Sociedad de Ingeniería de Audio

Artículo de Congreso

Congreso Latinoamericano de la AES 2018
24 a 26 de Septiembre de 2018
Montevideo, Uruguay

Influencia del acompañamiento en la identificación automática de cantante en música polifónica

Pablo Massaferro,[1] Martín Rocamora[1] y Pablo Cancela[1]

[1] Universidad de la República, Facultad de Ingeniería, Instituto de Ingeniería Eléctrica
Montevideo Uruguay

pmassaferro@fing.edu.uy, rocamora@@fing.edu.uy, pcancela@@fing.edu.uy

RESUMEN

En este trabajo se aborda el problema de identificación automática de cantantes en archivos de audio de música polifónica. Se estudia la incidencia del acompañamiento musical en la identificación del cantante, analizando los efectos del nivel de energía de la voz en la mezcla y de la duración de los archivos de audios. Se adopta un método de modelado y clasificación ampliamente utilizado y se propone una variante que muestra buenos resultados. Se realizan grabaciones para construir una base de datos de cantantes profesionales interpretando las mismas canciones. Se comparan resultados de identificación utilizando álbumes editados por los mismos artistas.

0. INTRODUCCIÓN

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 [1].

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 [2].

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 motivado el uso de herramientas automáticas de clasificación y recomendación, basadas en técnicas de procesamiento de señales y aprendizaje de máquina, para gestionar la enorme oferta musical existente. En este contexto, es de especial relevancia automatizar algunas tareas, como la identificación del cantante a partir de un archivo de audio.

0.1. Descripción del problema

La identificación automática de cantante se podría pensar en analogía con el problema de reconocimiento

del hablante. Sin embargo, hay diversos aspectos que dificultan este abordaje. En particular, la voz cantada utiliza rangos dinámicos y variaciones tímbricas significativamente mayores [3]. Sumado a esto, el acompañamiento musical tiene 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 del acompañamiento musical.

Otro efecto que dificulta la tarea de identificación automática es el llamado "efecto álbum". Introducido por Whitman et al. [4], ha sido poco estudiado en la literatura. Dicho efecto está vinculado a los procesos de producción y pos-producción de un álbum (selección de la instrumentación, efectos, mezcla y masterizado) que generan cierta similitud en el sonido final de las canciones que lo integran. El problema fue mostrado por Mandel en 2005 [5] al entrenar y clasificar canciones de diferentes álbumes de los mismos artistas, mostrando un deterioro considerable de desempeño frente a entrenar y clasificar con canciones de un mismo álbum. En particular la masterización fue estudiada por Kim et al. [6] 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.

En el presente trabajo se analiza la incidencia del acompañamiento musical típico de un intérprete en la identificación de cantantes, denominándolo "efecto banda". Otros efectos poco estudiados en la literatura como el nivel de energía de la voz en la mezcla y la duración de los archivos de audio a identificar son analizados experimentalmente. Para cumplir estos objetivos fue creada la base de datos *VoicesUy*, a nuestro entender, la primera en castellano apropiada para identificación de cantantes y separación de fuentes. Además, todos los intérpretes que integran la base de datos cuentan con su propia discografía editada, lo que permite realizar estudios adicionales con grabaciones comerciales.

1. TRABAJOS RELACIONADOS

Si bien desde los años 1970 se ha trabajado en el problema de identificación del hablante [7], no es hasta los primeros años de este siglo que se comienza a investigar el problema de reconocimiento automático de cantante. La solución al problema se puede plantear como un sistema típico de aprendizaje supervisado, lo que implica generar un conjunto de atributos que caracterice a la señal y un algoritmo que modele el espacio de características y permita discriminar entre las diferentes clases. 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, es decir, la voz de cada cantante. Se han utilizado diferentes enfoques para disminuir el efecto del acompañamiento musical en la caracterización de las voces. Entre los más usados están la selección de fragmentos de audio con sistemas de detección de presencia de voz [8, 9, 10], y la separación de fuentes (basada en frecuencia fundamental [8, 9, 10], factorización de matrices [11, 12, 13, 14] o aprendizaje profundo [15, 16, 17]), o incluso una combinación de ambos. 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 nivel de la voz respecto al acompañamiento en la mezcla y el "efecto álbum" [4].

