicesUy* descrita en el capítulo 3, con el objetivo de mostrar el efecto del acompañamiento musical en la clasificación de voces cantadas. Asimismo, se analiza la relación entre la verosimilitud de los datos de test y características del sonido. Se plantea una variante al sistema de clasificación de máxima verosimilitud que permite mejorar los resultados.

5.3.1. Metodología

Todos los experimentos son realizados en validación cruzada. Para clasificar las ocho voces se utilizan cuatro de las cinco canciones para entrenar y la quinta canción como validación. Esto se repite cinco veces teniendo un sistema de clasificación con validación cruzada de cinco particiones (*5-Folds*). En los experimentos realizados también se considera el problema de identificar la canción a partir de la interpretación de otros cantantes. En este caso el sistema también se entrena con ocho particiones (*8-Folds*).

Respecto a la clasificación de un intervalo de tiempo T, en este trabajo se evalúa el uso de la ecuación 5.2, que un método ampliamente utilizado. En varios de los trabajos reseñados T corresponde al largo total de cada canción de test [19, 20], otros utilizan intervalos de tiempo en la validación de entre 4 y 60 segundos [3, 14, 18, 22]. Lagrange [25] plantea fragmentar las canciones de validación en intervalos de 10 segundos de forma de tener más muestras para reportar resultados con

mayor resolución. De forma similar Wang et al. [31] dividen los datos de audio en fragmentos de 6 segundos con solapamiento para generar los espectrogramas a la entrada de la red CNN. El único trabajo que analiza diferentes duraciones (T) de validación es el de Tsai [20] donde los resultados usando 10 segundos son notoriamente inferiores comparados con usar la canción completa.

Los experimentos que aquí se presentan fueron realizados con intervalos de validación variando entre 100 ms y 40 s, con intervalos de tiempo solapados en las canciones de validación de forma de aumentar la cantidad de muestras para el reporte de resultados. De esta forma se puede tener una predicción variable a lo largo de la canción, lo que hace que el método pueda ser utilizado para archivos de audio donde la voz principal se alterna entre cantantes. Este análisis aporta información relevante para determinar la cantidad de datos a analizar cuando se trabaja con bases de datos de gran porte, por ejemplo, se podrían utilizar solo 15 segundos cualesquiera en lugar de los 3 minutos que puede durar una canción.

Para todos los experimentos de clasificación presentados se utiliza el clasificador GMM con 32 gaussianas por mezcla, al igual que en el trabajo de Lagrange [25]. Para el algoritmo EM se utiliza un máximo de 100 iteraciones. Los coeficientes MFCC del 2 a 20 y sus derivadas de primer orden modelan cada *frame*. El tamaño de la ventana de análisis es de 1024 muestras (23.2ms) y un paso de 512 muestras (11.6ms). El banco de filtros utilizado alcanza una frecuencia máxima de 10 kHz. De esta forma cada *frame* de audio de 23.2ms queda representado por un vector de 38 coeficientes.

5.3.2. Clasificación de voces

Una de las hipótesis planteadas en este trabajo es que la relación de energía del acompañamiento incide negativamente en la clasificación de voces. Para verificar dicho comportamiento y cuantificarlo se realiza la clasificación de las ocho voces de la base *VoicesUy* con cuatro niveles diferentes de mezcla (-3dB, 0dB, +3dB y +6dB). En este trabajo se utiliza la definición de Signal to Noise Ratio (SNR) para referirnos a la relación de energía entre la señal de la voz y el acompañamiento musical[1], que se define como

$$SNR = 10log_{10}\frac{||s_{voz}||^2}{||s_{acomp}||^2}.$$

Con los siguientes experimentos se pretende evaluar el efecto sobre la clasificación de dos aspectos diferentes: la duración T de los intervalos de evaluación y el nivel de mezcla SNR. En la figura 5.4 se presentan las matrices de confusión obtenidas al clasificar intervalos de 40 segundos con niveles de mezcla SNR de -3dB y +3dB, en validación cruzada. Se puede ver claramente cómo al aumentar la energía relativa de la voz la matriz se vuelve diagonal pasando de un acierto de 64.7 % a 97.6 %.

[1]En algunos trabajos se utiliza la sigla SAR, que en este trabajo será utilizada para definir métricas de calidad de separación de fuentes.

Tabla 5.1: Porcentajes de acierto en clasificación de voces para diferentes niveles de mezcla y diferentes intervalos de evaluación.

SNR(dB)	Intervalo de test (s)						
	0,1	0,5	1	2	5	10	40
-3	24,7 %	32,0 %	37,4 %	43,5 %	51,2 %	56,8 %	64,7 %
0	34,2 %	46,0 %	54,2 %	62,1 %	71,0 %	76,8 %	**85,8 %**
3	44,4 %	58,9 %	67,3 %	75,1 %	82,7 %	89,0 %	97,6 %
6	53,9 %	68,8 %	76,4 %	82,5 %	88,5 %	92,9 %	98,6 %

Sobre el efecto que la duración de los intervalos de evaluación tiene sobre el desempeño, en la tabla 5.1 se presentan los resultados para intervalos de tiempo desde 0.1 a 40 segundos. En el caso de una mezcla con SNR=0dB el cambio de pasar de 0.1 a 40 segundos, es de 34.2 % a 85.8 %. En dicha tabla se puede ver también, que el acierto en la clasificación es una función monótona creciente respecto a la duración del intervalo de clasificación T para todas las SNR presentadas. En la figura 5.5 se presenta la evolución del acierto en la clasificación en función de T para cada voz, la figura 5.5a presenta los resultados para un SNR=-3dB y la figura 5.5b para SNR=+3dB.

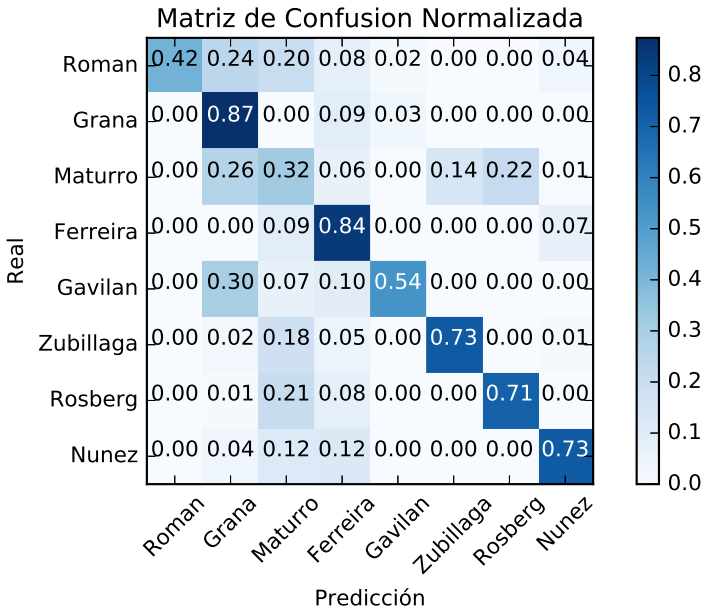
A efectos prácticos se toma como valor de referencia del problema real, una SNR de 0dB. En la figura 5.6a se presenta la matriz de confusión de la clasificación de voces para un SNR de 0dB. El acierto para la evaluación en validación cruzada es de 85.8 % cuando se evalúa con muestras de 40 segundos de duración. En la figura 5.6b se presenta el acierto para cada voz en los intervalos de tiempo T ya mencionados.

5.3.3. Clasificación de canciones

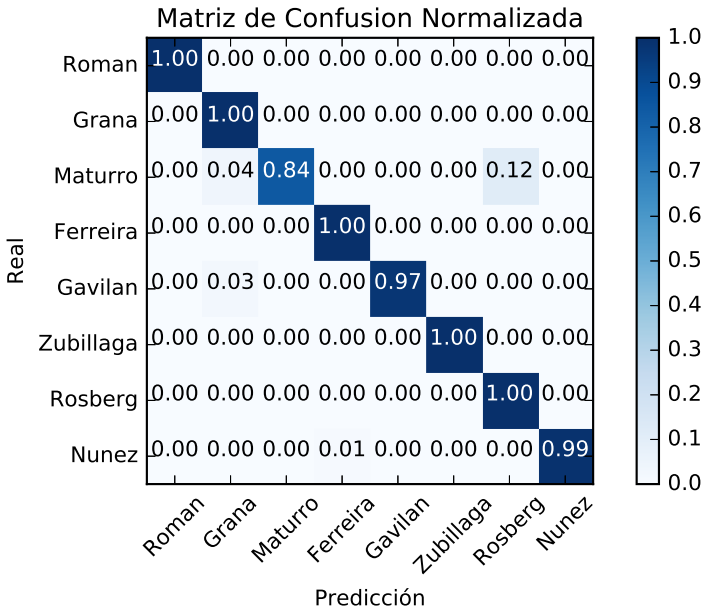
Los audios que componen la base *VoicesUy* no tienen pos-procesamiento ni masterización, por lo que las componentes espectrales de las voces solo tienen las modificaciones dadas por el procesos de grabación que fue exactamente el mismo para todas. De esta forma, el "efecto pos-producción" no está presente. Además, al utilizar canciones de diferentes compositores y con diferentes instrumentaciones el "efecto banda" tampoco está presente.

Un experimento que permite estudiar un caso extremo del "efecto banda" es identificar una canción con el mismo acompañamiento musical independientemente de quién la esté cantando. Además, se compara el resultado con la clasificación de canciones a partir de las voces a capella. De alguna manera se invierte el problema. Se pasa de intentar reconocer al cantante independientemente de la canción que cante, a identificar la canción independientemente de quién la canta.

Al igual que en los experimentos de la sección anterior, se presentan los resultados para diferentes niveles de mezcla (SNR) y diferentes largos de intervalos de clasificación (T) sobre los archivos de audios de test. Todos los resultados se

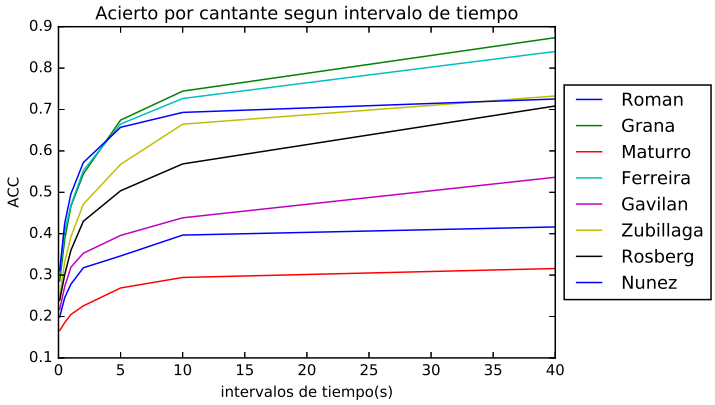


(a) Matriz de Confusión SNR=-3dB, T=40s

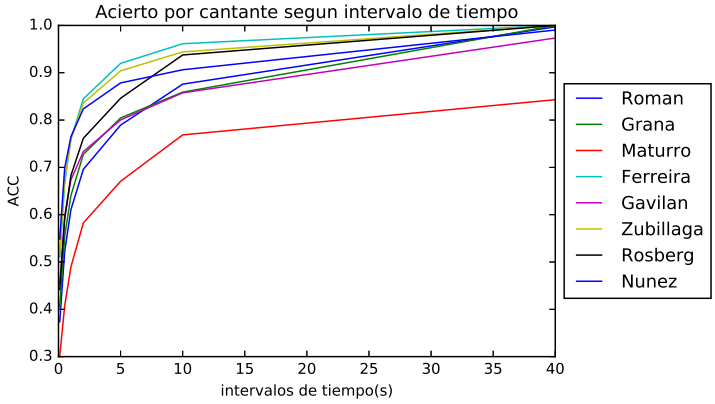


(b) Matriz de Confusión SNR=+3dB, T=40s

Figura 5.4: Matrices de confusión de clasificación de voces en validación cruzada para diferentes relaciones de energía del acompañamiento musical (SNR)



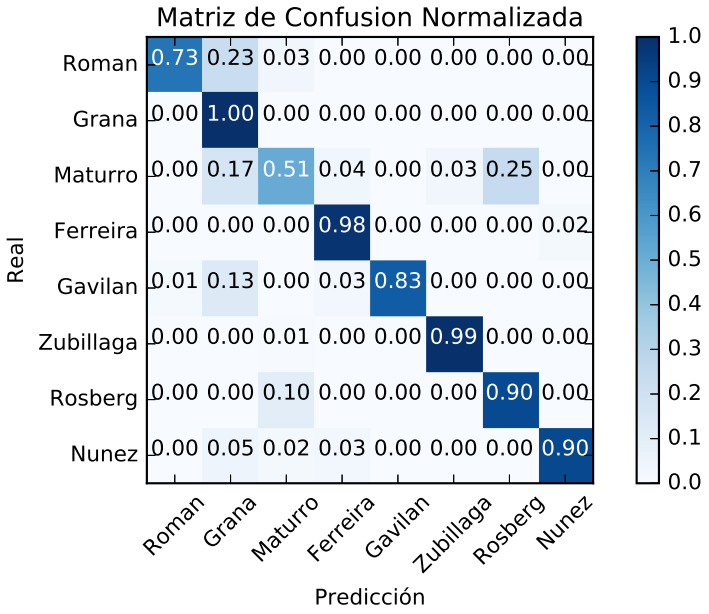
(a) Acierto vs Intervalo de tiempo SNR=-3dB



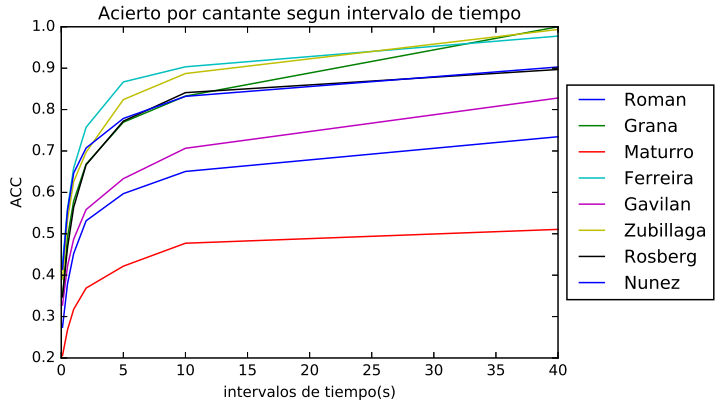
(b) Acierto vs Intervalo de tiempo SNR=3dB

Figura 5.5: Matriz de confusión de clasificación de voces en validación cruzada

obtienen realizando validación cruzada de cinco particiones (una por canción de test) y se presentan en la tabla 5.2. Se puede observar que con cinco segundos cualesquiera de una canción se puede identificar con un 99.8 % (en promedio) de acierto a qué canción corresponde, independientemente de quién esté cantando. A medida que el SNR aumenta el acierto disminuye, el caso extremo es clasificar la canción solo con las voces. A priori se podría pensar que, por las características de los coeficientes cepstrales, la melodía no debería modificar significativamente la energía en las bandas de frecuencia y por ende los valores de las características utilizadas. Sin embargo, más allá de las variaciones en las melodías, la letra y la intención de cada fonema sí está vinculado directamente a cada composición. En la tabla 5.3 se presentan los resultados de clasificar las canciones utilizando los archivos de audio de las voces solas. También se presentan los resultados de clasificar las voces entrenando también con los mismos archivos de audio, el caso ideal de clasificación de voces. Se puede ver que para el caso de utilizar 10 segundos de duración en los datos a clasificar, se pueden identificar las canciones con un 70.0 % de efectividad contra un 20.0 % de forma aleatoria.



(a) Matriz de confusión SNR=0dB, T=40s



(b) Acierto vs Intervalo de tiempo SNR=0dB

Figura 5.6: Clasificación de voces en validación cruzada con SNR=0dB para intervalos de evaluación de 40 segundos.

Tabla 5.2: Clasificación de canciones realizada con GMM de 32 gaussianas para diferentes niveles de mezcla (SNR) y diferentes intervalos de tiempo de los segmentos de test.

SNR(dB)	Intervalo de test (s)						
	0,1	0,5	1	2	5	10	40
-3	93,2 %	98,1 %	99,2 %	99,7 %	100,0 %	100,0 %	100,0 %
0	86,7 %	95,1 %	97,4 %	98,8 %	99,8 %	100,0 %	100,0 %
3	77,4 %	88,4 %	92,6 %	95,3 %	97,6 %	99,0 %	100,0 %
6	68,5 %	80,5 %	85,9 %	89,7 %	93,3 %	94,8 %	96,8 %

Tabla 5.3: Clasificación de voces y canciones solo utilizando los audios de las voces a capella para diferentes intervalos de tiempo de evaluación en validación cruzada de 5 y 8 particiones respectivamente.

Indentificar	Intervalo de test (s)						
	0,1	0,5	1	2	5	10	40
cantante	68,3 %	80,7 %	85,9 %	89,2 %	91,7 %	93,2 %	96,4 %
canción	44,3 %	52,0 %	56,9 %	61,4 %	67,2 %	70,0 %	74,2 %

Se explorará con más profundidad el "efecto banda" al clasificar canciones de un mismo álbum en la sección 5.5.

