VIII Congresso Ibero-americano de Acústica
acústica 2012
1-3 outubro
Évora - Portugal

ANÁLISIS DE LA SEÑAL ACÚSTICA MEDIANTE COEFICIENTES CEPSTRALES BIO-INSPIRADOS Y SU APLICACIÓN AL RECONOCIMIENTO DE PAISAJES SONOROS

Xavier Valero, Francesc Alías

Grup de Recerca en Tecnologies Mèdia, La Salle-Universitat Ramon Llull
xvalero@salleurl.edu, falias@salleurl.edu

Resumen
En el contexto del reconocimiento, la identificación y la clasificación de señales acústicas ambientales, resulta indispensable contar con un conjunto de descriptores capaces de representar con precisión las señales sonoras. En este trabajo, presentamos una modificación bio-inspirada de los tradicionales Mel Frequency Cepstral Coefficients (MFCC). La novedad radica en el uso de un conjunto de filtros originalmente diseñados para modelar la respuesta espectral del sistema auditivo humano: los llamados Gammatone Cepstral Coefficients (GTCC). En este trabajo los GTCC son aplicados sobre la descripción y reconocimiento de señales de paisajes sonoros que, con independencia del algoritmo de aprendizaje escogido, mejoran significativamente las tasas de reconocimiento obtenidas por los populares MFCC.

Palabras-clave: descriptores de la señal, filtros auditivos, Gammatone, reconocimiento de patrones, paisajes sonoros.

Abstract
When it comes to recognize, identify and classify environmental sound signals, it becomes essential to have a set of descriptors able to accurately represent the sound signals. In this work, we present a bio-inspired modification of the traditional Mel Frequency Cepstral Coefficients (MFCC). The novelty lies in the use of a set of filters originally designed to model the spectral response of the human auditory system: the so-called Gammatone Cepstral Coefficients (GTCC). In this work, the GTCC are applied to the description and recognition of soundscape signals, which, regardless of the selected learning algorithm, significantly improve the recognition rates obtained by the popular MFCC.

Keywords: signal descriptors, auditory filters, Gammatone, pattern recognition, soundscapes.

PACS no. 43.60.Bf, 43.72.Ar

1 Introducción

En aplicaciones como detección, reconocimiento, clasificación o computación del índice de similitud resulta esencial el análisis efectuado sobre la señal sonora [1]. La finalidad no es otra que extraer un conjunto de parámetros, descriptores o características relevantes de la señal, que la definan y permitan identificar su contenido acústico. Una práctica habitual es efectuar dicho análisis (mediante la Transformada Discreta de Fourier) en el dominio frecuencial de la señal, dada la relevancia de la información espectral para su identificación [2]. La aplicación de un banco compuesto por múltiples

filtros paso banda nos permite suavizar la forma del espectro, compactando el número de puntos de la Transformada Discreta de Fourier.

En el campo de la Acústica, los filtros de octava y de tercio de octava son de uso generalizado. Si nos referimos específicamente al campo del procesado de la señal sonora (y en especial el habla), los filtros perceptuales Mel son los más habituales. En concreto, el análisis se suele completar con el cálculo del logaritmo de la energía de cada banda de frecuencias y con el cálculo de la Transformada Directa del Coseno, obteniendo así los Mel Frequency Cepstral Coefficients (MFCC) [3].

Dichos MFCC, gracias a su simplicidad y eficacia, se han convertido en un estándar de facto en los campos del reconocimiento del habla, así como para la identificación de parlante o de los eventos sonoros ambientales [4]. En este trabajo, se propone una modificación de los ampliamente utilizados MFCC mediante la utilización de filtros Gammatone, los cuales fueron originalmente diseñados para modelar la respuesta del oído humano [5]. En concreto, se aplican para la descripción y reconocimiento de paisajes sonoros o *soundscapes*, como una mejora de trabajos anteriores [4].

Este documento está organizado como sigue. La segunda sección presenta los descriptores propuestos y detalla su cálculo. La tercera sección trata sobre la aplicación de dichos descriptores al problema del reconocimiento de paisajes sonoros. La cuarta sección muestra los resultados experimentales obtenidos y, finalmente, la quinta y última sección recoge las conclusiones y las líneas de futuro de este trabajo.

