TECNIACUSTICA Cáceres · 2011

42º CONGRESO ESPAÑOL DE ACÚSTICA
ENCUENTRO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON ENVIRONMENTAL ACOUSTICS
AND ON BUILDINGS ACOUSTICALLY SUSTAINABLE

MONITORIZACIÓN AUTOMÁTICA DE FUENTES DE RUIDO AMBIENTALES

PACS.: 43.50.Rq.

Valero González, Xavier; Alías Pujol, Francesc.
GTM - Grup de Tecnologies Mèdia, La Salle. Universitat Ramon Llull.
Quatre Camins, 2, 08022 Barcelona. España
Tel: +34 932 902 476
Fax: +34 932 902 470
E-mail: xvalero@salle.url.edu; falias@salle.url.edu

ABSTRACT

Traditionally, research in the audio recognition field has focused on the analysis of voice and music signals. However, in recent years a growing interest in the recognition of environmental sound has been observed, given its multiple potential applications such as automatic identification of acoustic scenes, their integration into surveillance and security systems or the automatic monitoring of environmental noise sources. This paper deals with the latter application, focusing on the search of acoustic signal descriptors that enable an optimal representation of the different noise sources.

RESUMEN

Tradicionalmente, la investigación en el ámbito del reconocimiento del audio se ha centrado en el análisis de señales de voz y música. Sin embargo, en los últimos años se está observando un interés creciente en el reconocimiento de sonidos ambientales, dadas sus múltiples aplicaciones potenciales tales como la identificación automática de escenas acústicas, su integración en sistemas de seguridad y televigilancia o la monitorización automática de fuentes de ruido ambientales. Este trabajo trata esta última aplicación, centrándose en la búsqueda de descriptores de la señal acústica que permitan representar las diferentes fuentes de ruido de manera óptima.

INTRODUCCIÓN

Dentro del campo del reconocimiento de audio, la identificación de sonidos ambientales ha recibido históricamente un interés menor comparado con el reconocimiento del habla o de los instrumentos musicales. La dificultad para identificar y clasificar los eventos sonoros parece ser mayor en este último caso, debido a las complejas características espectro-temporales de este tipo de señales [1]. Sin embargo, el interés en el reconocimiento fuentes sonoras ambientales está creciendo hoy en día, con aplicaciones tales como la identificación de escenas acústicas (*soundscapes*) [2], los sistemas de televigilancia basados en información sonora [3] o el reconocimiento de diversas fuentes de ruido ambiental [4]-[6].

Nuestra investigación plantea la monitorización automática de las fuentes de ruido ambiental que típicamente se pueden encontrar en ciudades y zonas urbanas. Nos referimos a las fuentes de ruido ambiental que provocan un impacto en la vida diaria de los ciudadanos y que, en cumplimiento de la Directiva Europea sobre ruido ambiental [7], los niveles de ruido que originan deben ser reportados en los mapas de ruido que exige la UE a los diferentes Estados miembros. Mediante el uso de sistemas de monitorización con reconocimiento de fuentes de ruido ambiental se puede obtener de forma automática una descripción del entorno acústico, incluyendo qué tipo de fuentes de ruido están presentes y cuál es el nivel equivalente generado por cada una de ellas. Todo ello sin ser necesaria la presencia de un técnico para validar la procedencia de los niveles de presión sonora registrados en cada instante por el equipo de medición.

Según nuestro conocimiento, hasta el momento solo se han presentado algunas propuestas con objetivos similares al de este trabajo. En [4], los autores se centran exclusivamente en fuentes de tráfico rodado, mientras que en [5] se obvian el ruido ferroviario y de aeronaves. En [6] el planteamiento es similar, no obstante, se trata de un trabajo que data de varios años y usa unos descriptores acústicos desfasados respecto al estado del arte actual. En el presente trabajo, se presenta el núcleo del sistema de reconocimiento de fuentes de ruido ambiental desarrollado, evaluando una larga serie de descriptores de la señal acústica y mostrando los resultados preliminares conseguidos hasta la fecha.