5.4. Análisis de resultados y método alternativo de clasificación

5.4.1. Introducción

En muchas ocasiones, los algoritmos de reconocimiento de patrones generan modelos que son tratados como una caja negra. Dependiendo del problema puede ser difícil saber qué es exactamente lo que están encontrando o qué tipo de datos son los causantes de los errores de predicción. En esta sección se explora la vinculación entre la variación temporal del contenido armónico de los archivos de audio y la log–verosimilitud de los datos. Se propone además una variación al algoritmo de clasificación presentado en 5.2.2.

5.4.2. Análisis de la log–verosimilitud como parámetro de clasificación

Como se ve en la ecuación 5.2, la forma de definir cuál es la voz correspondiente a un audio desconocido es determinando el máximo del promedio del logaritmo de las verosimilitudes sobre cada modelo de entrenamiento. Esto tiene un supuesto implícito, no todos los *frames* tienen el mismo peso a la hora de definir la clasifica-

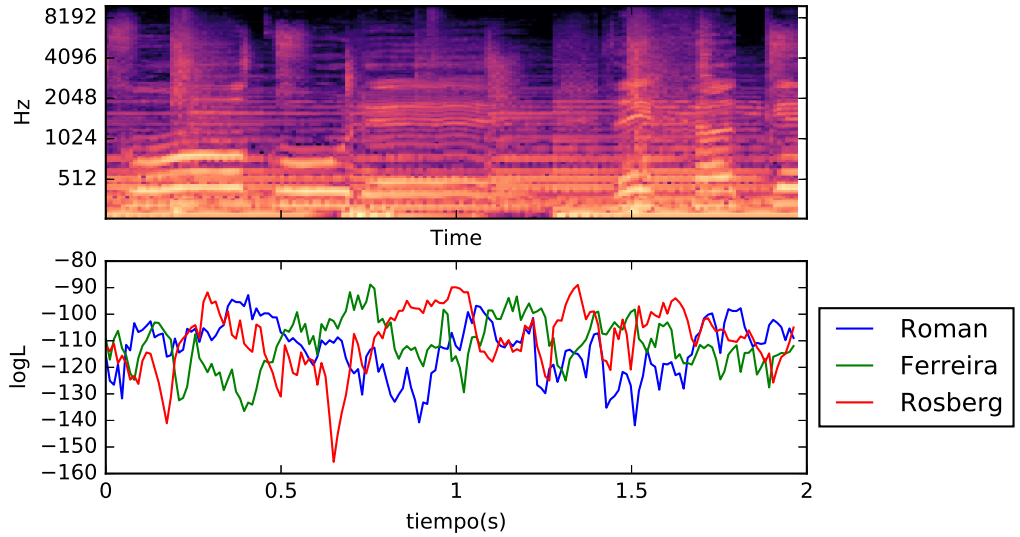


Figura 5.7: Log–verosimilitud de tres voces y espectrograma de la voz correcta (Rosberg). Fragmento de audio de la base *VoicesUy*

ción. Casos extremos pueden cambiar significativamente la decisión. En la figura 5.7 se muestra un fragmento de dos segundos con la log-verosimilitud de tres voces sobre un modelo (el correcto es el gráfico rojo). Observando la figura no resulta claro que realizar el promedio sea la mejor estrategia de clasificación. Cabe la pregunta de si existe alguna vinculación entre los *frames* donde existe más acierto en la clasificación y el nivel de armonicidad del audio en dicho *frame*.

Con el objetivo de evaluar la incidencia del nivel de armonicidad de cada *frame* en los resultados de clasificación, se analiza la relación entre la función de verosimilitud y la función Gathered Log Spectrum (GlogS).

GlogS es una de las técnicas utilizadas para determinar la frecuencia fundamental de una fuente armónica. Se calcula como el promedio del logaritmo de la magnitud del espectro en posiciones armónicas (ecuación 5.3). A efectos prácticos se computa el máximo de la función GlogS en un rango de frecuencias de 100 a 500 Hz con un paso de 10 Hz interpolando los valores de la STFT.

$$\rho_n(f_0) = \frac{1}{n_H} \sum_{i=1}^{n_H} log|X(if_0)| \tag{5.3}$$

En la tabla 5.4 se presenta el coeficiente de correlación de Pearson entre la función GlogS y la log-verosimilitud para cada una de las voces del ejemplo de la figura 5.7. Se puede ver que existe una correlación de bajo nivel por lo que no se puede concluir que los *frames* de mayor contenido armónico sean los más determinantes para clasificar correctamente las voces.

Tabla 5.4: Correlación de Pearson entre GlogS y log-verosimilitud

	Voz_1	Voz_2	Voz_3
ρ^2	0.22	0.16	0.19

Tabla 5.5: Porcentajes de acierto en clasificación de voces para diferentes niveles de mezcla y diferentes intervalos de evaluación para *VoicesUy*. Voto por mayoría (moda).

SNR(dB)	Intervalo T (s)						
	0,1	0,5	1	2	5	10	40
-3	22,8 %	30,3 %	35,8 %	42,2 %	51,3 %	58,3 %	66,1 %
0	31,4 %	43,3 %	51,6 %	60,1 %	70,3 %	76,6 %	87,7 %
3	41,5 %	56,7 %	65,8 %	74,3 %	83,6 %	89,3 %	98,5 %
6	50,4 %	66,8 %	75,4 %	82,7 %	89,2 %	94,1 %	99,6 %

5.4.3. Variante en método de clasificación

Se propone realizar la clasificación por votación por mayoría, de esta forma todos los *frames* pesan lo mismo y se evita la incidencia de valores extremos que afectan al promedio. El voto por mayoría en un intervalo T no es otra cosa que la moda de la clase con la máxima verosimilitud en cada *frame*,

$$voz_{pred} = \underset{n=n_k}{\overset{n_k+N}{\text{MODA}}} \left(\underset{i}{\text{argmax}} \left(\log p(x_n|\lambda_i) \right) \right), \tag{5.4}$$

donde N es la cantidad de frames dentro del intervalo de tiempo T, n_k el comienzo del intervalo k-ésimo y λ_i son los parámetros de la distribución de probabilidad de la clase i.

En la tabla 5.5 se presentan los resultados para los experimentos de clasificación de voces de la base *VoicesUy* para diferentes niveles de mezcla (SNR) y diferentes intervalos de tiempo de evaluación (T). Los experimentos son los mismos que los realizados en la sección anterior, pero en lugar de utilizar el promedio de la log–verosimilitud (ecuación 5.2) se clasifica cada *frame* y luego se computa la moda (ecuación 5.4). Comparando las tablas 5.1 y 5.5 se puede ver que se obtiene una mejora del orden del 2 % cuando el intervalo de clasificación es superior a 10 s, lo que puede considerarse un aumento significativo para este problema.

5.5. Clasificación de álbumes

Hasta este punto los experimentos de clasificación de voces son realizados con la base *VoicesUy*, de forma de que la única diferencia entre los audios a clasificar fueran las voces (los acompañamientos musicales son los mismos para cada cantante). Para evaluar el "efecto banda" se utiliza la base *AlbumsUy*. En dicha base, el acompañamiento musical de cada voz es completamente diferente, tanto las composiciones como la instrumentación y efectos, responden al estilo propio del artista.

Tabla 5.6: Comparación de clasificación de voces sobre las bases de datos *VoicesUy* y *AlbumsUy* paa ambos métodos de clasificación.

Método	Base	Intervalo T(s)						
		0,1	0,5	1	2	5	10	40
Moda	*VoicesUy* (0dB)	31,4 %	43,3 %	51,6 %	60,1 %	70,3 %	76,6 %	**87,7 %**
	AlbumsUy	45,4 %	60,7 %	68,0 %	74,2 %	81,0 %	85,3 %	**91,4 %**
Media	*VoicesUy* (0dB)	34,2 %	46,0 %	54,2 %	62,1 %	71,0 %	76,8 %	85,8 %
	AlbumsUy	50,5 %	63,7 %	69,8 %	75,1 %	80,8 %	83,9 %	89,1 %

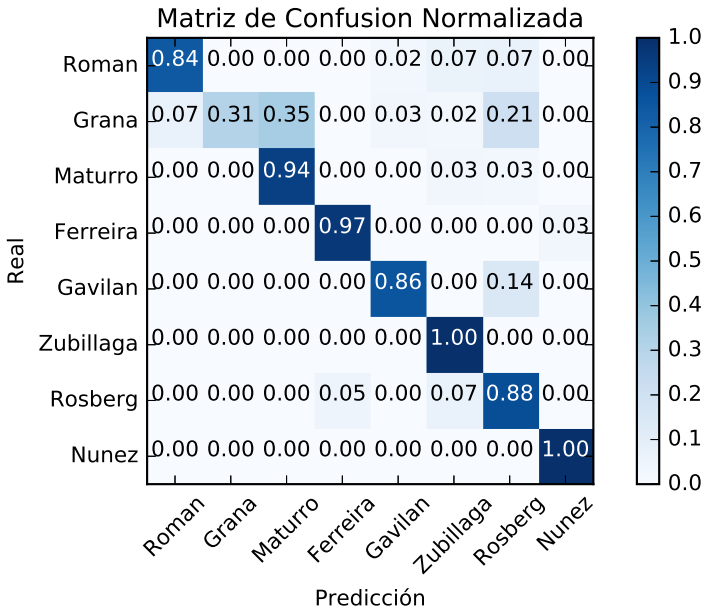
Lo que se quiere analizar es si estas variables dificultan o facilitan la identificación del cantante. Como ya se vio en los experimentos previos el acompañamiento musical disminuye el desempeño del algoritmo de clasificación automática. ¿Qué incidencia tiene el "efecto banda"? Para responder esta pregunta se realiza un experimento análogo al presentado en 5.3.2, clasificar las 8 voces utilizando 5 canciones por cantante (todas del mismo álbum) en validación cruzada (*5-Folds*).

La clasificación de voces sobre la base *AlbumsUy* fue realizada utilizando la forma clásica de clasificación, con la media de la log-verosimilitud y el método propuesto de la moda de cada intervalo. Las matrices de confusión se presentan en la figura 5.8. El acierto de clasificación es de 89.1 % y 91.4 % (media y moda, respectivamente). Estos resultados son superiores a los obtenidos con las mismas voces sobre la base *VoicesUy*, lo que sugiere que el "efecto banda" contribuye a una mejor identificación del cantante. En la tabla 5.6 se ve claramente que el acierto en la clasificación de voces sobre la base *AlbumsUy* es superior al mismo algoritmo corriendo sobre la base de datos *VoicesUy*.

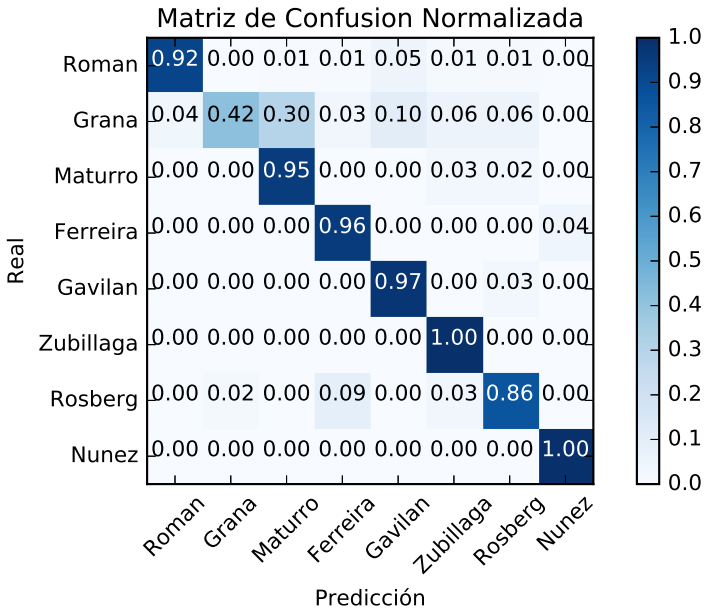
5.6. Discusión

En este capítulo se describió un sistema de clasificación de voces basado en la extracción de coeficientes cepstrales de frecuencia Mel MFCC y el modelado por mezcla de gaussianas GMM. Dicho método ha sido ampliamente utilizado en trabajos previos [19, 22, 24, 25, 28, 74]. Además, se analizó la incidencia del nivel de energía del acompañamiento musical en la clasificación de voces, mostrando una mejora de más de 10 % al reducir 3 dB el SNR de la mezcla (ver tabla 5.1). Habiendo detectado que la incidencia de la duración de los intervalos de audio a clasificar era un punto poco explorado previamente, todos los experimentos fueron realizados para siete intervalos de tiempo diferentes. Los resultados muestran una muy fuerte dependencia del acierto en la clasificación con el largo del intervalo a clasificar. La diferencia en la tasa de acierto entre utilizar 10 s o 40 s es en promedio mayor al 10 % de acierto.

Basado en la idea de que cada *frame* del intervalo de audio a ser clasificado tenga el mismo peso, se plantea una variante al sistema de clasificación al utilizar votación por mayoría clasificando *frames*. Se mostró que este método mejora los



(a) Matriz de confusión para método tradicional



(b) Matriz de confusión para método propuesto

Figura 5.8: Clasificación de voces de la base *AlbumsUy* en validación cruzada para intervalos de evaluación de 40 segundos. Comparación de métodos.

resultados de clasificación tanto sobre la base *VoicesUy* como sobre la base de datos *AlbumsUy* (ver tabla 5.6).

Con el objetivo de mostrar que el acompañamiento musical, si bien dificulta la identificación de cantante, aporta información relevante para el problema, se realizaron dos experimentos. El primero consistió en identificar las canciones sobre la base de datos *VoicesUy*, independientemente de quién sea el cantante. Este experimento permitió mostrar que la instrumentación es codificada en los coeficientes cepstrales y permite identificar las canciones aunque el nivel de la voz en la mezcla sea alto (ver tabla 5.2). El segundo experimento consistió en identificar las mismas voces pero sobre la base *AlbumsUy*, donde cada cantante está acompañado con su propia banda musical. En este experimento se pudo ver que la clasificación de voces alcanza valores superiores para cualquier intervalo de tiempo que se analice como se muestra en la tabla 5.6. Llamamos a este comportamiento "efecto banda".

Capítulo 6

Separación de voz cantada

6.1. Introducción

En este capítulo se aborda el problema de separar la voz de la mezcla de audio a los efectos de facilitar la identificación del cantante. Se parte del supuesto de que el acompañamiento puede resultar desfavorable para la identificación del cantante. Como se desprende de los experimentos del capítulo anterior, en muchos casos el acompañamiento dificulta el reconocimiento del cantante.

El uso de representaciones tiempo–frecuencia es de gran utilidad para el estudio de señales cuyo contenido espectral varía en el tiempo. En particular las señales provenientes de grabaciones musicales son claramente no estacionarias y el estudio de su contenido espectral en tiempos cortos es de gran importancia para el análisis y desarrollo de aplicaciones basadas en la extracción de información de dichas señales [75]. La representación más utilizada es el espectrograma que bajo el supuesto de que la señal en un tiempo corto (puede ser del orden de los 20 a 40 ms) no varía significativamente sus componentes espectrales, es una buena representación para una gran gama de aplicaciones. La definición del tamaño de la ventana determina la capacidad de ubicar un evento en el tiempo. Mientras que la resolución en frecuencia queda dada por la cantidad de muestras de señal a utilizar para calcular la transformada de Fourier y las propiedades de la ventana de análisis. En este punto se tiene un compromiso entre resolución temporal y frecuencial. En particular en música sería deseable contar con una mayor resolución frecuencial para bajas frecuencias, mientras que para altas frecuencias sería conveniente tener una mayor resolución temporal. Otra familia de representaciones fueron desarrolladas para tener en cuenta dicha variabilidad en la resolución y se denominan transformadas CQT (Constant Quality Transform) [76]. Otra particularidad que tienen las señales de audio provenientes de grabaciones musicales es que gran parte de las fuentes presentes contienen estructuras armónicas. Sonidos generados por instrumentos musicales tonales o la propia voz pueden ser bien aproximados, en intervalos de tiempo cortos (*frames*), como una serie de *chirps* lineales armónicamente relacionados. Visto en frecuencia, si la pendiente de la componente fundamental en t_0 es α_0, entonces la pendiente en el armónico n será $n\alpha_0$. Basado

en esta aproximación se puede utilizar la Fan Chirp Transform (FChT) como una buena representación del espectro de señales de audio [48].
En este capítulo se aborda el problema de separación de fuentes (Blind Source Separation (BSS)) en particular en la separación de la voz principal de grabaciones comerciales de música polifónica. En dicho contexto se evalúa la separación por enmascaramiento en tiempo-frecuencia utilizando las representaciones basadas en STFT y FChT. Se presentan también, experimentos en donde la separación de la voz se aplica al problema de identificación de cantante.

