TECNIACUSTICA ELCHE · 2022

53º CONGRESO ESPAÑOL DE ACÚSTICA
XII CONGRESO IBÉRICO DE ACÚSTICA

RELEVANCIA DE LOS CANALES ACÚSTICOS EN LA IDENTIFICACIÓN DE ESCENAS SONORAS

PACS: 43.60.Acoustic signal processing.

Fraile, Rubén; Gutiérrez-Arriola, Juana Mª; Sáenz-Lechón, Nicolás; Osma-Ruiz, Víctor J.; García-Barrios, Guillermo.
CITSEM, Universidad Politécnica de Madrid. Campus Sur. Campus Sur. Edificio La Arboleda. C/ Alan Turing 3. 28031 Madrid. España.
Tel: +34 910 673 379; E-mail: r.fraile@upm.es

Palabras Clave: Análisis cepstral, Acústica arquitectónica, Procesamiento de señales acústicas, Agrupación de patrones.

ABSTRACT.

Acoustic scenes comprise a set of acoustic sources surrounded by a common physical environment and recorded by a microphone. The acoustic channel between sources and microphone is mainly defined by the physical environment, and usually assumed to be independent of both sources and microphone. This paper addresses the question of whether the acoustic channel bears information that is relevant for the identification of the acoustic scene or not. For this purpose, the dataset corresponding to the Acoustic Scene Classification task (task 1A) in DCASE 2019 Challenge is analyzed by calculating the cepstral mean of all audio recordings corresponding to the same physical location. Non-parametric analyses of these cepstral means algorithm allow concluding that the acoustic channels bear information that is relevant for acoustic scene classification. A more detailed study of cluster means leads to the hypothesis that differences in the frequency responses of the estimated acoustic channels are related to the specific acoustic propagation effects that happen in enclosed spaces.

RESUMEN.

La escena sonora se puede definir como la situación en que un determinado sonido ha sido registrado, típicamente por un micrófono. Esta situación o escena incluye tanto al conjunto de fuentes sonoras que han generado el sonido como al entorno físico en que se encuentran tanto estas como el micrófono. A priori, este entorno físico debe resultar relevante para la catalogación de la escena sonora, ya que configura el canal acústico entre fuentes y receptor. En esta comunicación se analiza precisamente la relevancia del canal acústico en la identificación de escenas sonoras o, dicho con otras palabras, si de las señales de audio captadas por el micrófono se puede extraer información sobre el canal que sea significativa para la clasificación automática de escenas sonoras. Para realizar el análisis se procesa la base de datos de audios correspondiente al reto DCASE 2019 (tarea 1A) de modo que para cada localización geográfica se calcula la media cepstral como estimación del canal acústico. Posteriormente se realiza un análisis exploratorio de la distribución de medias cepstrales obtenidas mediante métodos no paramétricos. Del análisis realizado se concluye que en determinados escenarios el contenido frecuencial de los audios captados presenta características que pueden ser atribuidas a rasgos previsibles en los correspondientes canales acústicos.

1. INTRODUCCIÓN

La clasificación automática de escenas acústicas tiene diversas aplicaciones potenciales [1], por lo cual ha sido una tarea recurrentemente incluida en los retos DCASE celebrados hasta ahora [2]. Las tasas de acierto logradas hasta la fecha superan el 80% para conjuntos de escenas de entre 10 y 15 tipos. Estas tasas de acierto son comparables a las que pueden lograr oyentes humanos, cuando no mejores [1]. Los detalles de los sistemas de clasificación publicados por los organizadores de DCASE [2] indican que la mayoría de estos sistemas resultan muy complejos, con típicamente más de un millón de parámetros ajustables en cada clasificador. Tales clasificadores actúan sobre conjuntos de parámetros acústicos obtenidos mediante algoritmos muy diversos aplicados a las señales acústicas, combinando frecuentemente análisis en los dominios del tiempo, de la frecuencia y del cepstrum. Estas combinaciones unidas a la complejidad de los clasificadores hacen muy difícil la identificación de los rasgos de las señales acústicas más relevantes para discriminar escenas acústicas.