PRESENTACIÓN DEL SISTEMA DE RECONOCIMIENTO

En esencia, se trata de resolver un problema de reconocimiento de patrones aplicado al entorno acústico. La metodología a seguir se divide generalmente en dos partes, tal y como muestra la Figura 1. En primer lugar, se requiere de un proceso de parametrización de la señal sonora para hacerla tratable, extrayendo un conjunto de descriptores que permitan identificar la procedencia de las señales (fuentes de ruido ambiental). Estos descriptores serán óptimos si: *i)* ensalzan las diferencias entre muestras de sonidos producidos por fuentes distintas (ej. coche-avión), y *ii)* enfatizan la asociación de muestras de sonidos producidos por un mismo tipo de fuente (ej. coche). Mediante dicho proceso, se construyen patrones representativos de cada una de las fuentes de ruido a reconocer. Dado que dichos patrones pueden adquirir dimensionalidades muy elevadas, es una buena opción utilizar métodos de compactación de información que, además, contribuyen a reducir el coste computacional del sistema. En este trabajo, utilizamos el método de análisis de componentes principales (PCA) [8].

En segundo lugar, se requiere de un proceso de aprendizaje, de modo que el sistema, emulando el proceso cognitivo humano, aprenda a reconocer nuevas señales sonoras. Este proceso se ejecuta presentando patrones de diferentes eventos sonoros conocidos al sistema, el cual aprende a reconocer la fuente de ruido ambiental asociada a cada patrón. En concreto, los patrones usados son aquellos construidos a partir de los parámetros y descriptores de las señales sonoras extraídos en la fase anterior (una vez aplicado PCA). Con esta información, el sistema aprende y adquiere un conocimiento concreto sobre las características de las fuentes de ruido, el cual es utilizado para el reconocimiento de los nuevos eventos sonoros a identificar.

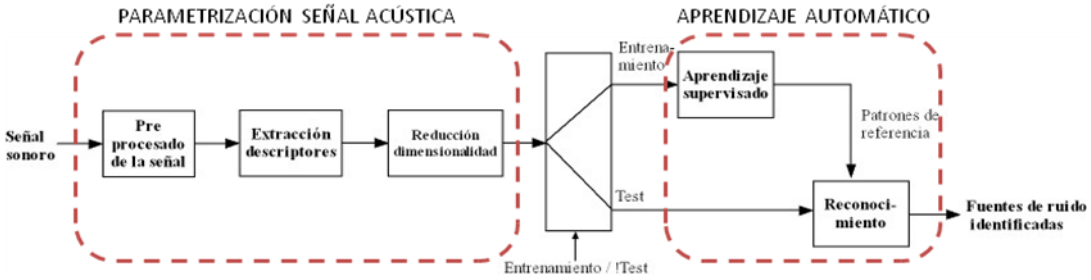

Figura 1. Diagrama de bloques de un sistema de reconocimiento de fuentes de ruido ambientales.

DESCRIPTORES DE LA SEÑAL ACÚSTICA

Según las diversas referencias bibliográficas consultadas, la selección de los descriptores de la señal acústica (primer bloque de la Figura 1) es un punto clave para conseguir un sistema reconocedor inteligente robusto y fiable [3]. Así pues, hemos procedido a realizar un estudio del estado del arte, seleccionado un conjunto de 13 descriptores de señal acústico utilizados en el campo del reconocimiento de audio. A continuación se describen cada uno de ellos:

Linear Predictive Coefficients (LPC): técnica que modela el espectro de potencia de la señal con una función de la forma (1), donde *w* es la frecuencia en radianes, σ es una constante y *A (p)* es un polinomio de orden *p*. En el presente trabajo, se han extraído 13 coeficientes empleando el método de la autocorrelación [9].

$$f = \sigma^2 / \left| A(e^{jw}) \right|^2 \qquad (1)$$

Linear Prediction Cepstral Coefficients (LPCC): coeficientes cepstrales c_i obtenidos a partir de los LPC a través de la fórmula recursiva (2) donde *G* es la ganancia del filtro, a_i son los coeficientes LPC y *n* su orden [9] (*n*=12 en este trabajo).