2 Coeficientes Cepstrales Bio-inspirados

En este capítulo se describen, en primer lugar, los filtros Gammatone, incluyendo: como modelan la respuesta del sistema auditivo humano, la definición de su función de transferencia y la escala frecuencial que emplean. En segundo lugar, se detalla la computación de los descriptores derivados que servirán para representar la señal acústica.

2.1 Filtros bio-inspirados

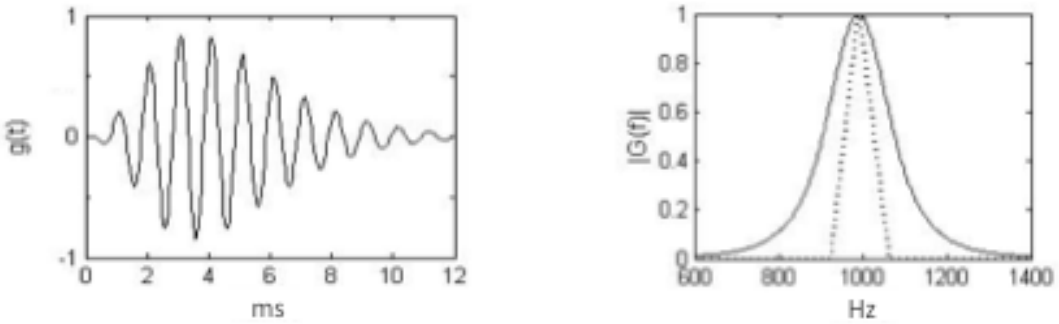
Hay varias razones fisiológicas, psicológicas y prácticas que apoyan el uso de la función Gammatone (GT) para modelar el análisis espectral realizado por el sistema auditivo humano [5]. En primer lugar, y desde el punto de vista fisiológico, la respuesta impulsional del filtro Gammatone se correlaciona con la obtenida por los mamíferos [6]. En segundo lugar, y desde el punto de vista psicológico, las propiedades de la selectividad de frecuencia medida fisiológicamente en la cóclea y aquellas psicofísicamente medidas en humanos convergen, ya que: *i)* la respuesta en magnitud de un filtro GT de orden cuatro es muy similar a la que comúnmente se utiliza para representar la respuesta del filtro auditivo humano, *ii)* el ancho de banda del filtro corresponde a una distancia fija en la membrana basilar, y *iii)* el GT es un filtro de fase mínima y, a pesar de que la fase del filtro auditivo humano es desconocida, parece razonable la suposición de que es cercana a la mínima [5]. Por último, y desde un punto de vista práctico, un filtro GT de orden n puede ser aproximado por un conjunto de n filtros GT de primer orden colocados en cascada [7], los cuales tienen una implementación digital particularmente eficiente. Por lo tanto, el banco de filtros GT resultante proporciona un buen compromiso entre la precisión con que simula el filtrado coclear y el coste computacional requerido para su implementación.

2.2 Función de transferencia

El filtro GT toma su nombre de la respuesta impulsional g(t), que es el producto de una función de distribución Gamma y un tono sinusoidal centrado en la frecuencia f_c, siendo calculado como [5]:

$$g(t) = K t^{(n-1)} e^{-2\pi Bt} \cos(2\pi f_c t + \varphi) \qquad t > 0 \tag{1}$$

donde K es un factor de amplitud; B determina la duración de la respuesta impulsional, y por lo tanto, el ancho de banda del filtro, n es el orden del filtro y determina su calidad; f_c es la frecuencia central, y φ la fase. Las diferencias en comparación con un filtro triangular Mel típico son notables en el dominio espectral, siendo la función de transferencia del filtro GT menos abrupta (véase la fig. 1.b).



(a) (b)

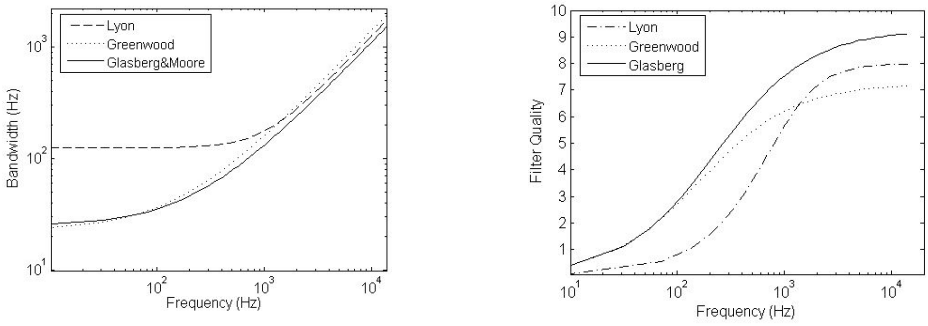
Figura 1 – Respuesta impulsional de un filtro Gammatone centrado a 1KHz; (b) respuesta en frecuencia del mismo filtro (línea continua) y de un filtro Mel a la misma frecuencia (línea discontinua).