6.2. Representaciones tiempo-frecuencia

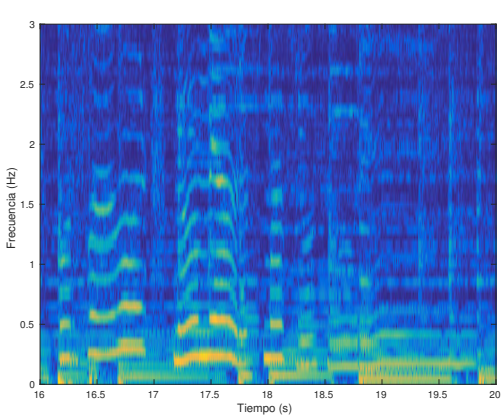
En este trabajo se utiliza dos tipos de representación: el espectrograma clásico, basado en la Transformada Discreta de Fourier (DFT por su sigle en inglés) y una representación alternativa basada en la transformada Fan Chirp.

6.2.1. Espectrograma

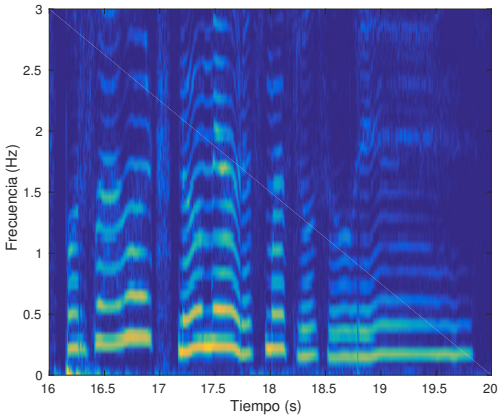
El espectrograma es una herramienta básica en el análisis espectral de señales de audio. Se puede definir como un gráfico de intensidad (tres dimensiones) del módulo de la transformada de Fourier de tiempo corto (STFT por sus siglas en inglés). Básicamente la STFT es la concatenación temporal de la transformada de Fourier de segmentos de datos enventanados. Las ventanas típicamente se solapan más del 50 % ya que esto tiene varias ventajas para trabajar con las señales de audio en el dominio de las frecuencias y luego volver al dominio del tiempo además de aumentar la resolución temporal. Los principales parámetros son entonces: el largo de la ventana a utilizar sobre la señal en el tiempo, el tipo de ventana, el paso (define el solapamiento) y la cantidad de puntos de la FFT (lo que determina el relleno de ceros).
En la figura 6.1a se puede ver el espectrograma de un fragmento de música con voz cantada utilizando una ventana de Hann de 1024 muestras con un paso de 256 muestras. Se puede observar el espectrograma del audio de la voz sola, del mismo fragmento en la figura 6.1b.

6.2.2. Fan Chirp Transform (FChT)

Una representación clásica como el espectrograma se basa en la suposición de que una señal no varía su contenido espectral en un lapso de tiempo corto. Una mejor representación de tiempo corto para una fuente que genera señales cuya frecuencia fundamental varía en el tiempo, es una serie de *chirps* lineales armónicamente relacionados, donde el espectrograma podría verse como un caso particular (cuando la pendiente del *chirp* es nula). Basado en esta idea es que Cancela et al [48] proponen un método para definir *frame* a *frame* el mejor conjunto de *chirps* que representan la señal de interés, logrando una representación en tiempo

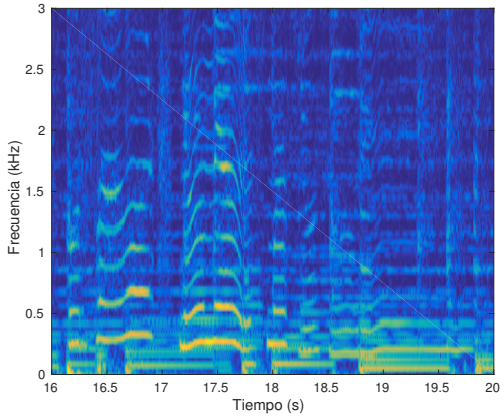


(a) Espectrograma de voz con acompañamiento musical.

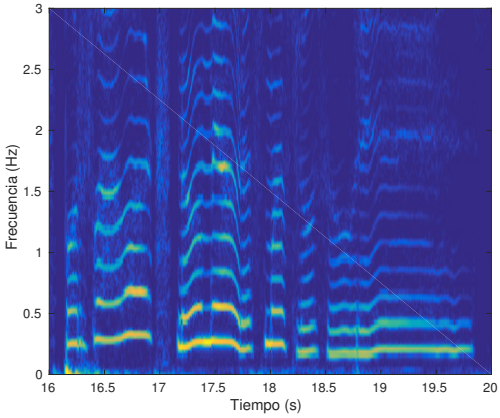


(b) Espectrograma solo la voz.

Figura 6.1: Espectrogramas de un fragmento de audio con ventanas de 1024 muestras y zero–padding a 2048.



(a) Transformada Fan Chirp de voz con acompañamiento musical.



(b) Transformada Fan Chirp solo la voz.

Figura 6.2: Transformada Fan Chirp sobre mismo fragmento de audio.

y frecuencia más definida. En la figura 6.2a se puede ver un espectrograma construido a partir de la FChT para el mismo fragmento de audio en el que se puede ver que los componentes armónico tienen mejor definición, particularmente en los *glissandos* de la voz y en la figura 6.2b solo la voz para igual parametrización.

Deformación temporal (time warping)

Uno de los puntos clave de la transformada está en realizar una deformación temporal de la señal de forma de utilizar las ventajas computacionales del cálculo de la transformada rápida de Fourier (FFT). Lo que se busca es realizar una deformación temporal de la señal de forma que un chirp lineal se convierta en una

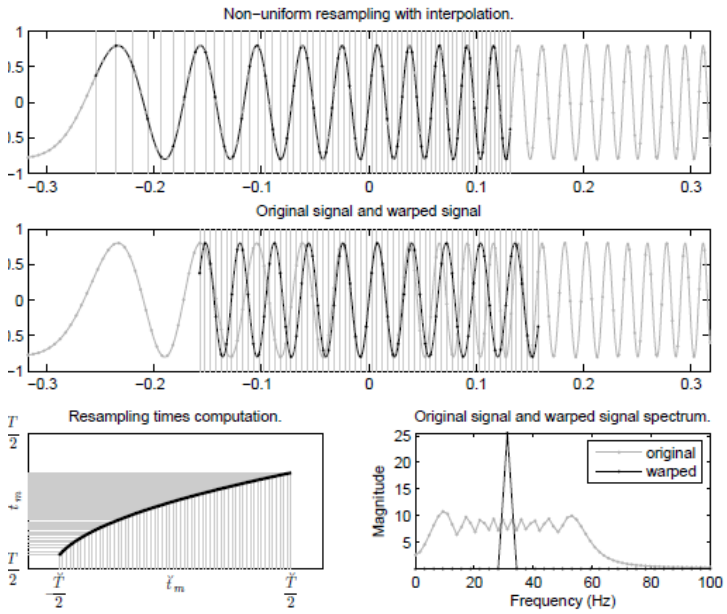


Figura 6.3: Proceso de Warping para un chirp lineal. Imagen tomada del artículo de Cancela et.al. [48].

sinusoidal estacionaria, para luego realizar la DFT.

Para el cálculo de la transformada se utiliza la siguiente definición:

$$X(f,\alpha) = \int_{-\infty}^{+\infty} x(t)\phi'_\alpha(t)e^{-j2\pi f\phi_\alpha(t)}dt, \tag{6.1}$$

donde $\phi_\alpha(t) = (1+\frac{1}{2}\alpha\, t)\, t$ es una función de deformación temporal (*warping*). Realizando un cambio de variable $\tau = \phi_\alpha(t)$ se puede observar que la transformación es equivalente a realizar la transformada de Fourier de la señal deformada en el tiempo.

Alcanza entonces con definir el α que mejor represente cada *frame* y realizar la transformada sobre muestras de la señal sobre el nuevo espacio temporal[1]. En la figura 6.3 se puede ver el proceso de *warping* para un *chirp* lineal.

Ajuste de *chirp* lineal con Gathered Log Spectrum (GlogS)

Una forma de estimar el mejor α para cada *frame* es realizar el cálculo para un conjunto finito de posibles α y maximizar una función de predominancia de tono. La función de predominancia de tono (pitch salience) debe estimarse teniendo en consideración la energía presente en cada tono y todos sus armónicos. Dada la estimación de la potencia espectral de un *frame* de la señal de análisis $|S_x(f)|$, la

[1]Por más detalles de la implementación ver [48].

predominancia de cada tono se calcula como la sumatoria de la potencia en todos los armónicos menores a la frecuencia máxima de interés.

$$\rho_0(f_0) = \frac{1}{n_H} \sum_{i=1}^{n_H} log|S(if_0)| \tag{6.2}$$

La inclusión de la función logaritmo en el cálculo genera resultados más robustos ya que disminuye el efecto de ruido y da relevancia a armónicos superiores de la serie armónica.

El máximo de la función 6.2 (variando F0 sobre el rango de frecuencias de interés) debería coincidir con la frecuencia fundamental buscada. En este punto hay que tener en cuenta que múltiplos o submúltiplos pueden tomar valores superiores por lo que una lógica de pos–procesamiento es necesaria.
En particular, la implementación presentada en [48] tiene la flexibilidad de poder modificar el cálculo de la norma de la estimación de la potencia espectral incluyendo un factor γ como se muestra en la ecuación 6.3.

$$\rho_0 = \frac{1}{n_H} \sum_{i=1}^{n_H} log(\gamma|S(if_0)| + 1) \tag{6.3}$$

Una vez hallado cada valor de α, la representación en tiempo y frecuencia es la concatenación de la FFT para la señal deformada en el tiempo para cada *frame*.

6.3. Separación de voz

En la figura 6.1b se puede observar que la mayor parte de la energía de la voz se distribuye en el plano TF en la frecuencia fundamental y sus armónicos. Si bien en el ejemplo presentado es claro, esto depende en gran medida de la forma en que maneja el aire el intérprete y fundamentalmente del tipo de sonidos emitidos. Como una aproximación a la separación de la voz cantada en este trabajo se busca extraer la parte tonal de la voz, desestimando los sonidos sordos y suponiendo que el espectro de la voz es armónico sobre la frecuencia fundamental de la melodía.

6.3.1. Separación por enmascaramiento

La metodología utilizada consiste en enmascarar la mezcla en su representación tiempo-frecuencia multiplicando punto a punto por una matriz de igual dimensiones que seleccione los *bins* de cada *frame* en el entorno de los múltiplos de la frecuencia fundamental.
Dicha máscara puede seleccionar de forma binaria o puede tener sus valores ponderados por algún criterio que utilice más información (no solamente la frecuencia fundamental). Como prueba de concepto, en este capítulo se utiliza la F0 de las anotaciones de las bases de datos.

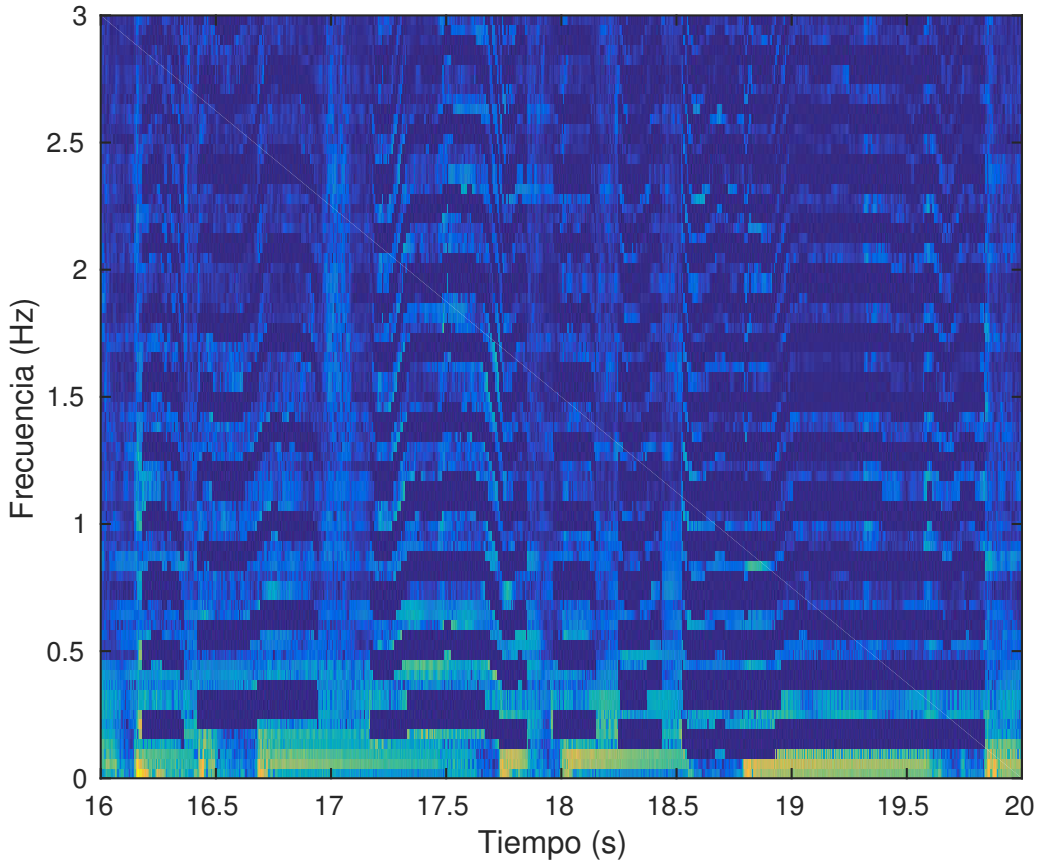


Figura 6.4: Residuo de enmascaramiento binario con máscara de tres *bins* por armónico.

Luego de realizar el enmascaramiento se pueden obtener dos representaciones, una para la señal de interés (en este caso la voz) y otra para el residuo (en este caso los remanentes no tonales de la voz y el resto del acompañamiento musical).

Hay que tener en cuenta que en el caso de utilizar una máscara binaria, toda la energía de acompañamiento musical presente en las frecuencias de interés se introduce como interferencia con la fuente que se intenta separar (ver espectrograma del residuo en figura 6.4). En las siguientes secciones se presentan dos métodos para reducir este efecto.

Síntesis

STFT: Para el caso de la STFT se pueden obtener las señales de audio sobre la representación obtenida en tiempo–frecuencia al calcular la transformada inversa de cada *frame* y aplicar *overlap–add* con el mismo paso con que se realiza el análisis. Dado que el espectrograma fue modificado pueden generarse artefactos al realizar la superposición de intervalos de señal (*overlap–add*). Para reducir dichos efectos se realiza *zero–padding* de la señal en la etapa de análisis de forma que la transformada inversa de cada *frame* tenga una envolvente temporal más parecida a la ventana utilizada al construir el espectrograma. En este punto puede

interpretarse el enmascaramiento como el producto por un tren de pulsos en frecuencia, lo que en el tiempo sería equivalente a la convolución con un *sinc* discreto.

FChT: La síntesis en el caso de la FChT es algo más compleja ya que la transformada inversa de un *frame* lo que devuelve son muestras de la señal deformada en el tiempo y por ende su envolvente (la ventana de análisis) también deformada. Se puede entonces resumir la síntesis en los siguientes siete pasos:

1. Realizar la transformada inversa de Fourier (IFFT) para cada *frame*.
2. Definir un vector de tiempo para recuperar la señal con la cantidad de muestras de la señal original (la cantidad de muestras no tiene necesariamente que coincidir con la IFFT).
3. Deformar dicho vector de tiempos con la ecuación $\phi_\alpha(t) = (1 + \frac{1}{2}\alpha t)t$, utilizando el α correspondiente a cada *frame*.
4. Interpolar el valor de la señal utilizando los valores de la señal con el vector de tiempo equiespaciado (en el espacio deformado) de la señal obtenida de la IFFT.
5. Interpolar también en dicho vector de tiempo los valores de la función de enventanado. Realizar *overlap–add* de la ventana en paralelo con la señal.
6. Realizar *overlap–add* de la señal.
7. Corregir el efecto de enventanado escalando la señal con la señal de enventanado construida.

6.3.2. Ponderación de máscara por Wiener

Como se vio en la sección anterior, cada *bin* de frecuencia filtrado es utilizado con su energía original para recuperar la señal de la voz sin tomar en cuenta el nivel de interferencia que el resto de las fuentes generan. En esta sección se busca mejorar el resultado de la separación, ponderando los coeficientes de la máscara mediante una estimación de la energía de las fuentes del residuo. Para tal fin se utiliza el método denominado separación por filtro de Wiener [77].

El filtro de Wiener es utilizado para estimar estadísticamente una fuente desconocida en presencia de ruido, bajo las hipótesis de proceso estacionario en la señal y en el ruido, siendo óptimo en estos casos para la minimización del error cuadrático medio. Conociendo la autocorrelación de la señal se calculan los coeficientes del filtro que minimizan el error con la señal buscada. El método que se utiliza a continuación no es estrictamente un filtro de Wiener, pero teniendo una buena estimación de la energía del ruido en cada punto del plano tiempo–frecuencia permite disminuir la interferencia.