En la tabla 1 se presenta un resumen de los principales trabajos en identificación de cantantes incluyendo información sobre los coeficientes utilizados para caracterizar las voces, cantidad de cantantes de la base de datos, sistema de clasificación y porcentaje de acierto de los experimentos presentados.

2. SISTEMA DE IDENTIFICACIÓN DE CANTANTE

En este trabajo se adopta el sistema de identificación de cantante más utilizado según la bibliografía existente. Se basa en generar modelos estadísticos de un conjunto de características que representen el espectro de la voz, de forma supervisada, para luego identificar al cantante según la verosimilitud de los modelos dados un conjunto de datos nuevos. En la figura 1 se presentan de forma esquemática los procesos involucrados en el sistema de clasificación de cantantes basado en un modelo de mezcla de Gaussianas (Gaussian Mixture Model, GMM). A continuación se describen cada una de las etapas de entrenamiento y clasificación.

2.1. Extracción de características

Diferentes características han sido utilizadas para clasificar voces cantadas, incluyendo: MFCC (Mel Frequency Cepstral Coefficients), LPCC (Linear Predictive Cepstral Coding), LPMCC (Linear Predictive Mel Frequency Cepstral Coefficients) y GFCC (Gammatone Frequency Cepstral Coefficients) entre otros [10, 8, 2, 21, 23, 20, 9]. En este trabajo se utilizan los MFCC por ser las características más utilizadas. En particular, se consideran los primeros 20 coeficientes MFCC, descartando el primero por ser una medida de energía de la señal. Además, se incluyen las derivadas de primer orden para capturar información sobre la variación de los coeficientes en el tiempo. El sistema utiliza un total de 38 características.

MFCC

Las características más utilizadas en el problema de reconocimiento de cantante, los MFCC, desde su introducción en los años 80 por Davis y Mermelstein, han sido el estado del arte en el área de clasificación de voces [26].

El diseño de los coeficientes logra codificar de forma compacta información sobre el espectro de la señal. El proceso de cálculo comienza fraccionando la señal

Tabla 1: Resumen de experimentos de los principales trabajos en identificación automática de cantantes (algunos acrónimos no están definidos, se refiere al lector a las publicaciones citadas).

Autor	Cantantes	Selección	Separación	Características	Clasificación	Acierto
Kim 2002 [2]	17	H	-	LPC/ W-LPC	SVM	41.5 %
Zhang 2003 [18]	8		-	12 LPMCC	GMM 10	84.4 %
Tsai 2003 [19]	23	GMM		20 MFCC	GMM 64	87.8 %
Fujihara 2005 [8]	10	GMM	FFT / máscara	15 LPMCC	GMM 64	95.0 %
Mesaros 2007 [9]	13	MFCC-0	F0 síntesis	12 MFCC	GMM 10	75.0 %
Fujihara 2010 [10]	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 [21]	10	SRC	-	13 MFCC / 15 LPMCC / 13 GFCC	GMM 5	90.0 %
Lagrange 2012 [22]	10	manual	Source-Filter / NMF	MFCC	GMM 32	94.0 %
Hu 2014 [23]	22	manual	Cocleagrama / MPL	64 GFCC	GMM 512	85.0 %
Kroher 2014 [24]	5	-	-	13 MFCC / 4 Vibrato / 13 Interp	SVM	88.0 %
Wang 2018 [25]	46	-	solo voz	CNN (3 capas)	Softmax	74.8 %