2.3 Equal Rectangular Bandwidth

La duración del filtro GT (B, en (1)), se relaciona con el *Equal Rectangular Bandwidth* (ERB), una medida del ancho de banda del filtro auditivo medido psicoacústicamente en cada punto de la cóclea [5]. Un filtro ERB modela la integración espectral derivada de la canalización efectuada por las células ciliadas internas, que envían señales de un cierto ancho de banda hacia el cerebro (véase (2)). Para el caso específico de un filtro GT de cuarto orden, B es 1,019 veces la ERB [29].

$$ERB = \left[\left(\frac{f_c}{EarQ} \right)^n + minBW^n \right]^{\frac{1}{n}} \tag{2}$$

donde f_c es la frecuencia central en Hertz; $EarQ$ es la calidad asintótica del filtro a altas frecuencias; $minBW$ es el ancho de banda mínimo a bajas frecuencias, y n es el orden de aproximación.
En la literatura encontramos tres modelos diferentes de filtros ERB. Greenwood propuso un filtro de primer orden, con $EarQ$ = 7,23 y $minBW$ = 22,85 [8]. Glasberg & Moore propuso filtros más abruptos ($EarQ$ = 9,26) de primer orden y ancho de banda mínimo similar ($minBW$ = 24,7) [9]. Por otra parte, Lyon sugirió usar filtros de segundo orden con ancho de banda más amplio a bajas frecuencias ($minBW$ = 125) y calidad media ($EarQ$ = 8) [10]. Como se representa en la Figura. 2, el modelo propuesto por Lyon produce los filtros más anchos y redondeados en las frecuencias más bajas, mientras que Glasberg&Moore presenta los filtros más agudos.



(a) (b)

Figura 2 - (a) Ancho de banda y (b) factor de calidad en función de la frecuencia central del filtro f_c para los 3 modelos ERB.

2.4 Banco de filtros Gammatone

La distribución de los filtros Gammatone a lo largo del ancho de banda espectral que se quiere analizar es un tema clave cuando queremos diseñar un banco de filtros de este tipo. En el sistema auditivo humano, existen alrededor de 3000 células ciliadas internas a lo largo de la cóclea, cada una con una frecuencia de resonancia y un ancho de banda (ERB) determinados. Por lo tanto, el sistema auditivo humano se compone de alrededor de 3000 filtros paso banda [11]. Normalmente, con el fin de reducir el coste computacional, el modelado se lleva a cabo con un número inferior de filtros solapados espectralmente. La frecuencia central de cada filtro f_{ci} viene dada por la ecuación (3).

$$f_{ci} = (f_{high} + EarQ\,minBW)\,e^{-\frac{i\,step}{EarQ}} - EarQ\,minBW \tag{3}$$

donde f_{high} es la frecuencia más alta del banco de filtros (típicamente, la frecuencia de muestreo de Nyquist); *EarQ* y *minBW* son los parámetros de ERB (véase (2)); *i* es el índice del filtro GT; *step* es la distancia entre los filtros, que depende tanto en la frecuencia más baja considerada, f_{low}, y el número de filtros, *N*. El paso espectral *step* puede calcularse como [11]:

$$step = \frac{EarQ}{N}\log\left(\frac{f_{high} + EarQ\ minBW}{f_{low} + EarQ\ minBW}\right) \tag{4}$$

Step toma el rango de valores [0,1], donde los valores cercanos a cero indican que los filtros se superponen casi por completo, mientras que los valores cercanos a uno significa que casi no hay solapamiento espectral. Por lo tanto, o bien el número de filtros *N* o bien el paso espectral *step* deben estar fijados de antemano cuando diseñamos un banco de filtros Gammatone.