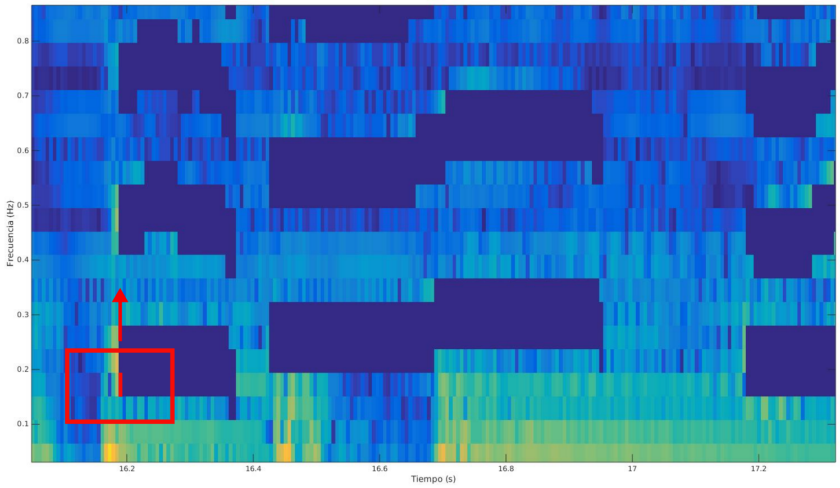


Figura 6.5: Filtro de media móvil.

Para ponderar cada coeficiente $m_{i,j}$ de la máscara binaria definida entorno a las frecuencias de interés se aplica la ecuación:

$$m_{i,j} = \frac{||S_{i,j}||^2}{||S_{i,j}||^2 + ||N_{i,j}||^2}, \tag{6.4}$$

donde $S_{i,j}$ es el valor en el *frame* i y en el *bin* de frecuencia j del espectro de la señal de voz estimada y $N_{i,j}$ el espectro del ruido (interferencia del acompañamiento) en el mismo punto.

El principal problema del método es conocer la energía de ruido (acompañamiento musical) en cada punto del plano tiempo–frecuencia. Se propone para este trabajo, utilizar el residuo del filtro con la máscara binaria como estimación del acompañamiento musical. Dado que la ponderación es punto a punto en el plano TF, es necesario estimar la energía del acompañamiento musical en las regiones de las frecuencias de interés (fundamental y armónicos de la voz). Se propone estimar la energía en las regiones de interés (hasta este punto en cero ver figura 6.4) aplicando un filtro de media móvil en la representación tiempo–frecuencia. Se tomará como ventana para el filtro de promedio los 10 *frames* anteriores y los 10 posteriores ($\pm 230\,ms$) sobre el *bin* de frecuencia de interés, y los *bins* superior e inferior. O sea, cada *bin* vacío tomará el valor promedio de un entorno de 20 *frames* y 3 *bin* de frecuencias (ver figura 6.5),los valores del entorno con energía cero no se utilizan para el cálculo del promedio. El resultado de la convolución con el núcleo, ya descrito, solo se utiliza para los puntos donde la máscara de filtrado genera ceros. La estimación realizada de la potencia de ruido permite ponderar el peso de cada coeficiente de la máscara utilizando la ecuación 6.4. El resultado para el mismo fragmento de audio se puede ver en la figura 6.6.

Una vez ponderados los coeficientes de la máscara el proceso de filtrado y la síntesis coinciden con lo ya explicado previamente, tanto para la STFT como para

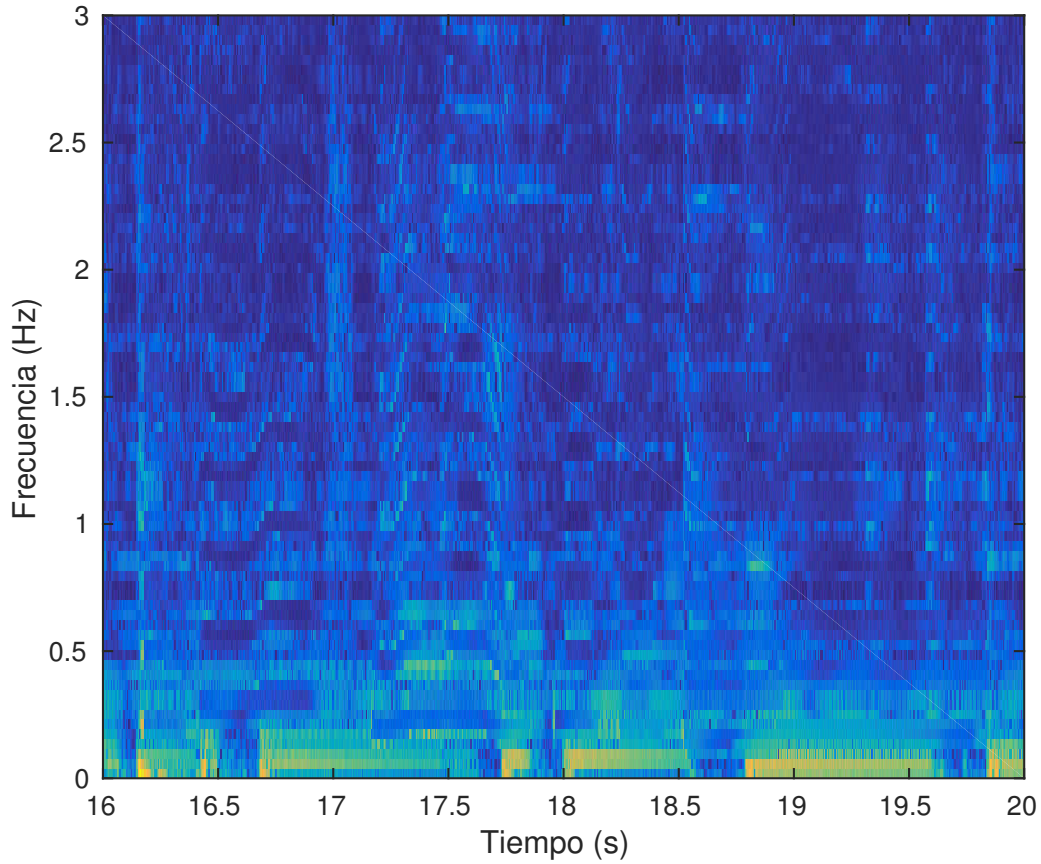


Figura 6.6: Resultado de aplicar media móvil sobre puntos faltantes en residuo de enmascaramiento.

la FChT.

6.3.3. Ponderación de máscara por ajuste de envolventes espectrales

Las interferencias de instrumentos musicales que coinciden en frecuencia con componentes tonales de la voz generan un nuevo sonido que, en la mayoría de los casos no podría haber sido generado por el cantante. Basado en que las componentes espectrales de la voz siguen algunos patrones que otros instrumentos no, se propone generar un diccionario de envolventes espectrales que sean representativas de la voz que se quiere separar para ajustar desvíos generados de posibles interferencias. El procedimiento utilizado en este trabajo es el que se resume a continuación:

1. Realizar la FChT sobre la pista de voz original.
2. Calcular la envolvente espectral para cada *frame* teniendo como parámetro la *Quefrency* de corte n_c en el filtrado del espectro *liftrado*.

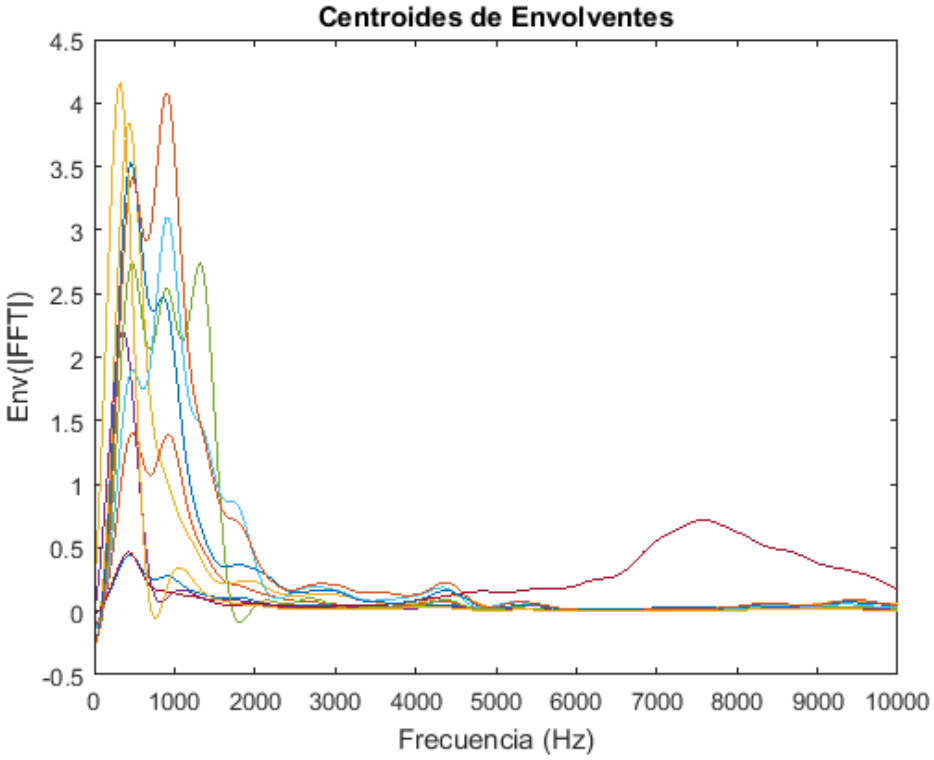


Figura 6.7: Centroides de envolventes espectrales sobre audio original de la voz

3. Realizar *clustering* de envolventes con K-means. Teniendo como parámetros la cantidad de centroides.

4. Enmascarar la FChT de la señal mezcla. Sobre el resultado calcular la envolvente espectral para cada *frame*.

5. Encontrar el centroide más próximo a la envolvente espectral de cada *frame* (distancia euclidiana).

6. Ajustar en valor absoluto de cada *bin* de frecuencia proporcionalmente a la diferencia con el centroide más próximo.

A modo de ejemplo, en la figura 6.7 se puede ver el conjunto de centroides obtenido con K-means para k=10 aplicado al audio de la voz sola. Se pueden identificar varias curvas con formantes similares a lo que se podría esperar de una vocal y una de estas con componentes por encima de los 5kHz vinculada a consonantes como la "s". En la figura 6.8 se puede ver el proceso de ajuste para un *frame*. Este proceso se puede repetir alterando los parámetros de ajuste de forma de encontrar un conjunto de parámetros que mejore el desempeño. Este punto se explora en 6.4.2. A continuación se presentan algunas métricas de desempeño.

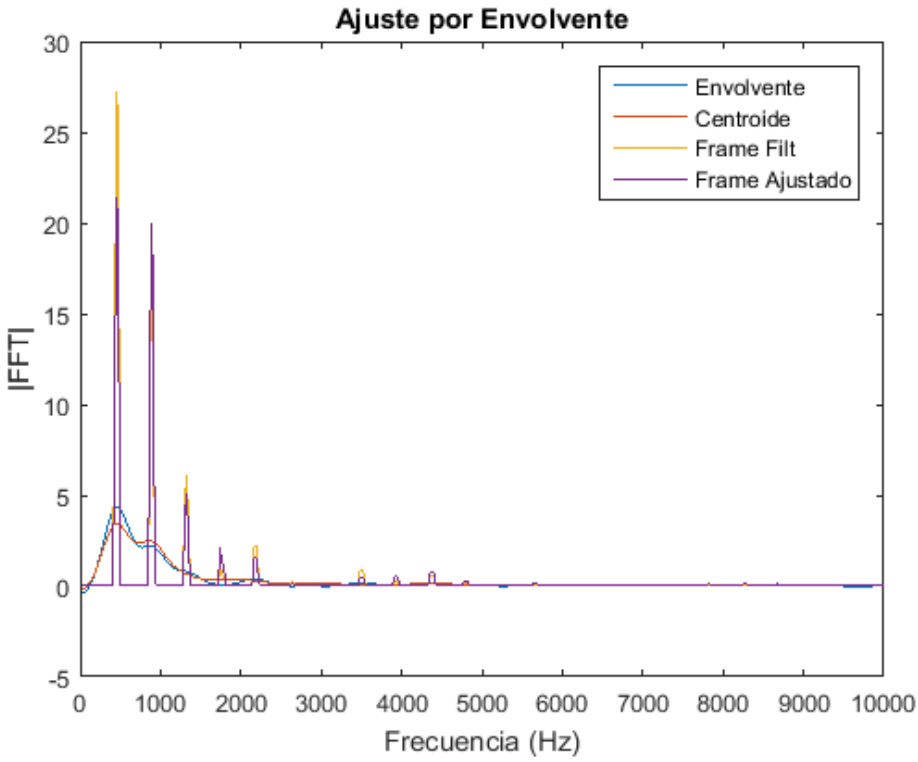


Figura 6.8: Proceso de ajuste de coeficientes de un *frame* por envolvente.

6.3.4. Medidas de calidad de separación de fuentes

Varios trabajos se han realizado para evaluar la calidad percibida de la separación de fuentes. Algunos implican el estudio de las respuestas de un conjunto de oyentes para obtener una correlación con coeficientes de desempeño [78]. Dentro de las medidas objetivas de desempeño, una de las más utilizadas es la Signal to Distortion Ratio (SDR) [79]. Emmanuel Vincente et al. (2006) dividen los efectos distorsivos productos de la separación en tres componentes, error de interferencia, error de artefactos y ruido. Por lo que la relación entre la fuente objetivo y la fuente obtenida en la separación se puede expresar como en la ecuación 6.5.

$$\hat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{6.5}$$

Dada esta separación en términos, se definen cuatro indicadores de relaciones de energía en decibelios: relación fuente a distorsión (SDR ec. 6.6), relación fuente a interferencia (SIR ec. 6.7), relación fuentes a ruido (SNR ec. 6.8) y relación fuentes a artefactos (SAR ec. 6.9). Es importante destacar que para calcular etos indicadores es necesario tener las pistas de audio originales (sin mezclar). El cálculo de cada uno de los indicadores con detalles sobre cómo se realiza la estimación del aporte de cada componente de ruido está explicado en un trabajo posterior de Emmanuel Vincent (2011) [80]. En este trabajo se utiliza la biblioteca de funciones BSS_EVAL [79] para calcular los indicadores mencionados anteriormente.

$$SDR = 10log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{6.6}$$

$$SIR = 10log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{6.7}$$

$$SNR = 10log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2} \tag{6.8}$$

$$SAR = 10log_{10}\frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2}. \tag{6.9}$$

6.4. Experimentos

El objetivo de los experimentos que aquí se presentan es comparar el desempeño de cinco métodos de separación de la voz cantada por enmascaramiento sobre dos bases de datos, utilizando las medidas SDR, SIR y SAR.
Los métodos utilizados son:

1. Máscara binaria sobre representación STFT.
2. Máscara ponderada por Wiener sobre representación STFT.
3. Máscara binaria sobre representación FChT.
4. Máscara ponderada por Wiener sobre representación FChT.
5. Máscara ponderada por envolventes espectrales sobre representación FChT.

6.4.1. Bases de datos

Para los experimentos se utilizan las bases de datos *VoicesUy* creada específicamente para esta tesis (ver capítulo 3) y la base de datos MedleyDB [58].

MedleyDB es una base de datos de audios multipista con anotaciones, desarrollada en 2014 por el laboratorio de investigación de audio de la Universidad de Nueva York (Music and Audio Research Lab, New York University). Contiene 122 canciones de las cuales 108 tienen anotaciones de melodías (60 vocales y 48 instrumentales). Las canciones son de géneros variados como rock, pop, jazz, cantautor.

Cada canción contiene los audios separados de cada pista, el audio de la mezcla y un archivo yaml (MetaData) conteniendo toda la información relevante (ej. anotaciones de *pitch*, anotaciones de melodía, etiquetas de comienzo por pista, nombre del instrumento de cada pista, genero, autor, etc.). Todos los audios están en formato wav con una frecuencia de muestreo de 44 100 Hz y 16 bits de cuantización.

En el presente trabajo se utilizan los archivos de audio de mezcla de 23 canciones de la base MedleyDB de diferentes géneros, los audios de la melodía sola y las

anotaciones de altura de la melodía. También se realizan los experimentos sobre las 40 canciones de la base *VoicesUy*.

6.4.2. Metodología

Para cada canción de la base MedleyDB se genera un archivo de audio de acompañamiento musical (sumando las pistas de audio de todos los instrumentos) y se mezcla con el archivo de audio de la voz (los niveles de mezcla son los originales de la canción). Para realizar los experimentos de separación sobre esta base se toman los primeros 10 s desde que la voz principal comienza a cantar.

Para la base *VoicesUy* se utilizan los audios completos de mezcla a 0dB de la carpeta `/Mixes`.