El objetivo que persigue el trabajo referido en esta comunicación es, precisamente, la identificación de algunos rasgos de las señales acústicas que sean relevantes para la clasificación de escenas acústicas y, a la vez, permitan una explicación cualitativa razonable de tal relevancia. Para conseguir este objetivo se realiza un análisis que parte del modelo de escena acústica ilustrado en la Figura 1, en el que la escena se compone de un conjunto de fuentes acústicas inmersas en un entorno físico común, y cuyas emisiones son captadas por un micrófono ubicado en el mismo entorno. Si bien la señal acústica emitida por cada fuente atravesará un canal acústico específico para alcanzar al micrófono, el hecho de que todas las fuentes compartan entorno físico permite suponer que los canales no serán independientes entre sí, sino que habrá cierta relación entre ellos, compartirán algunos rasgos. Con el fin de mantener el problema abordable desde el punto de vista matemático, se supondrá que cada canal acústico de una escena tiene dos partes: una primera que es específica de cada fuente, y una segunda, común a todas las fuentes, que modela el efecto del entorno físico común y del micrófono. Si bien el modelo puede parecer simplista y poco realista, nótese que los procesos de auralización en el campo de la acústica virtual frecuentemente hacen uso de este modelo, separando los efectos de las salas y del entorno inmediato de los oídos para conseguir una señal estéreo a partir de una monocanal [3].

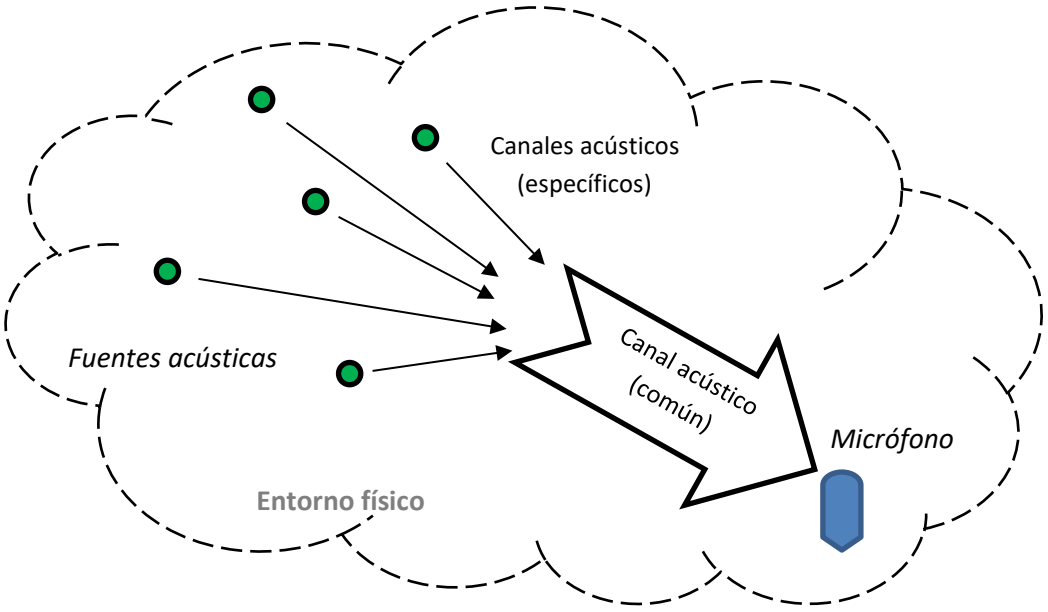

Figura 1 – Modelo de escena acústica.

Teniendo en cuenta el modelo anterior (Figura 1), la cuestión que se trata de responder se puede reformular como ¿es posible encontrar algún rasgo del canal acústico común que sea a la vez relevante en la identificación de la escena acústica y fácilmente interpretable? Para buscar una

respuesta a esta cuestión se analiza la base de datos de grabaciones correspondientes al reto DCASE 2019, concretamente a la tarea 1A (clasificación de escenas acústicas) [2]. La estimación de la componente común del canal acústico se realiza mediante el cálculo de la media cepstral de todas las grabaciones captadas con cada micrófono. Este proceso de estimación es típico del procesamiento del habla, cuando se intenta eliminar el efecto del canal para mejorar la identificación del contenido del discurso, o bien la identificación del hablante [4]. En este caso, no se trata de eliminar el efecto del canal, sino de analizar las estimaciones del canal. Este análisis se realiza mediante técnicas de aprendizaje automático no supervisado con el fin de evaluar si la distribución de los canales acústicos en el espacio cepstral está relacionada de algún modo con los tipos de escenas acústicas a que pertenecen. Esta evaluación, a su vez, se realiza en dos fases. En primer lugar, se proyectan los canales estimados sobre un espacio bidimensional con el fin de valorar visualmente su distribución. Esta proyección se realiza mediante la técnica t-SNE (*t-distributed stochastic neighbour embedding*) [5]. Posteriormente, se aplica la técnica de análisis de agrupamientos de las *k*-medias [6] para identificar los principales grupos de canales, comprobar si corresponden con tipos específicos de escenas acústicas, y analizar sus características.