2.5 Gammatone Cepstral Coefficients

En esta sección se discute el proceso de extracción de características de la señal utilizando los filtros Gammatone considerados en este trabajo. El análisis, tal y como se muestra en la Figura 3, comienza con el ventaneo de la señal acústica y el cálculo de la transformada rápida de Fourier. A continuación, el banco de filtros GT suaviza el espectro de la señal obtenida. Específicamente, dicho banco de filtros

puede ser configurado en base a los siguientes parámetros: modelo ERB (Lyon, Greenwood o Glasberg & Moore), número de filtros N y orden de los filtros n. Finalmente, siguiendo el mismo procedimiento para la obtención de los MFCC, se ejecutan dos procesos adicionales: el cálculo del logaritmo y la Transformada Discreta del Coseno (DCT). El primero modela la sonoridad de la señal percibida por el ser humano, mientras el segundo decorrela las salidas logarítmicas del banco de filtros, consiguiendo así una mejor compactación de la energía [3]. Los descriptores obtenidos son denominados Gammatone Cepstral Coefficients (GTCC), y se calculan como:

$$GTCC_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \log(X_n) \cos\left[\frac{\pi n}{N}\left(m - \frac{1}{2}\right)\right] \quad 1 \le m \le M \tag{5}$$

donde X_n es la señal sonora en el dominio espectral; N es el número de bandas del banco de filtros Gammatone; y M es el número de GTCC. Habitualmente $M<<N$, resultando en una reducción de la dimensionalidad de los datos.

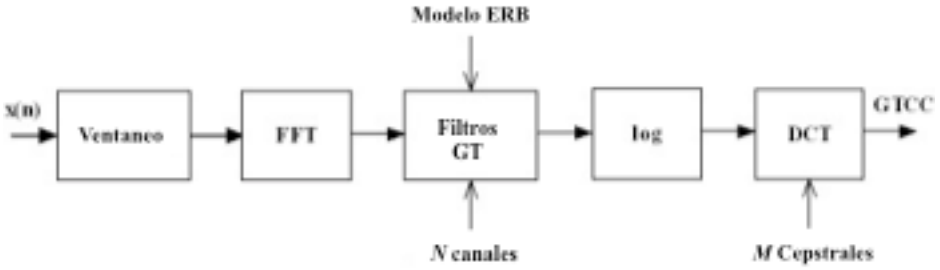


Figura 3 - Diagrama de bloques del proceso de cálculo de los Gammatone Cepstral Coefficients.

3 Evaluación experimental

En esta sección aplicaremos los GTCC al reconocimiento de señales de paisajes sonoros. Primero se describe el sistema diseñado para llevar a cabo dicha tarea. A continuación se detalla la base de datos empleada y se finaliza con la explicación acerca del procedimiento de evaluación.

3.1 Sistema de reconocimiento automático de paisajes sonoros

El sistema de reconocimiento automático de paisajes sonoros se compone de dos procesos principales: la parametrización de la señal acústica y el reconocimiento de patrones (véase la Figura 4). El primero comienza con proceso de ventaneo, por el cual la señal de audio se fracciona en segmentos de 30 ms [4]. Posteriormente, se extraen el conjunto de características (en este caso, GTCC o MFCC) de cada segmento de la señal, produciendo así patrones representativos de cada paisaje sonoro. Cabe señalar que los patrones resultantes contemplan la evolución temporal de las características de la señal, por lo que la dimensionalidad de los patrones resultantes es bastante elevada. Con la finalidad de compactar los patrones sin perder demasiada información temporal, se divide la señal en tres partes calculando el vector promedio en cada una de ellas, como en [12].

Finalmente, se utiliza un algoritmo de aprendizaje supervisado para realizar el reconocimiento del paisaje sonoro. En una primera etapa, el algoritmo aprende de una serie de patrones conocidos, y en una etapa posterior, en base al conocimiento adquirido, el algoritmo efectúa el reconocimiento de nuevas señales acústicas (es decir, nuevas grabaciones de paisajes sonoros).