$$c_0 = \log(G) \qquad c_m = -a_m + \frac{1}{m}\sum_{k=1}^{m-1}\left[-(m-k)a_k c_{(m-k)}\right] \quad 1 \le m \le p \qquad (2)$$

Perceptual Linear Predicition (PLP): añade los conceptos de psicofísica de la audición para hacer que la técnica LPC esté más acorde con la percepción humana. Los conceptos empleados en cuestión son: bandas críticas, curva de ponderación isofónicas y ley de potencia de intensidad-sonoridad [8].

Mel Frequency Cepstral Coefficients (MFCC): se obtienen aplicando un banco de filtros Mel (aproximan mejor la percepción auditiva humana que la escala Hertz) a la FFT de la señal. A la salida de cada filtro, se calcula el logaritmo de la energía resultante en cada banda frecuencial y, finalmente, se aplica la transformada discreta del coseno (DCT) con el fin de compactar mejor la información [8]. En el presente trabajo, hemos seleccionado un vector de 13 MFCC, como en [10].

Discrete Wavelet Coefficients (DWC): coeficientes basados en la Transformada de Wavelet, técnica de procesamiento de la señal que intenta mejorar uno de los inconvenientes de la Transformada de Fourier, que no es otro que la pérdida de localización espacial de la información. Mediante versiones desplazadas y escaladas de una función base (llamada "función madre") es capaz de reconstruir la señal original. En este trabajo, hemos empleado la función "Daubechies", como en [3].

Mel Frequency Discrete Wavelet Coefficients (MFDWC): modificación de los MFCC consistente en aplicar la transformada de Wavelet en lugar de la DCT sobre la salida del banco de filtros Mel (una vez aplicada a ésta la función logaritmo) [3].

MPEG-7 features: de entre todos los descriptores de audio definidos en el estándar MPEG-7 [11], consideramos la envolvente del espectro de audio (ASE), ya que es el descriptor que logró los mejores resultados en trabajos anteriores [10]. En este trabajo, proponemos realizar un post-procesamiento a este indicador consistente en tres pasos: *i)*, transformación en escala de decibelios; *ii)* normalización energética según su valor RMS; y *iii)* compactación de la energía mediante la DCT.

Spectral Flatness (SF): representa la desviación del espectro de potencia de la señal con respecto a un espectro plano para cada una de las bandas espectrales predefinidas *i*. Puede verse como una medida de la correlación de la señal con una señal aleatoria de ruido blanco. Se calcula como el cociente entre el promedio geométrico y el promedio energético del espectro de la señal *X[k]* en cada una de las bandas definidas entre las frecuencias de corte superior hi_i e inferior lo_i (3) [8]:

$$SF_i = \frac{\sqrt[hi_i - lo_i]{\prod_{k=lo_i}^{hi_i} X[k]}}{\frac{1}{hi_i - lo_i}\sum_{k=lo_i}^{hi_i} X[k]} \qquad (3)$$

Sub-Band Energy Ratio (SBER): se divide el espectro de potencia de la señal en una serie de sub-bandas. El parámetro SBER da una idea de la distribución de energía a lo largo de las *N* sub-bandas (en este trabajo, *N*=4 como en [4]) con respecto a la energía de la señal del espectro total [4] (ver ecuación (4)).

$$SBER_i = \frac{\sum_{k \in B_i} |X[k]|}{\sum_{k=0}^{N-1} |X[k]|} \qquad (4)$$

Spectral Centroid (SC): mide el centro de gravedad del espectro de potencia. Se calcula según la ecuación (5) [8]:

$$SC = \frac{\sum_{k=0}^{N-1} k\ |X[k]|}{\sum_{k=0}^{N-1} |X[k]|} \qquad (5)$$

Spectral Roll-off (SRO): ancho de banda en el que se concentra la mayor parte de la energía del espectro de potencia total. Informa sobre la asimetría de la forma espectral [8]. Se calcula como (6), donde *TH* es un umbral cuyo valor suele estar comprendido entre 0,85 y 0,99 [8]. En el presente trabajo tomamos el valor medio de 0,92.