2. MATERIALES

Las señales acústicas utilizadas para este análisis son las correspondientes al conjunto de grabaciones para el desarrollo de clasificadores (*development set*) de la tarea 1ª del reto DCASE 2019[1]. Todas las grabaciones de este conjunto forman parte de la base de datos TUT Urban Acoustic Scenes 2018 [7]. Concretamente, se seleccionaron 14400 grabaciones realizadas en 514 ubicaciones de 11 ciudades europeas. Cada ubicación se corresponde con uno de los diez tipos de escenas enumerados en la Tabla 1. La duración de cada grabación es de 10 s, y la duración total de todas las grabaciones realizadas en la misma ubicación oscila entre 5 y 6 minutos. Las grabaciones fueron realizadas con una grabadora Zoom F8 multitrack con una frecuencia de muestreo de 48 kHz y 24 bits/muestra. Se utilizó un micrófono biaural Soundman OKM II Klassik/studio A3, con respuesta aproximadamente plana entre 20 Hz y 20 kHz.

Tabla 1 – Escenas acústicas en la base de datos TUT Urban Acoustic Scenes 2018 [7]. El número de ubicaciones diferentes de cada escena se indica entre paréntesis. Las escenas se agrupan en tres tipos (interiores, exteriores y vehículos) con el fin de facilitar análisis posteriores.

#	Escena	Tipo de ubicación
1	Aeropuerto (40 ubicaciones)	Interior
2	Centro comercial cubierto (36)	
3	Parada de metro (57)	
4	Calle peatonal (46)	Exterior
5	Plaza (43)	
6	Calle con tráfico moderado (43)	
7	Parque público (41)	
8	Tranvía (70)	Interior de vehículo
9	Autobús (71)	
10	Metro (67)	

3. ALGORITMOS

El cálculo del cepstrum se propuso en primera instancia para la detección de ecos en señales de procedencia sísmica [8], pero se puede utilizar para estimar la respuesta de canales acústicos tal como se explica a continuación. De acuerdo con la teoría de sistemas lineales, el espectro de una señal $y(t)$ recibida a través de un cierto canal se puede expresar como:

$$S_y(f) = S_x(f) \cdot |H(f)|^2 \quad (1)$$

[1] Disponible en https://zenodo.org/record/2589280

donde $S_x(f)$ es la densidad espectral de potencia (DEP) de la señal $x(t)$ transmitida a través del canal, $S_y(t)$ es la DEP de $y(t)$, y $|H(f)|^2$ es la respuesta en frecuencia del canal, que puede estimarse como el cuadrado del módulo de la transformada de Fourier de la respuesta al impulso del propio canal.

Tomando logaritmos, el producto en el dominio del espectro se puede convertir en una suma:

$$\log\left(S_y(f)\right) = \log\left(S_x(f)\right) + \log(|H(f)|^2) \tag{2}$$

Esta conversión de producto a suma es válida en este dominio log-espectral, pero por linealidad se mantiene también si se aplica a (2) la transformada inversa de Fourier para obtener el cepstrum:

$$C_y(q) = C_x(q) \ + \mathcal{H}(q) \tag{3}$$

donde $C_y(q)$ y $C_x(q)$ son los cepstra de $y(t)$ y $x(t)$ respectivamente, $\mathcal{H}(q)$ contiene información del canal, puesto que es la transformada de Fourier inversa de $\log(|H(f)|^2)$, y q es la variable independiente del cepstrum. Considerando la ecuación anterior, si un conjunto suficientemente grande de J señales independientes entre sí $x_j(t)$ es transmitido a través del mismo canal, generando las correspondientes señales recibidas $y_j(t)$, entonces se puede suponer que la media cepstral de las señales recibidas es una estimación del cepstrum del canal:

$$\widehat{\mathcal{H}}(q) = \frac{1}{J}\sum_{j=1}^{J} C_{y_j}(q) = \frac{1}{J}\sum_{j=1}^{J} C_{x_j}(q) + \mathcal{H}(q) \approx \mathcal{H}(q) \tag{3}$$