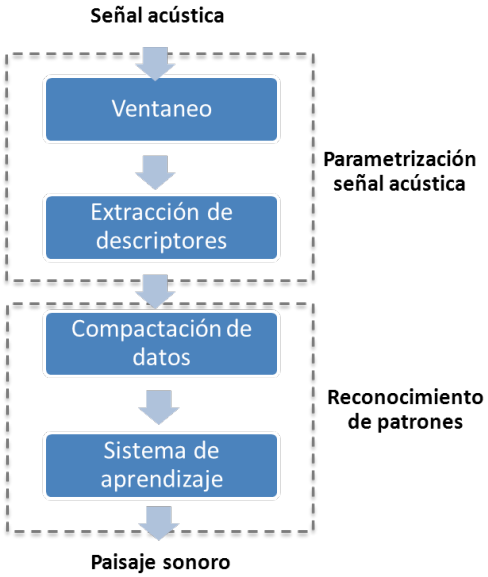


Figura 4 – Diagrama de bloques del sistema de reconocimiento de paisajes sonoros desarrollado.

3.2 Corpus de muestras sonoras

Con el fin de evaluar la técnica de parametrización propuesta, se preparó un corpus de datos compuesto por muestras de 15 paisajes sonoros distintos, incluyendo espacios interiores (ocio, trabajo y transporte) y exteriores (natural y urbano). Se realizaron grabaciones utilizando un sonómetro Bruel & Kjaer 2250 con módulo de grabación sonora, obteniendo grabaciones de alta calidad, con señal de audio muestreado a 48 KHz y codificado en un formato de codificación sin pérdidas (archivos WAV). El corpus se compone de las muestras de audio grabadas y se complementa con muestras extraídas de [13], a fin de incluir datos de diferentes orígenes.

Cada paisaje sonoro está representado por un conjunto de entre 150 y 300 muestras, grabadas en al menos cuatro localizaciones distintas. El tamaño total del corpus es de 3.500 muestras, lo que equivale a casi 4 horas de datos sonoros. Al igual que en trabajos anteriores [1], [4], la longitud de cada muestra se limita a 4 segundos. En la Tabla 1 se muestran los datos de audio comprendidos en el corpus, incluyendo el nombre, la categoría y el número de muestras de cada paisaje sonoro.

3.3 Procedimiento de evaluación

Los experimentos se efectuaron siguiendo el esquema de reconocimiento mostrado en la Figura 4. Además de los descriptores propuestos (GTCC), se utilizaron también los populares MFCC, con el fin de poder comparar el comportamiento de ambos. Los bancos de filtros GT y Mel utilizados para el cálculo de GTCC y MFCC, respectivamente, presentan el mismo ancho de banda (del mínimo audible, 20Hz, hasta 11KHz) y el mismo número de filtros (48). En cuanto al modelo ERB empleado para la computación de los GTCC, Glasberg&Moore fue elegido después de diversos test preliminares.
En cuanto al sistema de aprendizaje, se utilizaron hasta cuatro algoritmos diferentes, con el fin de desligar los resultados obtenidos al comportamiento de un determinado paradigma de aprendizaje. En concreto, se utilizaron los siguientes: Decision Tree (DT), K-Nearest Neighbours (KNN), Neural Networks (NN) y Support Vector Machines (SVM). Para más información sobre el funcionamiento de este tipo de algoritmos, se recomienda la lectura de [14]-[16].

La evaluación se realizó utilizando un sistema de validación cruzada 10-*kfold*, mediante el cual el sistema es entrenado con un 90% de los datos y testeado con el restante 10%, repitiendo el proceso 10 veces con diferentes instancias. Para cada repetición, se computa la tasa de reconocimiento, calculada como el porcentaje de muestras correctamente clasificadas con respecto al total de muestras testeadas.

Tabla 1 – Composición del corpus de paisajes sonoros empleado en la evaluación experimental de los coeficientes propuestos.

Categoría	Nombre	Muestras
Exterior - Rural	Playa	251
	Campo	150
Exterior - Ciudad	Calle - tráfico	253
	Calle peatonal	227
	Parque	200
Interior - Ocio	Biblioteca	173
	Restaurante	194
	Estadio	296
Interior – Entorno de trabajo	Aula	200
	Oficina	288
	Fábrica	250
Interior – Medios de transporte	Estación	198
	Coche - interior	300
	Autobús – interior	284
	Tren - interior	236
TOTAL		3500