$$SRO = \max_m \left(\sum_{k=0}^{m} |X[k]| \leq TH \cdot \sum_{k=0}^{N-1} |X[k]| \right) \qquad (6)$$

Short Term Energy (STE): energía de la señal a lo largo de los fragmentos analizados de la señal (en este trabajo, de 30 ms de duración), Se calcula como la suma de la amplitud al cuadrado de la señal [4].

$$STE = \sum_{n=0}^{N-1} |x[n]|^2 = \sum_{k=0}^{N-1} |X[k]|^2 \qquad (7)$$

Zero Crossing Rate (ZCR): número de veces que la señal cruza el cero en términos de amplitud [3]. El valor obtenido es proporcional a la frecuencia de la señal: valores altos de ZCR representan señales con presencia de altas frecuencias

SISTEMA DE CLASIFICACIÓN

El sistema de aprendizaje automático empleado para ejecutar el reconocimiento de patrones de fuentes de ruido ambiental es una red neuronal artificial (ANN). Las redes neuronales son modelos computacionales biológicamente inspirados que generan parametrizaciones no lineales entre un conjunto de variables de entrada y salida [12]. Como en anteriores trabajos [2], la red neuronal implementada presenta una arquitectura de perceptrón multicapa [12], el número de capas ocultas y de nodos ocultos (1 y 100, respectivamente) fueron seleccionados experimentalmente. La capa de salida contiene 6 nodos correspondientes a las 6 fuentes de ruido a ser identificadas, donde cada salida y_j puede calcularse mediante la ecuación (8). En todos los nodos se han seleccionado funciones de activación φ logarítmica-sigmoide (dado que los datos de entrada son mapeados previamente al rango [0,1]) y los pesos sinápticos w_{ki} y w_{jk} son inicializados de forma aleatoria.

$$y_j = \varphi_j \left(\sum_{k=0}^{N_2} w_{jk}^{L_2} \left\{ \varphi_k \left(\sum_{i=0}^{N_1} w_{ki}^{L_1} x_i \right) \right\} \right) \qquad (8)$$

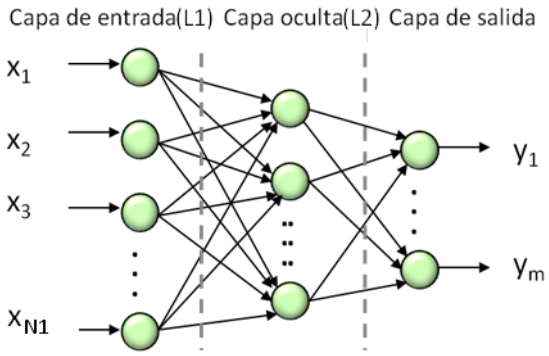

Figura 2. Esquema de la red neuronal reconocedora de fuentes de ruido ambientales.

EVALUACIÓN EXPERIMENTAL

Base de Datos Sonora

La base de datos sonora utilizada para entrenar y testear el sistema ha sido elaborada mediante la realización de una campaña de medidas para la grabación de las fuentes de ruido en entornos reales. Se consideran las siguientes categorías de fuentes: vehículos ligeros, vehículos pesados, motocicletas, aeronaves, trenes y ruido industrial (véanse sus características en la Tabla 1). Las grabaciones se realizaron utilizando un sonómetro Bruel & Kjaer 2250 equipado con módulo de grabación de audio, obteniendo grabaciones de alta calidad con frecuencia de muestreo de 48 KHz y usando un sistema de codificación sin pérdidas. Con el fin de garantizar la variabilidad de los datos, las fuentes de ruido se registraron en, al menos, seis ubicaciones diferentes, todas ellas con condiciones diversas. Los lugares seleccionados debían cumplir además una serie de requisitos. En primer lugar, el sonido debía ser grabado sin interrupción durante el intervalo representativo de ese tipo de fuente de ruido (ej., aproximación, paso y alejamiento de un vehículo). En segundo lugar, el nivel de ruido de fondo tenía que ser lo suficientemente bajo para que ninguna fuente sonora pudiera interferir en la grabación. En tercer lugar, se debía asegurar la ausencia de cualquier elemento que pudiera ejercer de barrera acústica, modificando el nivel y el espectro recibido en el micrófono.
La base de datos obtenida se compone de 90 muestras de audio para cada una de las 6 categorías de fuente de ruido consideradas, fijando la duración de las mismas en 4 segundos con el fin de poder mantener información sobre la evolución temporal de las señales.