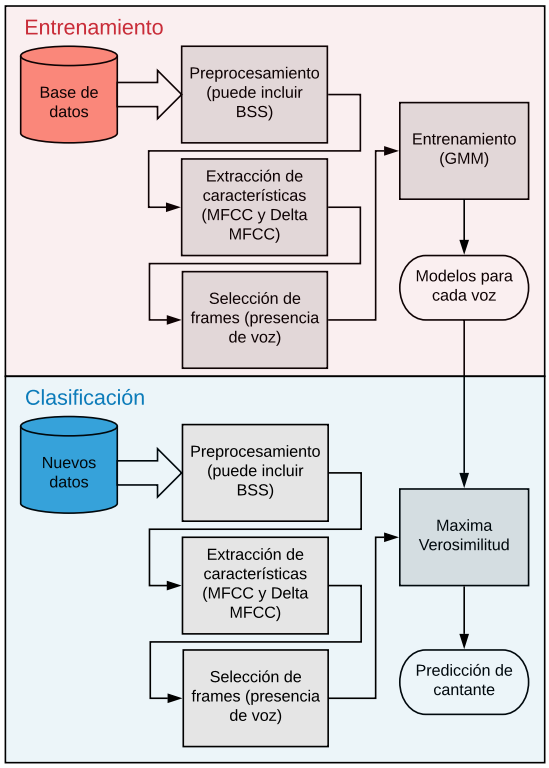


Figura 1: Esquema de sistema de clasificación de cantante basado en GMM.

en fragmentos (*frames*) utilizando una ventana de Hamming $w[n]$, en este trabajo de 23.2 ms. Para cada *frame* se calcula la transformada discreta de Fourier, DFT, como se muestra en la siguiente ecuación,

$$X_n[k] = \sum_{m=-\infty}^{+\infty} x[m]w[n-m]e^{-j\frac{2\pi k}{N}m} \quad (1)$$

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

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

De esta forma se busca aproximar la relación entre frecuencias a la percepción de alturas del oído humano. Se puede ver este proceso de filtrado como computar la energía $E[l]$ en cada filtro V_l.

$$E[l] = \sum_k |V_l[k]X_n[k]|^2 \quad (2)$$

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 (DCT) como se muestra en la ecuación 3 donde L es la cantidad de filtros triangulares.

$$MFCC[i] = \sum_{l=0}^{L-1} \log(E[l])\cos\left(\frac{2\pi i}{L}l\right) \quad (3)$$

2.2. 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 [20, 10, 8]. Otros trabajos lo hacen de forma manual utilizando una base de datos con anotaciones de presencia y ausencia de voz [22, 23]. También la selección de *frames* se puede dar dentro de un proceso de separación de fuentes [27]. En algunos trabajos de clasificación de cantantes este punto es trivial ya que se trabaja con archivos de audio de voces sin acompañamiento musical [25, 24].

En este trabajo se realiza la selección de *frames* utilizando las anotaciones de presencia de voz de las bases de datos, las cuales fueron generadas de forma semi-automática.

2.3. Clasificador GMM

Modelado por Mezcla de Gaussianas

Los Modelos de Mezcla de Gaussianas (GMM por su sigla en inglés) resultan efectivos para aproximar distribuciones de probabilidad de formas arbitrarias. 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 multivariadas como,

$$p(x) = \sum_{n=1}^{N} \omega_n \mathcal{N}(x; \mu_n, \Sigma_n) \tag{4}$$

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, Σ 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.

Clasificador por máxima verosimilitud para modelos de mezcla de gaussianas

Un sistema de clasificación basado en GMM implica el ajuste de un modelo para cada clase (cada voz en este caso). En el caso de GMM los modelos para cada clase se ajustan independientemente utilizando solo datos 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:

$$voz_{pred} = \underset{i}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log\, p(x_t|\lambda_i) \tag{5}$$

Variante en método de clasificación

Es probable que algunas modulaciones de tiempo corto de la voz no se modelen correctamente con la información de entrenamiento. Este echo puede generar que algunos *frames* tengan una muy baja probabilidad de pertenecer al modelo correcto y tengan una incidencia negativa en la clasificación usando la ecuación 5. Para evaluar el impacto de dicho comportamiento, se propone en este trabajo 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*, de la siguiente forma:

$$voz_{pred} = \underset{t=ini}{\overset{ini+T}{\operatorname{MODA}}} \left(\underset{i}{\operatorname{argmax}} \left(\log p(x_t|\lambda_i) \right) \right) \tag{6}$$

3. DATOS

Se crearon para este trabajo dos bases de datos de canciones con anotaciones: 1) *VoicesUy*, canciones populares grabadas en multipistas y cantadas por artistas profesionales y 2) *AlbumsUy*, conteniendo parte de la discografía de los mismos artistas.