Experimento 1

Se realiza el análisis del audio de mezcla con ventanas de Hann de 1024 muestras (23.2 ms) con un paso de 256 muestras determinando un vector de tiempos asociado a cada *frame*. Sobre dicho vector de tiempos se estima el valor de F0 con una interpolación lineal sobre el *ground truth*. Se construye el espectrograma realizando la transformada rápida de Fourier (FFT) con 2048 puntos sobre cada *frame* de señal.

Dada la resolución discreta en frecuencias, $22.050/1024 = 21.5\,\text{Hz}$, se busca el *bin* más cercano a la F0 y a cada uno de sus armónicos para todo tiempo. Se construye la máscara binaria con los *bins* seleccionados y un *bin* superior e inferior para cada *frame*.

En la etapa de síntesis, la ventana utilizada y el paso aseguran que al realizar la reconstrucción de la señal con *overlap–add* no se tenga un efecto de la ventana sobre la envolvente de amplitud. Se reconstruye la señal filtrada desde su espectrograma, $S_{filt}(i,j) = S(i,j).Mask(i,j)$ y la señal de residuo desde su representación T–F, $S_{residuo} = S - S_{filt}$.
Tomando las dos señales reconstruidas y los audios originales de voz y mezcla instrumental, se calculan las medidas de desempeño SDR (distorsión), SIR (interferencia) y SAR (artefactos).

Experimento 2

El experimento 2 utiliza el mismo procedimiento que el experimento 1 agregando, previo a la síntesis, una ponderación de la máscara por Wiener (visto en 6.3.2). El cálculo de las medidas de desempeño se realiza en relación a los audios del *ground truth*.

Experimento 3

En el contexto del grupo de Procesamiento de Audio del Instituto de Ingeniería Eléctrica[2] es realizado en 2010 un trabajo que incluye el desarrollo de un programa para calcular la transformada FChT en *Matlab* y optimizado con funciones en lenguaje C [48]. Dicho código es utilizado en este trabajo para computar la FChT. Los parámetros más relevantes utilizados para configurar la transformada son los siguientes:

- Cantidad de puntos para computar la FFT sobre la señal en el tiempo deformado = 1024
- Cantidad de α utilizados para seleccionar aquel que maximiza el GlogS = 21
- α máximo = 6
- Tipo de Warping = lineal
- Factor de sobremuestreo en el tiempo deformado = 2
- Paso (hop) = 512

Dado que α es la tangente del ángulo formado por la curva de F0 y los valores de α se eligen de manera lineal en todo el rango de valores considerado, se obtiene más resolución para F0 con pendientes altas. Se propone en este trabajo una distribución sobre la curva que sigue una función tangente en lugar de lineal, logrando una distribución de α que permite evaluar pendientes de F0 equiespaciadas entre $[-\pi/2; \pi/2]$. Se muestra con "x" en el gráfico en la figura 6.9. Con el mismo costo computacional se computan 15 puntos entre pendientes de F0 de -1 a 1 radianes (-57º a 57º) en lugar de 5 puntos.

El procedimiento de enmascaramiento es idéntico al utilizado en los experimentos anteriores y la síntesis es programada respetando el método presentado en la sección 6.3.3. Luego de la síntesis se computan SDR, SIR y SAR.

Experimento 4

Análogamente al experimento 2, aquí la única modificación respecto al experimento 3 es la ponderación de la máscara utilizando el método de Wiener. Luego de la síntesis se computan SDR, SIR y SAR.

Experimento 5

Este experimento se realiza exclusivamente sobre la base MedleyDB. Para este experimento se aplica el ajuste de la máscara por envolvente presentado en 6.3.3. Para definir los parámetros del ajuste: frecuencia de corte en el cálculo de la envolvente espectral (Q_c, *Quefrency de liftrado*), cantidad de centroides (k) y factor

[2]Facultad de Ingeniería, UdelaR.

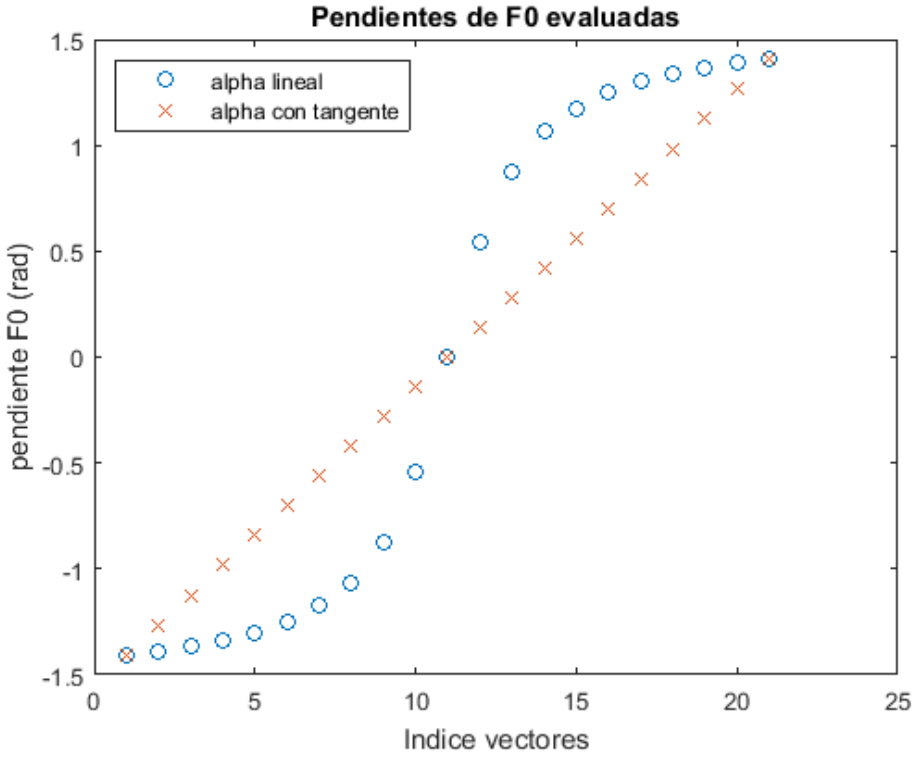


Figura 6.9: Distribución de pendientes de F0 evaluadas al computar FChT

máximo de ajuste de envolvente. Se propone un método simple para realizar una búsqueda de algún máximo local de la SDR. El método es computar SDR sobre una cantidad finita y aleatoria de puntos en dicho espacio de tres dimensiones. En cada iteración, se asigna aleatoriamente Q_c entre 0.01 y 0.1, la cantidad de centroides k entre 10 y 100, y el factor máximo de ajuste entre 1.0 y 2.0.

En al figura 6.10 se muestra el mejor conjunto de parámetros para cada una de las 23 canciones analizadas. Se presenta la *quefrency* de corte del *liftrado* de forma normalizada. Se puede ver que hay una tendencia a seleccionar valores de corte más altos y una mayor cantidad de centroides.

6.4.3. Resultados

Resultados sobre MedleyDB

Para cada uno de los experimentos fueron calculadas las medidas de desempeño de la separación, tanto para la voz como para el acompañamiento (tomando la mezcla sin la voz como una única señal). En la figura 6.11 se presenta el resultado en dB para SDR, SIR y SAR del promedio de las 23 canciones. Se puede ver cómo la SDR aumenta con el avance de los cinco experimentos (STFT, STFT+Wiener, FChT, FChT+Wiener, FChT+Env).

En la tabla 6.1 se presentan los resultados del ratio SDR en dB para las 23

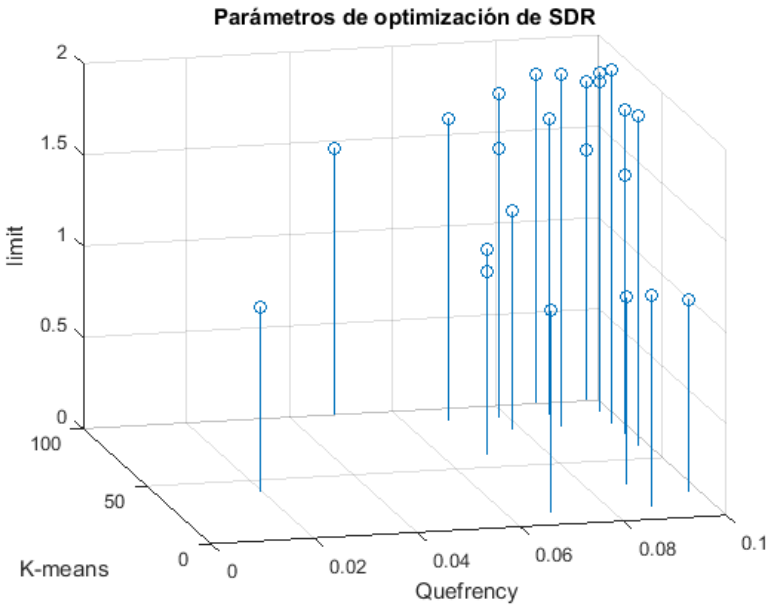


Figura 6.10: Búsqueda de parámetros óptimos para ponderación de máscara por ajuste de envolventes espectrales. Datos luego de 20 iteraciones sobre la base MedleyDB.

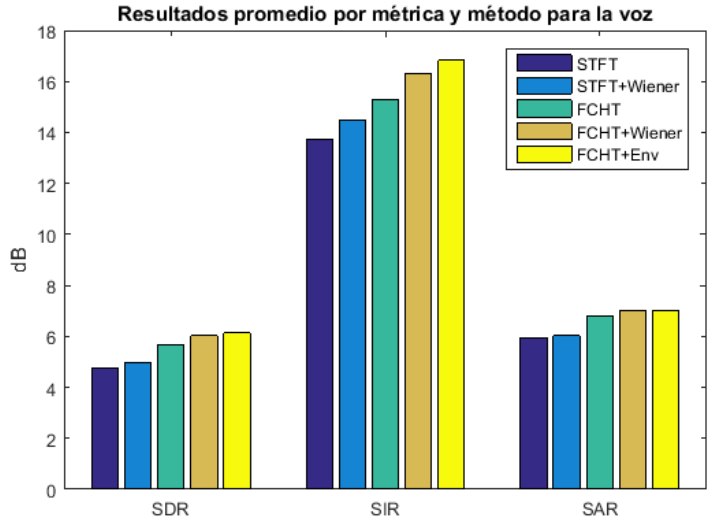


Figura 6.11: Resultados del desempeño de la separación de la voz, en promedio, para las canciones de la base MedleyDB

canciones analizadas en los cinco experimentos. Por completitud, el resto de los resultados quedan en anexos.

Resultados sobre *VoicesUy*

La separación de fuentes también es aplicada sobre las mezclas de la base de datos *VoicesUy*. Ocho cantantes diferentes cantando cinco canciones conforman dicha base (ver capítulo 3). Para la separación se utiliza la información de frecuencia fundamental de cada voz. Se realizan los experimentos de uno al cuatro presentados en 6.4, que básicamente son la separación por enmascaramiento utilizando las representaciones STFT y FChT con y sin ponderación por Wiener.

Los resultados de calidad de separación representados por los ratios SDR, SIR y SAR como un promedio de las 40 canciones de la base se presentan en la figura 6.12. Al igual que sobre la base MedleyDB, se ve una mejora clara en los tres indicadores de calidad al utilizar la representación FChT con ponderación por Wiener. El SDR promedio es de 9.4 dB alcanzando valores de hasta 13.2 dB, mientras que con las interferencias (SIR) se logra un promedio sobre la base de 18.7 dB con un máximo de 23.9 dB. Los resultados de SDR para 40 canciones de la base se presentan en la tabla 6.2, todos los resultados de separación sobre la base $VoicesUy$ incluyendo la calidad de la separación del acompañamiento musical se presentan en el apéndice B.

En la figura 6.13 se presentan los indicadores promedio para cada voz, donde se pueden ver mejores resultados en las voces femeninas (4 y 8) mientras que entre las voces masculinas el resultado es muy similar. Los resultados presentados no solo dependen de las características de cada voz sino también de la canción. En la figura 6.14 se puede ver la variación de la calidad de la separación sobre los 40 archivos de audio analizados, para cada uno de los cuatro métodos utilizados en el experimento.

6.5. Clasificación de voces con separación de fuentes

6.5.1. Introducción

Como ya quedó demostrado en este trabajo (ver capítulo 5) el acompañamiento musical disminuye el desempeño del algoritmo de clasificación de voces. La pregunta a responder es si las técnicas de separación de voces por enmascaramiento, que generan una disminución del acompañamiento musical, mejoran el resultado de la identificación de voces cantadas.

Varios trabajos han mostrado de forma experimental que aplicar técnicas de separación de fuentes mejora los resultados de la clasificación [14, 19, 25, 28]. Fujihara et al. [19] muestran que al utilizar MFCCs como características y aplicar separación por enmascaramiento de las componentes tonales de la voz (sobre el

Tabla 6.1: SDR (dB) de la señal de voz separada para todas las canciones analizadas

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	4,09	4,28	4,98	5,42	5,48
'AlexanderRoss_GoodbyeBolero'	1,90	2,40	2,45	3,00	4,51
'Auctioneer_OurFutureFaces'	3,14	3,59	3,82	4,38	3,79
'AvaLuna_Waterduct'	5,35	5,18	5,28	5,42	5,51
'ClaraBerryAndWooldog_Stella'	9,08	8,89	9,64	9,66	9,75
'ClaraBerryAndWooldog_WaltzForMyVictims'	7,92	7,95	8,30	8,48	8,30
'Creepoid_OldTree'	5,51	5,66	6,22	6,54	6,81
'DreamersOfTheGhetto_HeavyLove'	7,26	7,47	8,22	8,70	8,14
'HezekiahJones_BorrowedHeart'	1,91	2,18	2,59	2,95	2,60
'HopAlong_SisterCities'	4,24	4,20	4,88	5,14	5,31
'LizNelson_Coldwar'	1,98	2,32	2,34	2,62	4,48
'LizNelson_Rainfall'	5,99	6,35	6,32	6,45	8,15
'Meaxic_TakeAStep'	2,78	3,16	4,42	5,02	4,53
'MusicDelta_Country1'	5,03	5,11	5,62	5,96	6,20
'MusicDelta_Disco'	6,80	6,68	9,41	9,38	9,05
'MusicDelta_Grunge'	2,23	2,74	2,86	3,46	3,36
'MusicDelta_Punk'	4,60	4,68	6,22	6,32	5,91
'MusicDelta_Reggae'	5,85	6,09	7,50	7,79	7,69
'MusicDelta_Rockabilly'	4,54	5,04	5,50	6,36	6,49
'MusicDelta_Rock'	4,68	5,33	5,81	6,61	6,76
'SecretMountains_HighHorse'	4,06	3,88	3,91	3,95	3,84
'TheSoSoGlos_Emergency'	3,27	3,61	4,10	4,94	4,27
'Wolf_DieBekherte'	7,38	7,40	10,13	10,07	10,13
Media	4,77	4,97	5,67	6,03	6,13

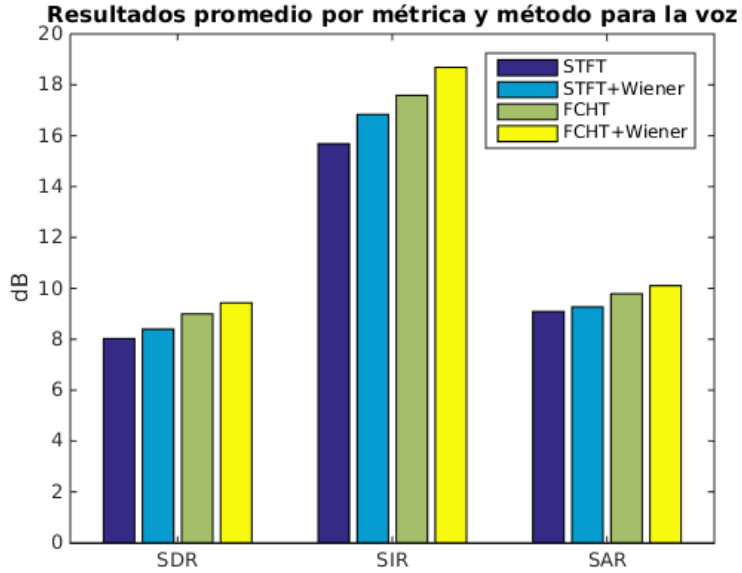


Figura 6.12: Resultados del desempeño de la separación de la voz, en promedio, para las canciones de la base *VoicesUy*.