Fuente de ruido	Características	
Vehículos ligeros	Turismos, todoterrenos, furgonetas.	Vías urbanas y carreteras secundarias
Vehículos pesados	Camiones	
Motocicletas	Ciclomotores (50 cc), motocicletas (>125 cc)	
Aeronaves	Operaciones de despegue y aterrizaje.	
Trenes	Cercanías, regionales, alta velocidad, de carga. Vías rectas y en curva.	
Industrial	Chimeneas, maquinaria, sistemas de refrigeración, etc.,	

Tabla 1. Características de las muestras sonoras grabadas de cada fuente de ruido ambiental.

Configuración del Test de Evaluación

Los archivos sonoros de 4s de duración son analizados usando ventanas de 30 ms, con una separación entre ventanas de 15 ms, como en [10]. De cada fragmento se extraen los descriptores sonoros correspondientes, creando así los patrones representativos de cada fuente de ruido ambiental. El grupo de 540 archivos sonoros son distribuidos siguiendo un esquema de validación cruzada de 4 conjuntos: el 75% de las muestras son usadas para el entrenamiento mientras que el 25% restante son usadas para el testeo del sistema de reconocimiento de fuentes sonoras implementado, siendo tal procedimiento repetido 4 veces con conjuntos diferentes de entrenamiento-testeo.

El espacio *n*-dimensional de los patrones sonoros se reduce mediante la aplicación de PCA basándose únicamente en la información aportada por el conjunto de archivos de entrenamiento. El mismo cambio de coordenadas obtenido de la fase de entrenamiento se aplica sobre el conjunto de archivos de testeo. En este punto, se entregan los archivos a la red neuronal la cual, una vez entrenada, efectúa el reconocimiento de las fuentes de ruido ambientales correspondientes a cada archivo sonoro. Las tasas de acierto se calculan como el promedio de las muestras correctamente clasificadas, en tanto por ciento.

6. RESULTADOS

Comparativa de Descriptores

Los resultados en términos de porcentaje de acierto medio (%) para cada descriptor sonoro se detallan en la Figura 3. Tal y como se puede observar, existen un conjunto formado por 4 descriptores (ZCR, SRO, SC y STE) cuyas tasas de aciertos son realmente bajas, debido a su baja capacidad para extraer información representativa de las fuentes de ruido. Este hecho puede explicarse por la sencillez de los descriptores, puesto que todos ellos se componen de un único valor numérico por fragmento de la señal analizado. En un segmento medio, encontramos los descriptores SBER, SF, LPC, LPCC y MFDWC, que presentan resultados notables con tasas de acierto alrededor del 75% (a excepción del primero, con un porcentaje de aciertos menor, del 67%). En un tercer grupo, encontramos los descriptores que demuestran un mejor rendimiento para el objetivo planteado en este trabajo: MPEG-7, MFCC, PLP y DWC.

A continuación, se ha realizado un análisis estadístico de los resultados obtenidos por este último grupo de descriptores con la finalidad de probar la representatividad de las tasas de aciertos medias anteriormente obtenidas. Tal y como se extrae de la figura 4, los descriptores MPEG-7 y MFCC conducen a resultados estadísticamente superiores a los descriptores PLP y DWC. Sin embargo, tanto las diferencias entre MPEG7 y MFCC como entre PLP y DWC no son significativas.