3.1. Base de datos *VoicesUy*

La base de datos cuenta con ocho voces de las cuales seis son masculinas y dos femeninas. Cada cantante interpreta cinco canciones, que son versiones de temas populares rioplatenses (ver tabla 2). 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. Los cantantes son artistas uruguayos profesionales interpretando las mismas 5 canciones y la grabación de todos los instrumentos es hecha en formato multipistas por músicos profesionales.

Tabla 2: Canciones que componen la base de datos *VoicesUy* y su instrumentación

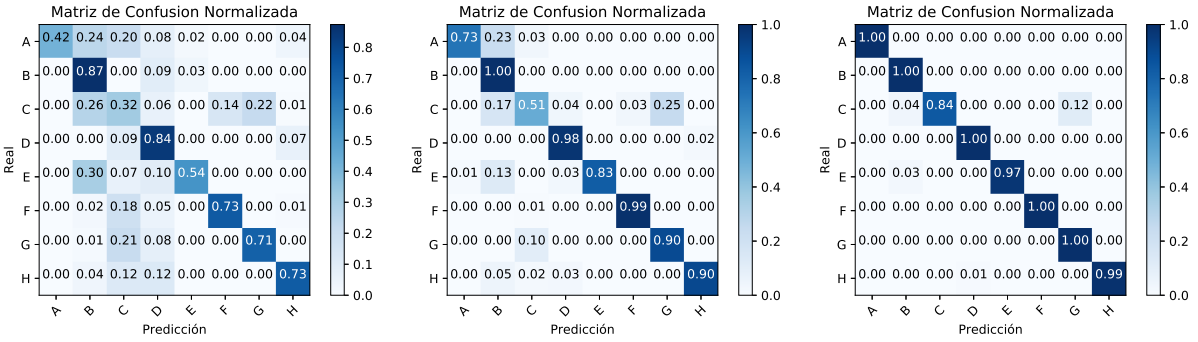
Canción	Compositor
Biromes y servilletas	Leo Masliah
La edad del cielo	Jorge Drexler
Pa' los músicos	Federico Graña
Príncipe azul	Eduardo Mateo
Promesas sobre el bidet	Charly García

3.2. Base de datos *AlbumsUy*

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* (ver tabla 3).

4. EXPERIMENTOS

En esta sección se estudia de forma experimental la incidencia sobre la identificación de cantante de los siguientes factores: nivel de energía de la voz en la mezcla, duración de los los archivos de audio a clasificar y



(a) Matriz de Confusión SNR=-3dB, T=40s (b) Matriz de Confusión SNR=0dB, T=40s (c) Matriz de Confusión SNR=+3dB, T=40s

Figura 2: Matrices de confusión de clasificación de voces en validación cruzada para diferentes relaciones de energía del acompañamiento musical (SNR)

Tabla 3: Cantantes que integran las bases *VoicesUy* y contenido de *AlbumsUy*

Voz	Cantante	Álbum
A	Nicolás Román	Rústico
B	Federico Graña	Feria
C	Diego Maturro	Mirando pa' la costa
D	Lucía Ferreira	Blues de los esclavos de ahora
E	Sebastián Gavilanes	Debut
F	Javier Zubillaga	Mentira
G	Diego Rosberg	Surcando
H	Florencia Núñez	Palabra clásica

el "efecto banda". También se presenta una comparación de desempeño del sistema de clasificación tradicional y la variante presentada en este trabajo.

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*). Respecto a la evaluación de la incidencia del duración de los archivos de audio, los experimentos fueron realizados con intervalos de validación variando entre $100\,\mathrm{ms}$ y $40\,\mathrm{s}$, 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 permite estudiar el tiempo mínimo necesario para realizar la clasificación.