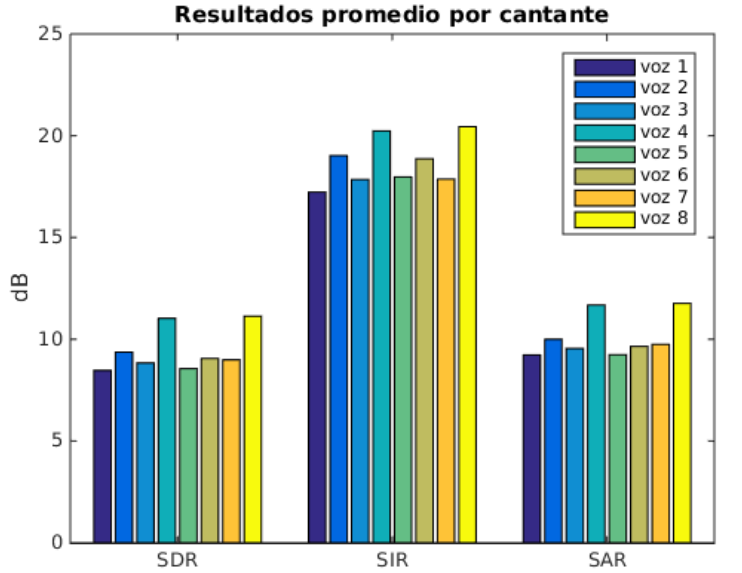
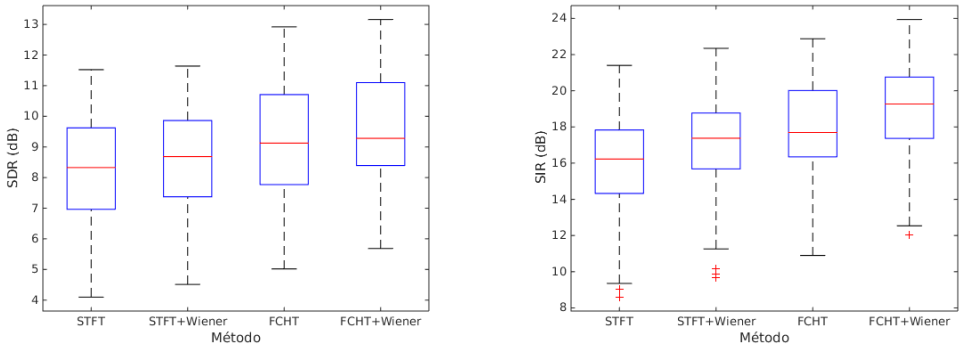


Figura 6.13: Resultados del desempeño de la separación de la voz con el método FChT+Wiener, en promedio, para cada uno de los ocho cantantes de *VoicesUy*.

Tabla 6.2: Resultados de separación SDR en dB sobre la base *VoicesUy*

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	8,9	9,2	9,6	9,6
	Graña	8,7	8,8	9,3	9,1
	Maturro	8,4	8,7	8,8	8,7
	Ferreira	10,8	10,9	12,6	12,8
	Gavilán	8,8	8,9	9,3	9,2
	Zubillaga	8,2	8,6	8,8	9,0
	Rosberg	8,3	8,7	8,7	8,8
	Núñez	11,5	11,6	12,4	12,5
La edad del cielo	Román	9,1	9,3	10,3	10,5
	Graña	9,9	10,1	10,7	11,0
	Maturro	9,8	9,9	10,9	11,0
	Ferreira	11,4	11,6	12,9	13,2
	Gavilán	9,8	9,9	11,3	11,5
	Zubillaga	8,8	8,8	9,7	9,8
	Rosberg	10,3	10,6	11,7	12,0
	Núñez	11,3	11,4	12,6	12,8
Pa' los músicos	Román	7,1	7,4	7,8	8,3
	Graña	9,7	9,9	10,7	11,3
	Maturro	8,2	8,6	9,2	9,8
	Ferreira	9,4	9,8	10,8	11,5
	Gavilán	8,7	9,1	9,5	10,4
	Zubillaga	9,6	9,8	10,6	11,2
	Rosberg	7,2	7,6	8,1	8,7
	Núñez	10,5	10,7	11,6	12,0
Príncipe azul	Román	5,7	6,3	6,8	7,5
	Graña	5,8	6,3	6,7	7,1
	Maturro	4,3	4,7	5,3	5,7
	Ferreira	7,6	7,9	9,3	9,6
	Gavilán	4,1	4,5	5,4	5,8
	Zubillaga	5,2	5,7	6,2	6,6
	Rosberg	4,4	5,1	5,4	6,1
	Núñez	6,9	7,3	8,3	8,7
Promesas sobre el bidet	Román	4,6	5,5	5,3	6,4
	Graña	7,1	7,7	7,8	8,5
	Maturro	7,0	8,0	7,8	8,9
	Ferreira	6,6	7,1	7,5	8,1
	Gavilán	4,4	5,2	5,0	5,9
	Zubillaga	7,2	7,7	8,0	8,7
	Rosberg	7,3	8,3	8,0	9,3
	Núñez	8,4	8,7	9,1	9,5



(a) Nivel de distorsión (SDR) por método sobre *VoicesUy*

(b) Nivel de interferencia (SIR) por método sobre *VoicesUy*

Figura 6.14: Estadística de resultados de separación de las 40 canciones de *VoicesUy* según método utilizado.

espectrograma), se pasa de 58 % a 65 % de acierto en la identificación de voces cantadas.

6.5.2. Experimentos de clasificación de voz con separación de fuentes.

Se presentan los resultados de clasificar las voces de la base *VoicesUy* sobre los audios producto de la separación. El método seleccionado es el que obtuvo los mejores resultados sobre las bases en 6.4.3, separación sobre la representación basada en FChT con ponderación de máscara por Wiener (experimento 4 descrito en 6.4.2).

Se compara el resultado obtenido con el mejor caso, que es clasificar los archivos de audios de las voces solas. Lo que se espera es que el resultado supere la clasificación de las voces en el contexto de música polifónica (mezcla a 0 dB), esto no necesariamente se tiene que cumplir ya que el proceso de separación podría generar distorsión y artefactos al disminuir la interferencia con el acompañamiento de forma que afecte el contenido espectral.

Para mantener coherencia con los experimentos planteados en el capítulo 5, se clasifica las voces utilizando los mismos siete intervalos de tiempo y los dos métodos ya analizados (máximo de la **media** de la log–verosimilitud y **moda** del máximo de la log–verosimilitud).

Los resultados calculados como el promedio de validación cruzada de cinco particiones se presentan en las tablas 6.3 y 6.4. Para ambos métodos se puede ver que la clasificación de las voces, luego de aplicar el proceso de separación, se encuentra entre los resultados obtenidos para la mezcla y los obtenidos para las voces solas.

Tabla 6.3: Comparación de resultados de clasificación de voces de la base *VoicesUy*. Método de clasificación de intervalos:"**media**".

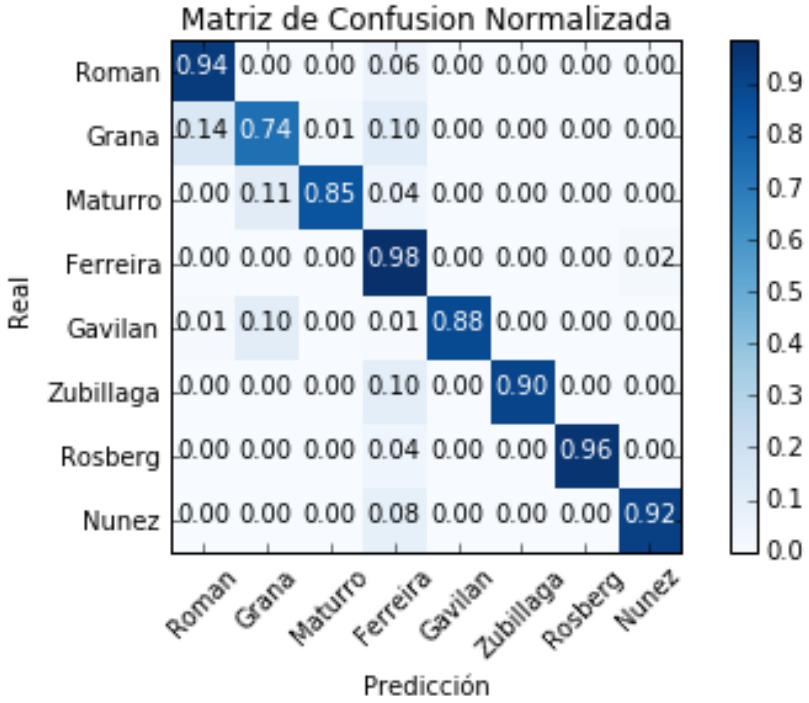
Audio	Duración intervalos (s)						
	0,1	0,5	1	2	5	10	40
Mezcla 0dB	34,2 %	46,0 %	54,2 %	62,1 %	71,0 %	76,8 %	85,8 %
Separación	39,7 %	53,3 %	61,6 %	69,7 %	77,6 %	82,6 %	**89,5 %**
Solo voz	68,3 %	80,7 %	85,9 %	89,2 %	91,7 %	93,2 %	96,4 %

Tabla 6.4: Comparación de resultados de clasificación de voces de la base *VoicesUy*. Método de clasificación de intervalos:"**moda**".

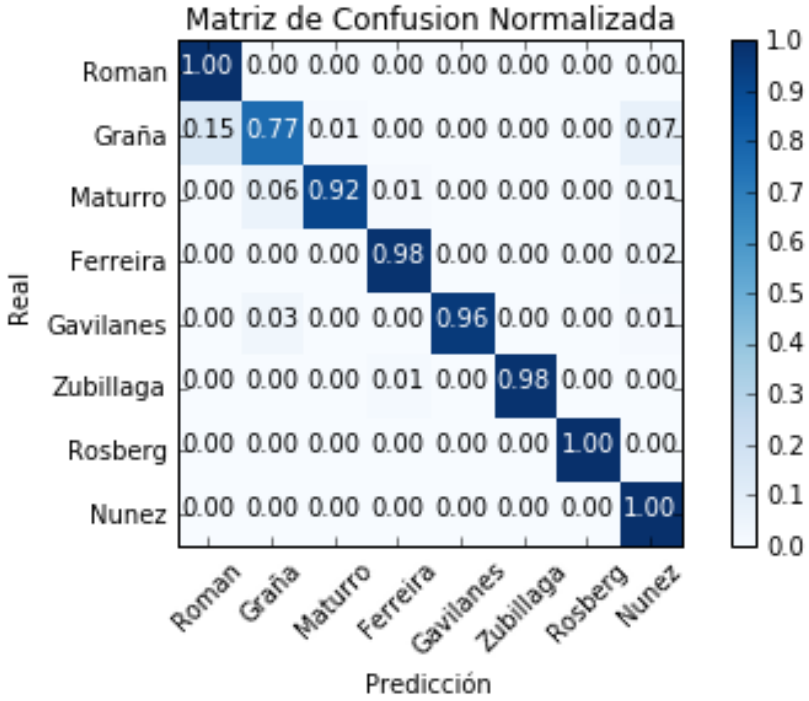
Audio	Duración intervalos (s)						
	0.1	0.5	1	2	5	10	40
Mezcla 0dB	31.4 %	43.3 %	51.6 %	60.1 %	70.3 %	76.6 %	87.7 %
Separación	37.2 %	51.0 %	59.0 %	67.6 %	79.4 %	86.7 %	**95.1 %**
Solo Voz	65.3 %	79.3 %	85.3 %	89.3 %	93.1 %	96.1 %	99.0 %

Se destaca que el método de clasificación propuesto genera una mejora de 7.4 % respecto a los audios de mezcla (la canción original) alcanzando un acierto en la clasificación de 95.1 % contra un 89.5 % obtenido con el método utilizado en los trabajos revisados [10, 11, 14, 18–20, 22, 28].

Por último, se presentan los resultados de la identificación de cantantes. La figura 6.15 muestra las matrices de confusión para el caso de clasificar intervalos de 40 segundos de duración. Al comparar los métodos se puede ver que la clasificación mejora en las 8 voces al utilizar la votación de *frames* (**moda**). Asimismo, se verifica nuevamente la mejora de la clasificación al usar intervalos de test de mayor duración (ver figura 6.16).



(a) Matriz de confusión. Método estándar, media de intervalo.



(b) Matriz de confusión. Método de voto por mayoría (Moda)

Figura 6.15: Matrices de confusión de clasificación de voces producto se separación con FChT-Wiener para intervalos de evaluación de 40 segundos.

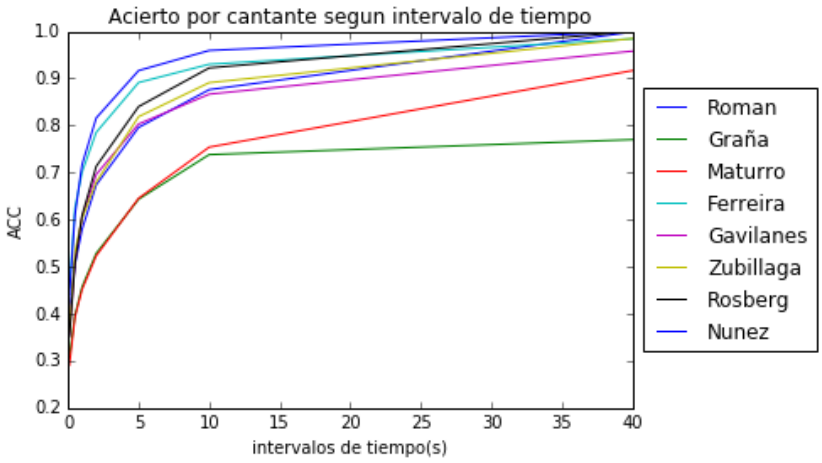


Figura 6.16: Acierto en clasificación por cantante según duración de intervalos utilizados en evaluación para una mezcla de audio SS. Método de moda (voto por mayoría).

6.6. Discusión

Se probaron diferentes técnicas de separación de fuentes por enmascaramiento en representaciones tiempo-frecuencia sobre dos bases de datos de música popular de diferentes géneros musicales. Analizando los resultados presentados, es claro que la representación FChT permite separar la voz con mayor efectividad que aplicar las mismas técnicas de enmascaramiento sobre un espectrograma. Asimismo, la técnica propuesta para la ponderación de máscaras por Wiener también genera mejoras en los indicadores SDR, SIR y SAR para la separación de la voz cantada sobre las dos bases de datos.

Se aplicó la técnica de separación FChT+Wiener como pre-procesamiento en el problema de identificación de voz cantada mostrando mejoras significativas en los resultados.

Si bien no existe una base de datos común a todos los trabajos realizados sobre el problema de clasificación de voces cantadas, se obtuvieron resultados comparables al estado del arte. Se puede ubicar el resultado obtenido en la tabla 6.5 que resume los principales trabajos en el área.

Tabla 6.5: Resumen de experimentos de los principales trabajos en identificación automática de cantantes.

Autor	N cantantes	Selección de *frames*	Separación	Características	Clasificación	Resultado
Kim 2002	17	H	-	LPC/ W-LPC	SVM	41.5 %
Zhang 2003	8		-	12 LPMCC	GMM 10	84.40 %
Tsai 2003	23	GMM		20 MFCC	GMM 64	87.8 %
Fujihara 2005	10	GMM	FFT / máscara	15 LPMCC	GMM 64	95.0 %
Mesaros 2007	13	MFCC-0	F0 síntesis	12 MFCC	GMM 10	75.0 %
Fujihara 2010	20	GMM	FFT / máscara	15 LPMCC/ DF0	GMM 64	95.3 %
Tsai 2011	10	GMM	Cepstrum	20 MFCC	GMM 64	92.5 %
Cai 2011	10	SRC	-	13 MFCC / 15 LPMCC / 13 GTCC	GMM 5	90.0 %
Lagrange 2012	10	manual	Source-Filter / NMF	MFCC	GMM 32	94.0 %
Hu 2014	22	manual	Cocleagrama / MPL	64GTCC	GMM 512	85.0 %
Kroher 2014	5	-	-	13 MFCC / 4 Vibrato / 13 Interp	SVM	88.0 %
Wang 2018	46	-	solo voz	CNN (3 capas)	Softmax	74.8 %
Massaferro 2018	**8**	**manual**	**FChT+Wiener**	**19 MFCC / Δ MFCC**	**GMM 32 (moda)**	**95.1 %**

Capítulo 7

Conclusiones y trabajos futuros

7.1. Conclusiones

En este trabajo se abordó el problema de detección automática de cantantes en archivos de audio de música polifónica. Para comprender los límites de las técnicas existentes se realizó un revelamiento detallado de las publicaciones más relevantes en la materia. Dicho revelamiento incluye el estado del arte de las técnicas de separación de fuentes y las bases de datos disponibles para el análisis de este tipo de problemas. Se pudo comprobar que las bases de datos existentes no permitían el análisis de la influencia de varias etapas de la producción fonográfica en la identificación de voces. Para poder evaluar efectos vinculados a la instrumentación, pos-producción, mezcla y masterizado es necesario contar con una base de datos con varias canciones por artista en multipistas. Este trabajo incluyó sesiones de grabación de estudio con ocho cantantes y compositores de música popular uruguaya interpretando versiones de canciones rioplatenses. De esta forma se creó la base de datos *VoicesUy*. Una selección de temas de la discografía de dichos artistas fue también analizada para crear la base de datos *AlbumsUy*. Ambas bases fueron etiquetadas de forma semiautomática para poder ser utilizadas en los experimentos de esta tesis.