4 Resultados

En este capítulo se detallan los resultados de los experimentos realizados para evaluar el comportamiento de los descriptores GTCC con respecto a los MFCC de referencia. Tal y como recoge la Tabla 2, los descriptores propuestos obtienen unas tasas de reconocimiento superiores a las obtenidas por los tradicionales MFCC usando cualquiera de los cuatro algoritmos de aprendizaje. Empleando árboles de decisión (DT) (el algoritmo que presenta unos resultados más pobres), la mejora media de los GTCC respecto a los MFCC es de un 3.9%. Dicho margen aumenta al emplear el algoritmo KNN (mejora del 4,7%) y las redes neuronales (NN) (mejora del 5,6%). Finalmente, SVM conducen a las tasas de reconocimiento más elevadas: un 85,9% con los MFCC y un 88,1% con los GTCC. Si promediamos los resultados con los cuatro algoritmos, se obtiene una mejora promedio del 4,1% usando los GTCC en lugar de los tradicionales MFCC.

Adicionalmente, se ha procedido a realizar un estudio detallado con tal de demostrar la significancia estadística de los resultados obtenidos. Además de las distribuciones de las tasas de reconocimiento obtenidas al combinar ambos descriptores con cada uno de los cuatro algoritmos (véase la Figura 4), se realizó la prueba estadística t-Student [17]. La mejora introducida por los GTCC fue probada estadísticamente, dado que la probabilidad obtenida en dicho test fue inferior a 0.001 en los cuatro casos estudiados.

Tabla 2 – Tasa de reconocimiento media obtenida al combinar los coeficientes MFCC y GTCC con cada uno de los cuatro algoritmos de aprendizaje usados sobre el corpus de paisajes sonoros considerado.

Algoritmo	MFCC (%)	GTCC (%)
DT	67,6	71,5
KNN	81,6	86,3
NN	75,4	81,0
SVM	85,9	88,1
Promedio	77,6	81,7