4.1. Efectos del nivel de mezcla

Este experimento estudia el efecto de la relación de energía entre la voz y el acompañamiento musical en la mezcla. Para caracterizarlo se utiliza la definición de SNR (Signal to Noise Ratio) y se realizan experimentos de identificación de las ocho voces de la base de datos *VoicesUy* para tres niveles de mezcla diferentes: -3dB, 0dB y +3dB, usando intervalos de tiempo T de 40 segundos.

$$SNR = 10log_{10}\frac{||s_{voz}||^2}{||s_{acomp}||^2} \quad (7)$$

En la figura 2 se muestran las tres matrices de confusión de la identificación de voces para los SNR definidos. Se ve claramente que los errores en la identificación de cantante disminuyen al aumentar la energía de la voz en la mezcla.

Dado que el nivel de acompañamiento incide negativamente en la identificación de cantante, también se evalúa el caso ideal. Se realiza el mismo experimento pero utilizando únicamente las pistas de la las voces (lo que equivale a SNR=∞). Los resultados de la tabla 4 muestran que si se cuenta con cinco segundos de audio para identificar al cantante utilizando solamente la pista de la voz el nivel de acierto es de 91.7 %, mientras que para la mezcla con igual nivel de energía el acierto es de 76.8 %.

4.2. Efectos del la duración de los archivos de audio

La mayoría de los trabajos existentes en la literatura utilizan o bien el audio completo de una canción para identificar la voz o un intervalo de tiempo T fijo. En la tabla 4 se presentan los resultados de clasificación para siete intervalos de tiempo T diferentes. Se puede ver claramente cómo aumenta el acierto en la clasificación cuanto mayor es T. El análisis de los resultados muestra una mejora en la clasificación para todas las voces de la base de datos *VoicesUy*, como se puede ver en la figura 3.

4.3. "Efecto Banda"

Los experimentos presentados hasta este punto se realizaron utilizando la base de datos $VoicesUy$, donde todos los cantantes utilizan el mismo acompañamiento. La única diferencia entre los archivos de audio es

Tabla 4: Porcentajes de acierto en clasificación de voces para diferentes niveles de mezcla y diferentes intervalos de evaluación.

T(s)	SNR=-3dB	SNR=0dB	SNR=+3dB	SNR=∞
0,1	24,7 %	34,2 %	44,4 %	68,3 %
0,5	32,0 %	46,0 %	58,9 %	80,7 %
1	37,4 %	54,2 %	67,3 %	85,9 %
2	43,5 %	62,1 %	75,1 %	89,2 %
5	51,2 %	71,0 %	82,7 %	91,7 %
10	56,8 %	76,8 %	89,0 %	93,2 %
40	64,7 %	85,8 %	97,6 %	96,4 %

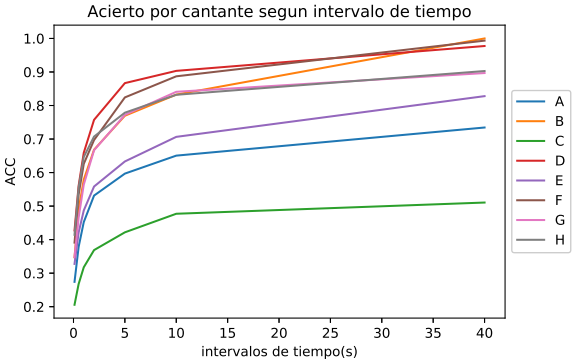


Figura 3: Acierto en identificación según largo del intervalo de tiempo T para una relación de mezcla SNR=0 dB.

la voz. Cabe señalar, que el acompañamiento modifica los coeficientes cepstrales de la mezcla, incidiendo en el modelo de cada cantante. Cuando analizamos el problema de identificación en música comercial hay que tener en cuenta todas las etapas de la producción fonográfica que inciden en el sonido final de las canciones que componen un álbum. Dentro de las etapas de producción de un álbum podemos identificar al menos cinco etapas: 1) Composición, 2) selección de instrumentación y arreglos musicales, 3) grabación, 4) mezcla, y 5) masterizado. En este trabajo definimos entonces el efecto de las etapas 1 y 2 sobre la identificación del cantante como "efecto banda".