Se estudió e implementó un sistema de identificación de cantante, siguiendo el enfoque más habitual que se desprende de la revisión bibliográfica. El sistema se basa en la caracterización de las voces a través de los coeficientes cepstrales MFCC (además se estudian y presentan otras características, a saber LPC, LPCC, GFCC) y el modelado con mezclas de Gaussianas (GMM). Varios de los trabajos de referencia [10, 11, 14, 18–20, 22, 28] clasifican los intervalos de audio de test según el valor promedio máximo de la log-verosimilitud, en este trabajo se muestra que utilizar votación por mayoría sobre intervalos de tiempo corto (*frames*) alcanza mejores resultados sobre las bases analizadas.

Se puede comprobar experimentalmente que la clasificación automática de voces cantadas empeora en presencia de acompañamiento musical. Varios experimentos realizados con diferentes niveles de mezcla (voz-música) permiten mostrar que cuanto menor es la relación de energía entre la voz y el acompañamiento musical,

menor es el acierto en la clasificación. Otro factor importante, poco explorado en trabajos previos, es la relación entre la duración de los archivos de audio a clasificar y el acierto de la clasificación. En esta tesis se pudo mostrar experimentalmente, sobre las bases *VoicesUy* y *AlbumsUy*, dicha vinculación, mostrando mejoras de más de diez puntos porcentuales al pasar de 10 s a 40 s de duración.

La comparación de resultados de identificación de cantantes entre las bases de datos *VoicesUy* y *AlbumsUy* permite concluir que, si bien el acompañamiento musical dificulta la identificación de cantante, un artista interpretando sus composiciones junto con su banda es más fácil de identificar que interpretando versiones. Denominamos a este comportamiento "efecto banda".

Otro resultado importante de esta tesis, es la comprobación de que el uso de la representación de tiempo–frecuencia basada en la transformada FChT permite realizar una separación de la voz cantada en contexto de música polifónica de mejor calidad que si se utilizara una representación tradicional como el espectrograma. Este resultado no solo se mostró con experimentos sobre la base *VoicesUy* sino también sobre la base de datos MedleyDB de la Universidad de Nueva York [58]. Se propuso también una técnica simple para generar una máscara suave ajustando los coeficientes por Wiener. Los resultados sobre ambas bases muestran mejoras en los indicadores SDR, SIR y SAR (distorsión, interferencia y artefactos) para la separación de la voz cantada.

Por último, se pudo mostrar empíricamente que, al aplicar la técnica de separación FChT+Wiener como pre-procesamiento en el problema de identificación de voz cantada, se obtienen mejoras significativas en el porcentaje de acierto. Los resultados de identificación de cantante obtenidos, dadas las características de la base de datos utilizada, son comparables con el estado del arte en el problema [19, 22, 24, 25, 28, 74].

7.2. Trabajos futuros

A partir del trabajo realizado en esta tesis se abren nuevas posibilidades de investigación. Respecto al análisis de variables de la producción fonográfica que pueden afectar a los algoritmos de identificación de cantante, las bases de datos creadas permitirían realizar varios experimentos. En trabajos futuros se puede evaluar la aplicación de procesos de ecualización y compresión sobre la base de datos *VoicesUy* para entrenar algoritmos de identificación de voces y comparar resultados con la base *AlbumsUy*. De esta forma se podría estudiar el "efecto pos–producción" y el "efecto álbum". En referencia a la separación de fuentes, no se han hallado antecedentes de un análisis que vincule la calidad de la separación con el desempeño de los algoritmos de identificación. Por lo que un estudio que aplique otras técnicas de separación, como las revisadas en el capítulo 2, y compare calidad de separación y desempeño de clasificación podría ser un trabajo a futuro.

Apéndice A

Resultados de experimentos de separación sobre base MedleyDB

Tabla A.1: SIR (dB) para la voz separada por los 5 métodos

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	10,42	10,64	11,16	11,58	12,27
'AlexanderRoss_GoodbyeBolero'	6,13	6,83	6,85	7,49	10,42
'Auctioneer_OurFutureFaces'	9,65	10,78	10,36	11,42	10,41
'AvaLuna_Waterduct'	20,45	20,99	23,34	25,16	26,18
'ClaraBerryAndWooldog_Stella'	18,60	18,55	19,44	19,55	20,51
'ClaraBerryAndWooldog_WaltzForMyVictims'	14,24	14,42	14,80	15,06	14,86
'Creepoid_OldTree'	13,36	13,89	14,91	15,50	16,82
'DreamersOfTheGhetto_HeavyLove'	21,04	21,64	23,78	25,03	24,01
'HezekiahJones_BorrowedHeart'	7,18	7,67	7,40	7,98	7,65
'HopAlong_SisterCities'	12,99	13,25	14,50	15,20	15,86
'LizNelson_Coldwar'	5,71	6,30	5,62	6,08	8,67
'LizNelson_Rainfall'	13,73	14,40	13,72	14,42	15,84
'Meaxic_TakeAStep'	13,79	15,27	16,63	18,58	17,91
'MusicDelta_Country1'	10,80	11,35	11,95	12,74	13,81
'MusicDelta_Disco'	21,18	21,95	25,09	25,65	25,56
'MusicDelta_Grunge'	9,48	10,74	11,21	12,56	13,17
'MusicDelta_Punk'	20,42	22,02	24,78	26,27	25,57
'MusicDelta_Reggae'	19,50	21,09	23,36	25,47	28,59
'MusicDelta_Rockabilly'	8,44	9,46	9,32	10,77	11,26
'MusicDelta_Rock'	9,86	11,20	11,31	12,76	13,69
'SecretMountains_HighHorse'	18,57	19,21	19,32	21,28	19,61
'TheSoSoGlos_Emergency'	15,14	15,98	15,98	17,87	17,07
'Wolf_DieBekherte'	15,25	15,37	17,08	17,13	17,14
Media	13,74	14,48	15,30	16,33	16,82

Tabla A.2: SAR (dB) para la voz separada por los cinco métodos

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	5,62	5,78	6,49	6,92	6,75
'AlexanderRoss_GoodbyeBolero'	4,90	5,16	5,23	5,62	6,17
'Auctioneer_OurFutureFaces'	4,68	4,86	5,29	5,64	5,23
'AvaLuna_Waterduct'	5,53	5,33	5,37	5,48	5,56
'ClaraBerryAndWooldog_Stella'	9,66	9,45	10,17	10,18	10,17
'ClaraBerryAndWooldog_WaltzForMyVictims'	9,24	9,22	9,54	9,69	9,52
'Creepoid_OldTree'	6,48	6,54	6,99	7,25	7,35
'DreamersOfTheGhetto_HeavyLove'	7,48	7,67	8,36	8,81	8,27
'HezekiahJones_BorrowedHeart'	4,21	4,31	5,06	5,23	4,92
'HopAlong_SisterCities'	5,08	4,98	5,54	5,72	5,82
'LizNelson_Coldwar'	5,41	5,45	6,14	6,18	7,12
'LizNelson_Rainfall'	6,97	7,25	7,38	7,36	9,07
'Meaxic_TakeAStep'	3,31	3,56	4,79	5,27	4,80
'MusicDelta_Country1'	6,72	6,59	7,04	7,21	7,20
'MusicDelta_Disco'	6,99	6,84	9,54	9,49	9,16
'MusicDelta_Grunge'	3,60	3,85	3,86	4,26	4,05
'MusicDelta_Punk'	4,75	4,78	6,30	6,37	5,97
'MusicDelta_Reggae'	6,09	6,27	7,63	7,88	7,73
'MusicDelta_Rockabilly'	7,39	7,45	8,30	8,66	8,57
'MusicDelta_Rock'	6,68	6,95	7,56	8,04	7,93
'SecretMountains_HighHorse'	4,27	4,06	4,08	4,07	4,00
'TheSoSoGlos_Emergency'	3,69	3,98	4,50	5,24	4,59
'Wolf_DieBekherte'	8,29	8,28	11,19	11,10	11,18
Media	5,96	6,03	6,80	7,03	7,01

Tabla A.3: SDR (dB) para la los audios de acompañamiento (residuo)

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	11,59	11,86	12,02	12,19	12,12
'AlexanderRoss_GoodbyeBolero'	5,36	6,30	5,82	6,42	7,59
'Auctioneer_OurFutureFaces'	5,71	6,17	6,43	6,97	6,86
'AvaLuna_Waterduct'	9,91	9,75	10,55	10,53	11,20
'ClaraBerryAndWooldog_Stella'	5,54	5,24	6,27	6,28	6,40
'ClaraBerryAndWooldog_WaltzForMyVictims'	6,30	6,23	6,87	6,99	7,11
'Creepoid_OldTree'	3,47	3,70	4,02	4,14	4,17
'DreamersOfTheGhetto_HeavyLove'	5,83	5,64	6,86	6,96	6,85
'HezekiahJones_BorrowedHeart'	3,00	4,37	3,78	4,59	4,87
'HopAlong_SisterCities'	4,36	4,27	4,52	4,55	4,72
'LizNelson_Coldwar'	3,01	5,01	3,75	4,68	6,43
'LizNelson_Rainfall'	9,08	10,10	9,84	10,41	11,92
'Meaxic_TakeAStep'	3,64	3,59	3,88	3,92	3,98
'MusicDelta_Country1'	7,00	7,34	7,81	8,31	8,69
'MusicDelta_Disco'	9,20	8,71	9,49	9,39	10,13
'MusicDelta_Grunge'	7,49	8,00	7,34	7,68	7,95
'MusicDelta_Punk'	8,42	8,16	7,33	7,24	7,92
'MusicDelta_Reggae'	9,44	9,32	10,47	10,52	10,14
'MusicDelta_Rockabilly'	7,05	7,75	7,97	8,69	8,54
'MusicDelta_Rock'	5,35	6,28	6,22	7,08	7,01
'SecretMountains_HighHorse'	9,76	9,67	9,77	9,81	9,83
'TheSoSoGlos_Emergency'	6,01	6,11	6,48	6,69	6,54
'Wolf_DieBekherte'	4,84	4,73	6,67	6,66	6,60
Media	6,58	6,88	7,14	7,42	7,72

Tabla A.4: SIR (dB) para la los audios de acompañamiento (residuo)

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	23,59	17,63	25,00	20,79	24,35
'AlexanderRoss_GoodbyeBolero'	18,43	12,46	22,09	17,13	20,20
'Auctioneer_OurFutureFaces'	15,22	9,84	17,70	13,51	14,89
'AvaLuna_Waterduct'	14,18	12,67	16,51	15,86	19,11
'ClaraBerryAndWooldog_Stella'	16,65	10,49	20,03	17,85	21,69
'ClaraBerryAndWooldog_WaltzForMyVictims'	18,76	10,86	21,34	18,27	18,63
'Creepoid_OldTree'	16,90	11,03	19,64	15,05	12,65
'DreamersOfTheGhetto_HeavyLove'	14,58	9,98	15,15	13,55	15,71
'HezekiahJones_BorrowedHeart'	16,34	9,86	18,00	13,43	9,82
'HopAlong_SisterCities'	15,97	11,17	16,66	13,57	11,91
'LizNelson_Coldwar'	20,66	10,70	23,91	15,37	15,58
'LizNelson_Rainfall'	20,21	14,43	23,30	19,55	23,74
'Meaxic_TakeAStep'	13,95	10,46	17,61	15,00	19,40
'MusicDelta_Country1'	15,45	10,84	19,39	16,27	20,55
'MusicDelta_Disco'	14,56	11,65	19,87	18,42	22,53
'MusicDelta_Grunge'	15,55	11,53	17,00	14,64	17,48
'MusicDelta_Punk'	13,01	10,75	13,84	12,89	16,45
'MusicDelta_Reggae'	14,19	11,93	14,76	14,27	14,95
'MusicDelta_Rockabilly'	24,04	13,01	28,34	19,81	22,80
'MusicDelta_Rock'	21,41	11,39	24,19	16,26	19,83
'SecretMountains_HighHorse'	20,56	16,75	21,64	19,61	19,68
'TheSoSoGlos_Emergency'	14,60	12,11	16,98	15,53	12,14
'Wolf_DieBekherte'	9,94	7,87	11,61	11,47	10,58
Media	16,90	11,71	19,33	16,00	17,59

Tabla A.5: SAR (dB) para la los audios de acompañamiento (residuo)

Canción	STFT	STFT Wiener	FCHT	FCHT Wiener	FCHT env
'AimeeNorwich_Child'	11,90	13,27	12,26	12,88	12,40
'AlexanderRoss_GoodbyeBolero'	5,64	7,74	5,95	6,89	7,88
'Auctioneer_OurFutureFaces'	6,35	9,03	6,84	8,24	7,75
'AvaLuna_Waterduct'	12,11	13,08	11,91	12,15	12,02
'ClaraBerryAndWooldog_Stella'	5,99	7,15	6,50	6,66	6,56
'ClaraBerryAndWooldog_WaltzForMyVictims'	6,61	8,41	7,06	7,39	7,49
'Creepoid_OldTree'	3,76	4,91	4,19	4,64	5,07
'DreamersOfTheGhetto_HeavyLove'	6,60	8,05	7,69	8,22	7,57
'HezekiahJones_BorrowedHeart'	3,30	6,25	4,01	5,39	6,97
'HopAlong_SisterCities'	4,78	5,59	4,89	5,32	5,91
'LizNelson_Coldwar'	3,12	6,73	3,81	5,19	7,11
'LizNelson_Rainfall'	9,46	12,26	10,06	11,02	12,23
'Meaxic_TakeAStep'	4,24	4,96	4,14	4,41	4,16
'MusicDelta_Country1'	7,80	10,25	8,17	9,17	9,02
'MusicDelta_Disco'	10,84	12,07	9,95	10,03	10,41
'MusicDelta_Grunge'	8,35	10,84	7,93	8,81	8,54
'MusicDelta_Punk'	10,49	11,99	8,60	8,83	8,67
'MusicDelta_Reggae'	11,38	13,04	12,64	13,07	12,02
'MusicDelta_Rockabilly'	7,16	9,50	8,02	9,08	8,73
'MusicDelta_Rock'	5,49	8,18	6,31	7,74	7,29
'SecretMountains_HighHorse'	10,18	10,70	10,09	10,34	10,36
'TheSoSoGlos_Emergency'	6,80	7,63	6,97	7,42	8,20
'Wolf_DieBekherte'	6,86	8,27	8,64	8,70	9,19
Media	7,36	9,13	7,68	8,33	8,50

Apéndice B

Resultados de experimentos de separación sobre base *VoicesUy*

Tabla B.1: Resultados de separación, interferencia SIR en dB sobre la base *VoicesUy* según método de separación.

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	15,8	17,0	17,3	18,2
	Graña	16,2	17,4	17,6	18,4
	Maturro	15,5	16,9	16,9	17,9
	Ferreira	19,5	20,3	21,0	21,6
	Gavilán	15,9	17,1	17,6	18,5
	Zubillaga	15,7	17,1	17,2	18,4
	Rosberg	15,2	16,5	16,7	17,7
	Núñez	20,7	21,7	22,1	22,7
La edad del cielo	Román	16,8	17,9	19,2	19,9
	Graña	17,9	19,1	20,1	21,1
	Maturro	16,8	17,6	18,8	19,3
	Ferreira	17,8	18,6	20,0	20,5
	Gavilán	17,1	17,9	19,3	19,8
	Zubillaga	17,9	18,9	20,3	20,9
	Rosberg	17,6	18,7	20,0	20,7
	Núñez	18,2	18,8	20,3	20,8
Pa' los músicos	Román	15,1	16,3	16,1	17,5
	Graña	19,1	20,4	21,0	22,4
	Maturro	17,0	18,2	18,8	20,1
	Ferreira	19,2	20,6	21,3	23,0
	Gavilán	19,2	20,6	20,6	22,3
	Zubillaga	18,4	19,4	20,0	21,4
	Rosberg	16,8	17,9	18,3	19,7
	Núñez	21,4	22,3	22,9	23,9
Príncipe azul	Román	10,2	11,3	12,5	13,7
	Graña	10,4	11,3	12,5	13,4
	Maturro	9,4	10,2	11,7	12,5
	Ferreira	13,5	14,4	16,5	17,3
	Gavilán	9,0	9,9	11,6	12,6
	Zubillaga	10,3	11,3	12,7	13,7
	Rosberg	8,6	9,7	10,9	12,0
	Núñez	12,5	13,4	15,3	16,2
Promesas sobre el bidet	Román	13,3	15,1	14,9	16,9
	Graña	16,6	18,0	18,2	19,8
	Maturro	15,4	17,3	17,3	19,3
	Ferreira	15,6	16,8	17,4	18,7
	Gavilán	13,5	15,1	15,0	16,7
	Zubillaga	16,6	17,9	18,4	19,9
	Rosberg	15,6	17,3	17,4	19,2

Tabla B.2: Resultados de separación, artefactos SAR en dB sobre la base *VoicesUy* según método de separación.