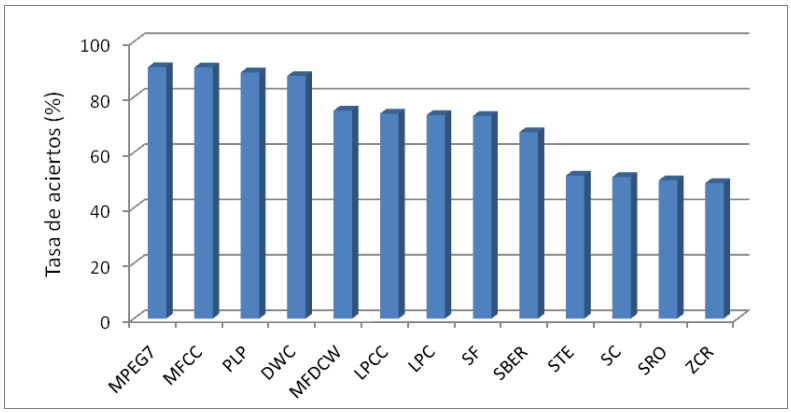

Figura 3. Promedio de tasa de aciertos obtenidos por cada uno de los descriptores de audio considerados

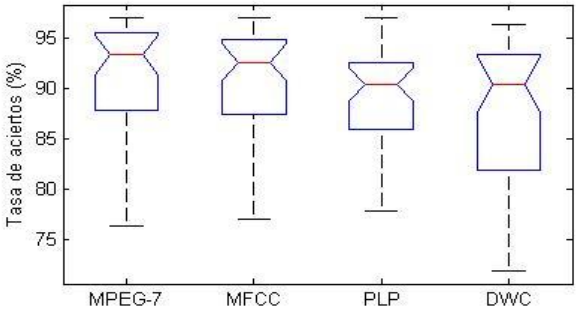

Figura 4. Representación de la distribución de los resultados (en tasa de acierto) obtenidos por los cuatro descriptores de audio con un mayor rendimiento.

Reconocimiento por fuente de ruido

Adicionalmente, se ha calculado (usando el descriptor MPEG-7) las tasas de acierto específicas para cada una de las fuentes de ruido ambiental con la finalidad de conocer si existen diferencias significativas entre ellas. Para ello tomamos los coeficientes MPEG-7, puesto son el descriptor que han obtenido una tasa de aciertos promedio mayor (según hemos visto en el apartado anterior. Tal y como muestra la figura 5, las fuentes de ruido con unas tasas de acierto menores son aquellas referidas a las fuentes de tráfico rodado (vehículos ligeros, motocicletas y vehículos pesados).

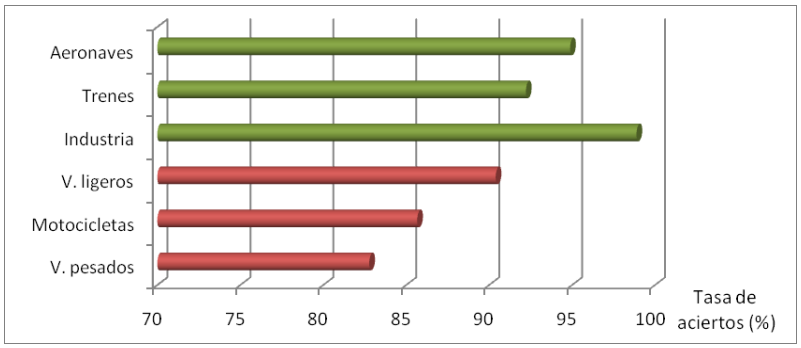

Figura 5. Promedio de tasa de acierto para cada una de las fuentes de ruido.

Con el fin de analizar los motivos por los que dicha tipología de fuentes de ruido obtienen unas tasas de acierto menores, se ha procedido al cálculo de la matriz de confusión (véase la Tabla

2). Esta herramienta relaciona las soluciones entregadas por el sistema con las soluciones reales, y nos permite visualizar rápidamente el origen de los errores producidos. En este caso, los errores del sistema más frecuentes son debidos a las confusiones entre vehículos pesados y vehículos ligeros y entre vehículos pesados y motocicletas, respectivamente.