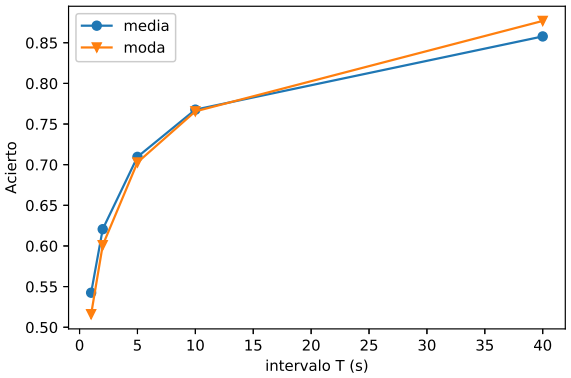
La base de datos *AlbumsUy* cuenta con exactamente las mismas ocho voces pero, en este caso, cada cantante interpreta canciones propias con su banda y su estilo propio. En este caso no se cuenta con las pistas originales. El experimento de identificación es análogo al realizado sobre *VoicesUy* pero sin alterar el nivel de mezcla. Para comparar los resultados obtenidos se utiliza como referencia el experimento con SNR=0 dB sobre *VoicesUy*. En la tabla 5 se puede ver cómo independientemente de la duración de los archivos de audio la identificación del cantante es mayor cuando éste se encuentra acompañado por su banda interpretando canciones de su repertorio.

Tabla 5: Acierto en la identificación de cantante sobre la base *AlbumsUy* vs *VoicesUy* con SNR=0 dB para diferentes intervalos de análisis T

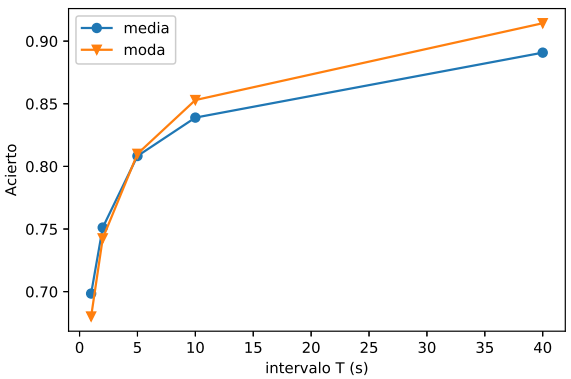
T(s)	*VoicesUy*	*AlbumsUy*
0,1	34,2 %	50,5 %
0,5	46,0 %	63,7 %
1	54,2 %	69,8 %
2	62,1 %	75,1 %
5	71,0 %	80,8 %
10	76,8 %	83,9 %
40	85,8 %	89,1 %

4.4. Comparación de métodos de clasificación

En este trabajo se propone un método diferente de identificación de cantante dada la verosimilitud de los datos en un sistema de clasificación GMM. Se clasifica con ambos métodos para las dos bases de datos. Las gráficas de la figura 4 muestran el nivel de acierto al variar T.



(a) Comparación de métodos de clasificación sobre *VoicesUy*



(b) Comparación de métodos de clasificación sobre *AlbumsUy*

Figura 4: Comparación de desempeño de métodos propuesto (moda) vs método tradicional (media)

5. DISCUSIÓN

Los experimentos realizados permiten estudiar el efecto del acompañamiento en el problema de reconocimiento de cantante en audio polifónico. La base de datos *VoicesUy* fue diseñada de forma que la única variación entre las clases a identificar fuera la voz del interprete de forma de disminuir el efecto del acompañamiento musical en la identificación. Los experimentos presentados sobre estos datos muestran que el acompañamiento musical incide negativamente en la clasificación, disminuyendo 16.4 % el acierto en el caso de clasificar fragmentos de 10 segundos de duración. Sin embargo, los resultados de los experimentos sobre la base *AlbumsUy* muestran que el acompañamiento musical puede facilitar la identificación de un cantante en contexto de música polifónica. Esto quiere decir que un cantante interpretando temas de su repertorio junto a su banda es más fácil de identificar que interpretando versiones. Al identificar voces en fragmento de audio de 10 segundos de duración sobre la base *AlbumsUy* se obtiene un acierto 7.1 % mayor que sobre la base *VoicesUy*.