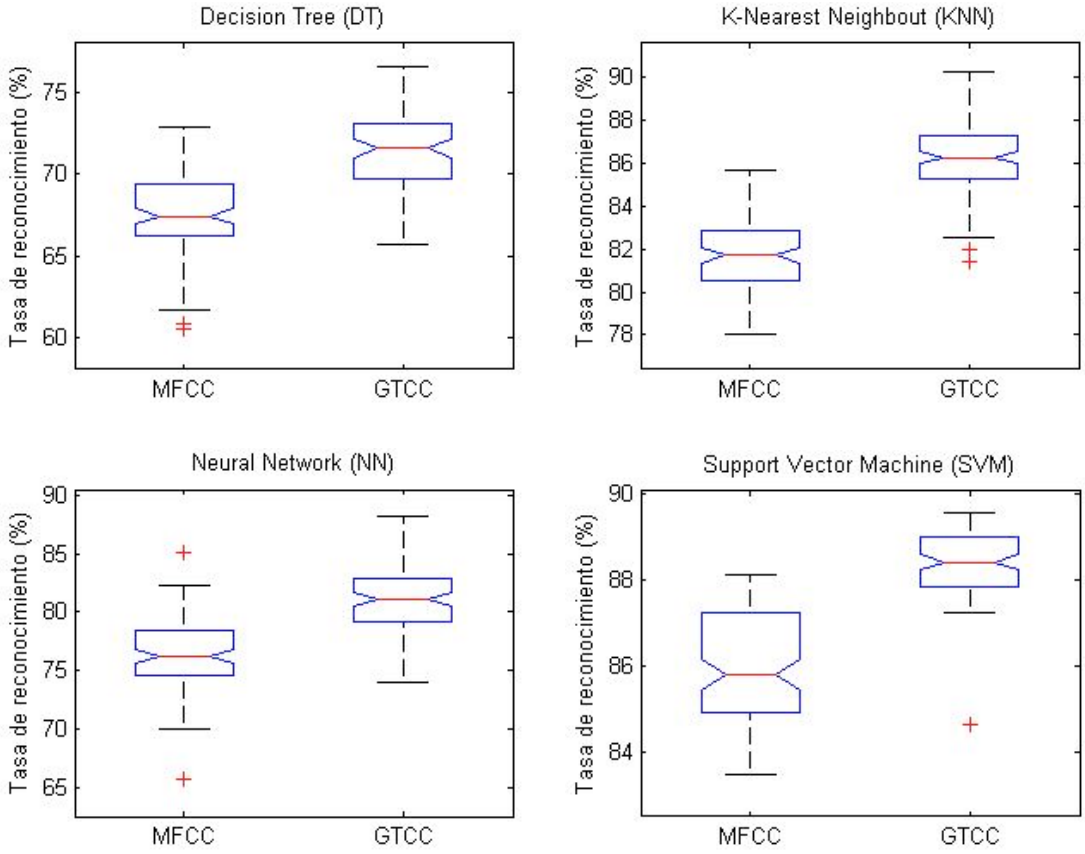


Figura 4 – Diagramas de caja de las tasas de reconocimiento obtenidas al combinar los coeficientes MFCC y GTCC con cada uno de los cuatro algoritmos de aprendizaje utilizados sobre el corpus de paisajes sonoros considerado.

5 Conclusiones

En la presente comunicación se ha presentado una modificación bio-inspirada de los tradicionales MFCC. Los descriptores propuestos, llamados Gammatone Cepstral Coefficients (GTCC), han sido utilizados para representar señales acústicas grabadas en diferentes paisajes sonoros. Los experimentos llevados a cabo demuestran que la utilización de dichos descriptores permiten un reconocimiento

satisfactorio de los distintos paisajes sonoros, mejorando las tasas de reconocimiento obtenidas cuando se consideran los populares MFCC junto con cada uno de los cuatro algoritmos de aprendizaje testeados (en promedio, en un 4.1%). Un estudio estadístico de los resultados mediante diagramas de cajas y pruebas t-Student demuestran la significancia de los mismos. En trabajos sucesivos, nos planteamos aplicar los descriptores propuestos a otros tipos de señales acústicas, tales como las fuentes de ruido ambientales, y estudiar cómo implementar la técnica para su funcionamiento en tiempo real.

Referencias

[1] Chu, S.; Narayanan, S.; Jay Kuo, C.-C. Environmental sound recognition with time-frequency audio features. *IEEE Trans. Audio, Speech and Lang. Processing*. Vol. 17(6), 2009, pp. 1142-1158.

[2] Rabiner, L.; Juang, B. *Fundamentals of speech recognition*, Prentince Hall, 1993.

[3] Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process*ing. Vol. 28(4), 1980, pp. 357–366.

[4] Valero, X.; Farré, P.; Alías, F. Comparison of Machine Learning Techniques for the Automatic Recognition of Soundscapes. *Forum Acusticum 2011*, Aalborg (Dinamarca), Junio 2011.

[5] Patterson, R. D.; Holdsworth, J. A functional model of neural activity patterns and auditory images, *W. A. Ainsworth (Ed.), Advances in Speech, Hearing and Language Processing*, Vol. 3 part B. London: JAI Press, 1996, pp. 554-562.

[6] Carney, L.H.; Yin, C.T. Temporal encoding of resonances by low-frequency auditory nerve fibers: Single fibre responses and a population model. *Journal of Neurophysiology.* Vol. 60, 1998, pp. 1653-1677.

[7] Holdsworth, J.; Nimmo-Smith, I.; Patterson, R.D.; Rice, P. Spiral Vos Final Report, Part A: The Auditory Filter bank (Annex C). *APU report 2341*, 1988.

[8] Greenwood, DD. A cochlear frequency-position function for several species—29 years later. *Journal of the Acoustical Society of America*. Vol. 87(6), 1990, pp. 2592-2605.

[9] Glasberg, B.R.; Moore, B.C.J. Derivation of auditory filter shapes from notched-noise data. *Hearing Research.* Vol. 47, 1990, pp. 103-108.

[10] Slaney, M. Lyon's Cochlear Model. *Apple Technical Report #13*, Apple Computer Corporate Library, Cupertino, CA 95014, 1988.

[11] Slaney, M. An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. *Apple Technical Report #35*, Apple Computer Library, Cupertino, CA 95014, 1993.

[12] Valero, X.; Alías, F. Gammatone Wavelet Coefficients for sound classification in surveillance applications. *EUSIPCO'12,* Bucarest, Agosto 2012.

[13] The Freesound Project [Online]. Available: http://www.freesound.org/.

[14] Bishop, C. M. *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 2003.

[15] Jones, M.T. *Artificial Intelligence - A Systems Approach*. Infinity Science Press, Higham, 2008.

[16] Cristianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2000.

[17] Jackson, S. L. *Research Methods and Statistics: A Critical Thinking Approach*, John Benjamins Publishing Company, 2009.