		Solución real					
		Aeronave	Tren	Industria	V. ligero	Motocicleta	V. pesado
Salida del sistema	Aeronave	94,9	1,7	0,4	0,1	1,1	0,4
	Tren	1,6	92,2	0,1	2,4	1,0	3,1
	Industria	0,7	1,3	98,9	1,0	1,4	0,7
	V. ligero	0,0	2,3	0,0	90,4	1,0	6,1
	Motocicleta	0,3	0,6	0,1	1,1	85,5	7,0
	V. pesado	2,6	1,9	0,4	5,0	10,0	82,7

Tabla 2. Matriz de confusión de la clasificación efectuada por el sistema, en %.

CONCLUSIONES

En este trabajo, se ha propuesto un sistema de reconocimiento de patrones usando una técnica de aprendizaje automático basada en una red neuronal para la identificación automática de fuentes de ruido ambiental. En concreto, se han comparado distintos descriptores sonoros del estado del arte .Como se puede observar a partir de los resultados obtenidos, la correcta parametrización de la señal acústica se trata de un punto fundamental dentro del sistema de reconocimiento, comportando una variación muy acusada en las tasas de aciertos recopiladas en los experimentos. De entre los distintos descriptores testeados, tanto MPEG-7 como MFCC conducen a unas tasas de acierto muy buenas (del orden del 90%). En futuros trabajos, nos centraremos en mejorar el reconocimiento de fuentes de tráfico rodado, ya que según los resultados obtenidos, son las fuentes de ruido ambientales con un mayor número de confusiones en la fase de identificación de fuentes sonoras ambientales.

REFERENCIAS

[1] B. Gygi, “Factors in the identification of environmental sounds”, PhD Thesis, Indiana University. July 2001.
[2] X. Valero, P. Farré, F. Alías, “Comparison of Machine Learning Techniques for the Automatic Recognition of Soundscapes”, *Proc. Forum Acusticum 2011*.
[3] A. Rabaoui, Z. Lachiri, N. Ellouze, "Towards an optimal feature set for robustness improvement of sounds classification in a HMM-based classifier adapted to real world background noise", *Proc. 4th Int. Multi-Conference on Systems, Signals & Devices*, 2007.
[4] M. Sobreira Seoane, A. Rodriguez Molares, J.L. Alba Castro, "Automatic classification of traffic noise", *Proc. Acoustics'08 Paris*, 2008.
[5] B. Defréville, P. Roy, C. Rosin, F. Pachet, "Automatic recognition of urban sound sources", *Proc. 120th AES Convention*, 2006.
[6] C. Couvreur, V. Fontaine, P. Gaunard, C.G. Mubikangiey, "Automatic classification of environmental noise events by Hidden Markov Models", *Applied Acoustics*, vol. 54, no. 3, pp. 187-206, 1998.
[7] EU Directive, "Directive 2002/49/EC of the European parliament and the Council of 25 June 2002 relating to the assessment and management of environmental noise", *Official Journal of the European Communities, L 189/12* , July 2002.
[8] H. Kim, N. Moreau, T. Sikora, “MPEG-7 audio and beyond, Audio content indexing and retrieval”. *John Wiley & Sons*, 2005.
[9] L.Rabiner, B. Juang, “Fundamentals of speech recognition”, *Prentince Hall*, 1993.
[10] X. Valero, F. Alías: Applicability of MPEG-7 low level descriptors to environmental sound source recognition”, *Proc. 2010 EAA Euroregio Congress*.
[11] ISO/IEC FDIS 15938 4:2001, “Information Technology Multimedia Content Description Interface - Part 4 : Audio”.
[12] C. M. Bishop, “Neural Networks for Pattern Recognition”, *New York: Oxford Univ. Press*, 2003.