Por otro lado, los experimentos realizados permiten ver la relación entre la duración de los archivos de audio a clasificar y el nivel de acierto, lo que permite evaluar aplicaciones para detección de varios cantantes cantando intercalados en una misma canción. Por último los resultados experimentales sobre las bases de datos presentadas muestran que la variante propuesta al método de clasificación GMM genera un mejor desempeño cuando la duración de los archivos de audio a clasificar son mayores a 10 segundos. Sobre la base *AlbumsUy* se logra un desempeño de 91.4 % con el método propuesto contra un 87.7 % del método tradicional.

6. CONCLUSIONES Y TRABAJOS FUTUROS

Este trabajo se concentra en el estudio de la influencia del acompañamiento musical en el problema de identificación de cantantes en música polifónica. Para ello, se construyó una base de datos de grabaciones multipista, *VoicesUy*, con ocho cantantes interpretando cinco canciones en idioma español. Esta base de datos es la primera de este tipo en idioma español y resulta apropiada para la investigación en los problemas de identificación de cantantes y separación de fuentes.

En este trabajo se adoptó un sistema de clasificación de voces basado en la extracción de coeficientes cepstrales de frecuencias Mel (MFCC) y el modelado por mezcla de gaussianas (GMM). 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. 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.

Se mostró de forma experimental que el acompañamiento musical, si bien dificulta la identificación de cantante, aporta información relevante para el problema. Se pudo ver que la clasificación de voces alcanza valores superiores, para cualquier intervalo de tiempo que se analice, cuando el acompañamiento musical es el típico de un artista. Para realizar dichos experimentos se creo una base de datos (*AlbumsUy*) con álbumes comerciales de los mismos ocho cantantes. Se propuso una variante en el mecanismo de integración temporal de la verosimilitud de los datos, que para el problema de identificación de cantantes genera mejoras significativas de desempeño.

En trabajos futuros se aplicarán diferentes técnicas de separación de fuentes de forma de medir su impacto en la identificación de cantantes y se explorarán posibles modificaciones al sistema de clasificación.

REFERENCIAS

[1] Pascal Belin, Shirley Fecteau, and Catherine Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[2] 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*, 2002, vol. 13, p. 17.

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

[4] 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*. IEEE, 2001, pp. 559–568.

[5] Michael I Mandel and Dan Ellis, "Song-level features and support vector machines for music classification.," in *ISMIR*, 2005, vol. 2005, pp. 594–599.

[6] Youngmoo E Kim, Donald S Williamson, and Sridhar Pilli, "Towards quantifying the.ªlbum effectïn artist identification.," in *ISMIR*, 2006, pp. 393–394.

[7] Bishnu S Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[8] 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 *ISMIR*, 2005, pp. 329–336.

[9] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods.," in *ISMIR*, 2007, pp. 375–378.

[10] 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*, vol. 18, no. 3, pp. 638–648, 2010.

[11] 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*. IEEE, 2012, pp. 49–52.

[12] Zafar Rafii and Bryan Pardo, "Music/voice separation using the similarity matrix.," in *ISMIR*, 2012, pp. 583–588.

[13] 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*, vol. 31, no. 3, pp. 107–115, 2014.

[14] 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*, vol. 31, no. 3, pp. 66–75, 2014.

[15] 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*, vol. 24, no. 9, pp. 1652–1664, 2016.

[16] 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*. Springer, 2017, pp. 258–266.

[17] 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)*, 2017, pp. 323–332.

[18] Tong Zhang, "Automatic singer identification," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 1, pp. I–33.

[19] 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.

[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*, vol. 19, no. 5, pp. 1196–1205, 2011.

[21] Wei Cai, Qiang Li, and Xin Guan, "Automatic singer identification based on auditory features," in *Natural Computation (ICNC), 2011 Seventh International Conference on*. IEEE, 2011, vol. 3, pp. 1624–1628.

[22] 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.

[23] Ying Hu and Guizhong Liu, "Singer identification based on computational auditory scene analysis and missing feature methods," *Journal of Intelligent Information Systems*, vol. 42, no. 3, pp. 333–352, 2014.

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

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

[26] 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*, vol. 28, no. 4, pp. 357–366, 1980.

[27] 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*, vol. 23, no. 4, pp. 643–653, 2015.