Por supuesto, si J no es lo suficientemente grande, o bien las señales transmitidas no son independientes entre sí, entonces no se puede esperar que la suma $\frac{1}{J}\sum_{j=1}^{J} C_{x_j}(q)$ sea nula y $\widehat{\mathcal{H}}(q)$ inevitablemente contendrá información de la parte común de estas señales.

Interpretando la Figura 1, teniendo en cuenta esta técnica de estimación de la media cepstral, y suponiendo que las señales acústicas emitidas por cada una de las fuentes de una escena son independientes entre sí, la media cepstral de todas las grabaciones realizadas en una cierta localización se puede interpretar como una estimación de las características comunes de todos los canales acústicos asociados a esa localización. Coherentemente con este planteamiento, los dos canales disponibles de cada grabación (recuérdese que las grabaciones fueron biaurales) fueron procesados como sigue. En primer lugar, los 10 s de cada canal en cada grabación fueron normalizados a media nula y potencia unidad. Posteriormente, la señal de cada canal fue dividida en tramas de 200 ms con un solapamiento del 50% entre tramas consecutivas, y se calculó la transformada discreta de Fourier (DFT) de cada trama. El orden de la DFT fue dos veces el número de muestras de cada trama, con el fin de evitar solapamientos en el dominio del cepstrum [9]. Se calculó el cuadrado del módulo de cada DFT para obtener una estimación basada en el periodograma del espectro de cada trama [10] y, finalmente, se calculó el logaritmo de este módulo antes de calcular la DFT inversa para obtener el cepstrum de potencia [9]. Del cepstrum resultante se conservó la primera mitad de las muestras (recuérdese que al calcular la DFT se había duplicado el orden con respecto al número de muestras) y se calculó el promedio de todas las tramas de cada canal, de los dos canales de cada grabación, y de todas las grabaciones obtenidas en la misma localización. Ello dio lugar a 514 medias cepstrales, cada una de ellas correspondiente a una localización distinta.

4. RESULTADOS Y DISCUSIÓN

La Figura 2 muestra la distribución de las 514 localizaciones en el espacio definido por sus correspondientes medias cepstrales. Concretamente, los gráficos mostrados se corresponden con el resultado de calcular un mapa t-SNE [5] con las 514 medias cepstrales. De este modo se consigue una representación bidimensional de datos de un espacio de muchas más dimensiones. Tras una mera inspección visual se puede apreciar que las distribuciones de los datos son

diferentes para cada uno de los tres tipos de localizaciones definidos en la Tabla 1 (interiores, exteriores y vehículos). Concretamente, las localizaciones correspondientes a vehículos tienen la distribución más diferente. Entre las localizaciones interiores, se puede apreciar que las paradas de metro tienen una distribución mucho más dispersa que aeropuertos y centros comerciales. Estos, por otra parte, tienen distribuciones muy cercanas entre sí. De modo similar, parques y plazas son las localizaciones con distribuciones más dispersas en exteriores.