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	10,0	10,1	10,5	10,3
	Graña	9,7	9,5	10,0	9,6
	Maturro	9,4	9,5	9,6	9,3
	Ferreira	11,5	11,5	13,3	13,5
	Gavilán	9,8	9,7	10,1	9,8
	Zubillaga	9,2	9,3	9,5	9,6
	Rosberg	9,4	9,6	9,6	9,5
	Núñez	12,1	12,1	12,9	13,0
La edad del cielo	Román	10,0	10,0	11,0	11,1
	Graña	10,7	10,7	11,2	11,4
	Maturro	10,8	10,8	11,7	11,8
	Ferreira	12,6	12,7	13,9	14,1
	Gavilán	10,7	10,7	12,1	12,3
	Zubillaga	9,5	9,3	10,1	10,2
	Rosberg	11,3	11,4	12,5	12,6
	Núñez	12,3	12,3	13,5	13,6
Pa' los músicos	Román	7,9	8,1	8,6	9,0
	Graña	10,3	10,4	11,2	11,7
	Maturro	8,9	9,2	9,7	10,3
	Ferreira	9,9	10,2	11,2	11,8
	Gavilán	9,1	9,4	9,9	10,7
	Zubillaga	10,2	10,4	11,1	11,6
	Rosberg	7,8	8,1	8,6	9,1
	Núñez	10,9	11,0	12,0	12,4
Príncipe azul	Román	7,9	8,3	8,4	8,9
	Graña	8,1	8,2	8,2	8,4
	Maturro	6,4	6,5	6,7	6,9
	Ferreira	9,0	9,1	10,3	10,5
	Gavilán	6,3	6,4	6,8	7,0
	Zubillaga	7,2	7,4	7,5	7,8
	Rosberg	7,1	7,4	7,2	7,6
	Núñez	8,6	8,7	9,4	9,7
Promesas sobre el bidet	Román	5,5	6,1	6,0	6,9
	Graña	7,8	8,2	8,3	8,9
	Maturro	7,8	8,6	8,4	9,4
	Ferreira	7,3	7,6	8,0	8,5
	Gavilán	5,2	5,8	5,6	6,3
	Zubillaga	7,9	8,2	8,5	9,1
	Rosberg	8,1	8,9	8,7	9,8

Tabla B.3: Evaluación de calidad de separación sobre el acompañamiento musical en la base *VoicesUy*, SDR en dB

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	11,0	10,8	11,7	11,5
	Graña	10,6	10,3	11,1	11,0
	Maturro	10,6	10,6	11,2	11,1
	Ferreira	12,6	12,1	14,9	15,2
	Gavilán	10,4	10,1	11,3	11,1
	Zubillaga	9,9	9,8	10,8	10,8
	Rosberg	10,9	10,8	11,6	11,5
	Núñez	11,1	10,6	12,5	12,7
La edad del cielo	Román	11,5	11,3	12,2	12,4
	Graña	12,1	11,9	12,8	13,1
	Maturro	13,2	13,1	14,2	14,4
	Ferreira	13,0	12,3	14,4	14,6
	Gavilán	12,7	12,5	14,0	14,2
	Zubillaga	10,5	10,2	11,6	11,7
	Rosberg	12,9	12,6	14,5	14,7
	Núñez	12,6	12,0	13,7	13,9
Pa' los músicos	Román	4,9	5,6	5,7	6,5
	Graña	7,2	7,2	8,4	9,1
	Maturro	8,3	8,8	9,4	10,3
	Ferreira	7,8	8,0	9,3	10,1
	Gavilán	10,0	10,5	10,7	11,6
	Zubillaga	8,4	8,4	9,5	10,3
	Rosberg	7,9	8,5	8,8	9,5
	Núñez	9,6	9,4	10,6	11,2
Príncipe azul	Román	4,5	6,5	5,6	6,8
	Graña	3,7	5,1	4,7	5,6
	Maturro	5,0	6,7	5,9	6,8
	Ferreira	6,7	7,4	8,2	8,8
	Gavilán	4,4	6,1	5,6	6,5
	Zubillaga	4,5	6,2	5,7	6,7
	Rosberg	4,1	6,1	5,2	6,4
	Núñez	5,8	7,0	7,0	7,8
Promesas sobre el bidet	Román	8,8	10,1	9,4	10,6
	Graña	9,7	10,4	10,3	11,2
	Maturro	9,6	10,8	10,6	11,9
	Ferreira	9,1	9,7	9,8	10,4
	Gavilán	9,1	10,4	9,7	10,6
	Zubillaga	9,5	10,0	10,5	11,4
	Rosberg	10,7	12,0	11,5	12,9

Tabla B.4: Evaluación de calidad de separación sobre el acompañamiento musical en la base *VoicesUy*, SIR en dB

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	25,6	14,3	27,4	18,3
	Graña	21,0	13,6	22,1	17,6
	Maturro	21,8	14,0	24,1	17,9
	Ferreira	25,1	15,5	28,3	27,9
	Gavilán	18,9	13,1	20,7	16,9
	Zubillaga	22,2	13,4	25,6	18,4
	Rosberg	25,2	14,3	28,6	17,8
	Núñez	17,3	13,3	17,9	17,9
La edad del cielo	Román	18,4	14,3	19,0	18,5
	Graña	22,7	15,5	23,8	22,5
	Maturro	21,7	16,5	24,8	23,7
	Ferreira	24,7	15,3	27,8	26,4
	Gavilán	21,5	15,8	25,6	24,0
	Zubillaga	18,1	13,4	20,3	19,2
	Rosberg	25,2	16,0	31,9	27,4
	Núñez	21,5	14,9	23,3	22,9
Pa' los músicos	Román	18,2	10,0	20,2	18,5
	Graña	20,5	10,9	23,1	21,7
	Maturro	20,6	13,0	23,4	22,2
	Ferreira	22,2	11,8	28,2	24,4
	Gavilán	25,4	14,9	29,7	25,5
	Zubillaga	22,2	12,1	28,5	24,5
	Rosberg	22,5	13,0	25,9	23,1
	Núñez	21,2	13,0	22,6	22,6
Príncipe azul	Román	24,2	11,8	28,8	21,5
	Graña	16,0	9,4	19,3	16,3
	Maturro	17,9	11,8	19,7	17,2
	Ferreira	21,2	11,6	25,7	23,1
	Gavilán	18,8	11,1	22,6	17,5
	Zubillaga	17,0	10,9	20,9	18,3
	Rosberg	21,0	11,4	27,3	19,6
	Núñez	19,2	11,4	21,2	19,5
Promesas sobre el bidet	Román	21,6	14,8	23,0	20,2
	Graña	22,1	14,8	23,1	21,2
	Maturro	22,6	15,3	25,9	23,5
	Ferreira	19,7	13,6	20,6	18,9
	Gavilán	19,3	14,7	19,9	18,6
	Zubillaga	21,4	14,0	24,7	22,1
	Rosberg	28,0	16,9	32,3	25,2

Tabla B.5: Evaluación de calidad de separación sobre el acompañamiento musical en la base *VoicesUy*, SAR en dB

Canción	Voz	STFT	STFT+W	FChT	FChT+W
Biromes y servilletas	Román	11,2	13,6	11,8	12,5
	Graña	11,0	13,2	11,5	12,1
	Maturro	10,9	13,4	11,5	12,2
	Ferreira	12,9	14,8	15,1	15,5
	Gavilán	11,1	13,3	11,8	12,5
	Zubillaga	10,2	12,5	10,9	11,7
	Rosberg	11,0	13,6	11,7	12,7
	Núñez	12,3	14,3	14,0	14,3
La edad del cielo	Román	12,5	14,6	13,3	13,7
	Graña	12,5	14,6	13,2	13,7
	Maturro	13,9	16,0	14,6	14,9
	Ferreira	13,3	15,4	14,6	14,9
	Gavilán	13,3	15,4	14,4	14,7
	Zubillaga	11,4	13,3	12,3	12,6
	Rosberg	13,1	15,3	14,5	15,0
	Núñez	13,3	15,3	14,3	14,5
Pa' los músicos	Román	5,2	8,0	5,9	6,8
	Graña	7,4	9,9	8,6	9,4
	Maturro	8,6	11,1	9,6	10,6
	Ferreira	8,0	10,6	9,3	10,3
	Gavilán	10,1	12,6	10,8	11,8
	Zubillaga	8,6	11,1	9,6	10,5
	Rosberg	8,1	10,6	8,9	9,8
	Núñez	9,9	12,1	10,9	11,6
Príncipe azul	Román	4,6	8,2	5,6	7,0
	Graña	4,1	7,6	4,9	6,1
	Maturro	5,3	8,6	6,1	7,3
	Ferreira	6,9	9,7	8,3	9,0
	Gavilán	4,6	8,1	5,7	6,9
	Zubillaga	4,9	8,3	5,9	7,0
	Rosberg	4,2	8,0	5,2	6,7
	Núñez	6,1	9,2	7,3	8,2
Promesas sobre el bidet	Román	9,0	12,0	9,6	11,2
	Graña	9,9	12,5	10,6	11,7
	Maturro	9,8	12,8	10,7	12,3
	Ferreira	9,5	12,1	10,2	11,2
	Gavilán	9,6	12,5	10,1	11,5
	Zubillaga	9,8	12,4	10,7	11,8
	Rosberg	10,7	13,7	11,5	13,2

Referencias

[1] S Fernández González, F Vázquez de la Iglesia, M Marqués Girbau, and R García-Tapia Urrutia. La historia de la voz. *Rev Med Univ Navarra*, 50(3):9–13, 2006.

[2] Pascal Belin, Shirley Fecteau, and Catherine Bedard. Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, 8(3):129–135, 2004.

[3] Youngmoo E Kim and Brian Whitman. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, volume 13, page 17, 2002.

[4] Youngmoo E Kim. Excitation codebook design for coding of the singing voice. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 155–158. IEEE, 2001.

[5] Andre Millard. *America on record: a history of recorded sound*. Cambridge University Press, 2005.

[6] Johan Sundberg and Thomas D Rossing. The science of singing voice, 1990.

[7] Brian Whitman, Gary Flake, and Steve Lawrence. Artist detection in music with minnowmatch. In *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pages 559–568. IEEE, 2001.

[8] Michael I Mandel and Dan Ellis. Song-level features and support vector machines for music classification. In *International Society for Music Information Retrieval Conference (ISMIR)*, volume 2005, pages 594–599, 2005.

[9] Youngmoo E Kim, Donald S Williamson, and Sridhar Pilli. Towards quantifying the "album effect" in artist identification. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 393–394, 2006.

[10] Tong Zhang. Automatic singer identification. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–33. IEEE, 2003.

[11] Wei-Ho Tsai, Hsin-Min Wang, and Dwight 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*, 2003.

[12] Joseph P Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[13] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[14] Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 329–336, 2005.

[15] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *International Society for Music Information Retrieval Conference (ISMIR)*, volume 2, pages 287–288, 2002.

[16] Mathieu Ramona, Gaël Richard, and Bertrand David. Vocal detection in music with support vector machines. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1885–1888. IEEE, 2008.

[17] Martın Rocamora and Perfecto Herrera. Comparing audio descriptors for singing voice detection in music audio files. In *Brazilian symposium on computer music, 11th. san pablo, brazil*, volume 26, page 27, 2007.

[18] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 375–378, 2007.

[19] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):638–648, 2010.

[20] Wei-Ho Tsai and Hao-Ping Lin. Background music removal based on cepstrum transformation for popular singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1196–1205, 2011.

[21] C Nithin and Jini Cheriyan. A novel approach to automatic singer identification in duet recordings with background accompaniments. In *Emerging*

Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD), 2014 Annual International Conference on, pages 1–6. IEEE, 2014.

[22] Wei Cai, Qiang Li, and Xin Guan. Automatic singer identification based on auditory features. In *Natural Computation (ICNC), 2011 Seventh International Conference on*, volume 3, pages 1624–1628. IEEE, 2011.

[23] AMHJ Aertsen, PIM Johannesma, and DJ Hermes. Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38(4):235–248, 1980.

[24] Wei-Ho Tsai and Hsin-Chieh Lee. Automatic singer identification based on speech-derived models. *International Journal of Future Computer and Communication*, 1(2):94, 2012.

[25] Mathieu Lagrange, Alexey Ozerov, and Emmanuel Vincent. Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning. In *13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[26] Li Deng, Jasha Droppo, and Alex Acero. Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. *IEEE Transactions on Speech and Audio Processing*, 13(3):412–421, 2005.

[27] Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[28] Ying Hu and Guizhong Liu. Singer identification based on computational auditory scene analysis and missing feature methods. *Journal of Intelligent Information Systems*, 42(3):333–352, 2014.

[29] Nadine Kroher and Emilia Gómez. Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors. In *ICMC*, 2014.

[30] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[31] Cheng-i Wang and George Tzanetakis. Learning audio features for singer identification and embedding. 2018.

[32] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[33] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[34] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *SAPA@ INTERSPEECH*, pages 17–22, 2008.

[35] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1180–1191, 2011.

[36] Pablo Sprechmann, Pablo Cancela, and Guillermo Sapiro. Gaussian mixture models for score-informed instrument separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 49–52. IEEE, 2012.

[37] Romain Hennequin, Bertrand David, and Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 45–48. IEEE, 2011.

[38] Zafar Rafii and Bryan Pardo. Music/voice separation using the similarity matrix. In *International Society for Music Information Retrieval Conference (ISMIR), pages=583–588, year=2012*.

[39] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, 2014.

[40] Paris Smaragdis, Cedric Fevotte, Gautham J Mysore, Nasser Mohammadiha, and Matthew Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014.

[41] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 477–482, 2014.

[42] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.

[43] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

[44] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[45] Andreas Jansson, Eric J Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 323–332, 2017.

[46] Emad M Grais and Mark D Plumbley. Single channel audio source separation using convolutional denoising autoencoders. *arXiv preprint arXiv:1703.08019*, 2017.

[47] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[48] Pablo Cancela, Ernesto López, and Martín Rocamora. Fan chirp transform for music representation. In *13th Int. Conf. on Digital Audio Effects, Austria*, 2010.

[49] Martín Rocamora, Pablo Cancela, and Alvaro Pardo. Query by humming: Automatically building the database from music recordings. *Pattern Recognition Letters*, 36:272–280, 2014.

[50] Bin Gao, Wai Lok Woo, and Satnam Singh Dlay. Adaptive sparsity non-negative matrix factorization for single-channel source separation. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):989–1001, 2011.

[51] Zafar Rafii and Bryan Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 221–224. IEEE, 2011.

[52] Zafar Rafii, Antoine Liutkus, and Bryan Pardo. Repet for background/foreground separation in audio. In *Blind Source Separation*, pages 395–411. Springer, 2014.

[53] Yann Salaün, Emmanuel Vincent, Nancy Bertin, Nathan Souviraa-Labastie, Xabier Jaureguiberry, Dung T Tran, and Frédéric Bimbot. The flexible audio source separation toolbox version 2.0. In *ICASSP*, 2014.

[54] Antoine Liutkus, Derry Fitzgerald, Zafar Rafii, Bryan Pardo, and Laurent Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16):4298–4310, 2014.

[55] Bernhard Lehner and Gerhard Widmer. Monaural blind source separation in the context of vocal detection. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 309–315, 2015.

[56] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. The second 'chime'speech separation and recognition challenge: Datasets, tasks and baselines. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 126–130. IEEE, 2013.

[57] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'chime'speech separation and recognition challenge: Dataset, task and baselines. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 504–511. IEEE, 2015.

[58] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *International Society for Music Information Retrieval Conference (ISMIR)*, volume 14, pages 155–160, 2014.

[59] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang. Vocal activity informed singing voice separation with the ikala dataset. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 718–722. IEEE, 2015.

[60] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, 2010.

[61] A Michael Noll. Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309, 1967.

[62] Alan V Oppenheim. Speech analysis-synthesis system based on homomorphic filtering. *The Journal of the Acoustical Society of America*, 45(2):458–465, 1969.

[63] Bishnu S Atal and Suzanne L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The journal of the acoustical society of America*, 50(2B):637–655, 1971.

[64] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

[65] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986.

[66] Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

[67] DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[68] Shixiong CHEN, Qin GONG, and Huijun JIN. Gammatone filter bank to simulate the characteristics of the human basilar membrane [j]. *Journal of Tsinghua University (Science and Technology)*, 6:034, 2008.

[69] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription*. Springer Science & Business Media, 2007.

[70] Malcolm Slaney. Auditory toolbox. *Interval Research Corporation, Tech. Rep*, 10:1998, 1998.

[71] Anil K. Jain. Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8):651–666, June 2010.

[72] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, 1 edition, August 2013.

[73] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[74] Ying Hu and Guizhong Liu. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(4):643–653, 2015.

[75] Leon Cohen. *Time-frequency analysis*, volume 778. Prentice hall, 1995.

[76] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[77] Jonathan Le Roux, Emmanuel Vincent, Yuu Mizuno, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama. Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 89–96. Springer, 2010.

[78] Brendan Fox, Andrew Sabin, Bryan Pardo, and Alec Zopf. Modeling perceptual similarity of audio signals for blind source separation evaluation. *Independent Component Analysis and Signal Separation*, pages 454–461, 2007.

[79] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[80] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2011.

Esta es la última página.
Compilado el martes 31 julio, 2018.
http://iie.fing.edu.uy/