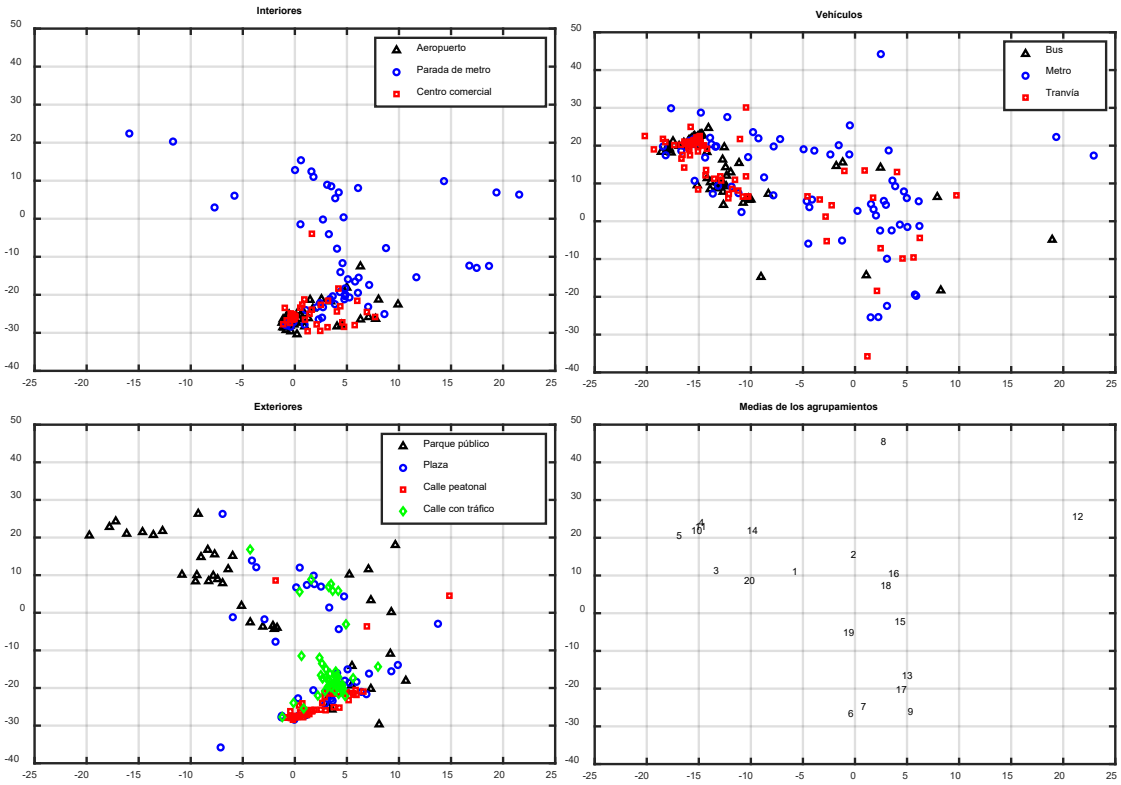

Figura 2 – Mapa t-SNE (*t-distributed stochastic neighbour embedding*) representando las medias ceptrales correspondientes a todas las localizaciones en las que se grabó. Por claridad se han separado los datos correspondientes a interiores (gráfica superior izquierda), exteriores (inferior izquierda), vehículos (superior derecha), y se ha añadido una cuarta gráfica representando los centros de los 20 principales agrupamientos de datos identificados (inferior derecha).

Con el fin de conocer mejor la organización de los datos se ejecutó sobre ellos una detección de agrupamientos basada en el algoritmo de las *k*-medias, con $k = 20$. El vector correspondiente a la media de cada uno de los 20 agrupamientos también se incorporó en la generación del mapa t-SNE y su posición en este mapa se indica en la gráfica inferior derecha de la Figura 2. Un aspecto relevante de este análisis es conocer si cada agrupamiento incluye principalmente datos correspondientes a escenas del mismo tipo o, alternativamente, escenas de tipos diversos aparecen mezcladas en los agrupamientos. Con este fin se identificaron las localizaciones incluidas en cada uno de los 20 agrupamientos. Del análisis posterior se descartaron los agrupamientos 8, 12, y 16 porque incluían cada uno menos de 10 localizaciones (menos de un 2% del total) y se les consideró poco relevantes. De entre los restantes, los agrupamientos 15, 18 y 19 (parte central del mapa) incluían una mezcla de escenas en la que ningún tipo aparecía como dominante. En esos casos, por tanto, se puede decir que la media cepstral no es útil para la identificación de la escena acústica. Los 14 agrupamientos restantes incluían más de 10 localizaciones cada uno y en todos se identificó un tipo de escena dominante. La composición de estos 14 agrupamientos se resume en la Tabla 2.

Los resultados de la Tabla 2 indican que en buena medida la estimación del canal acústico proporcionada por la media cepstral es útil para discriminar el tipo de escena acústica, si bien se

pueden identificar algunas limitaciones. Por ejemplo, 27 localizaciones correspondientes a calles peatonales tienen medias cepstrales similares a las de localizaciones interiores. Una posible explicación es que en las medias cepstrales se cuele información sobre las fuentes sonoras, que tanto en calles peatonales como en localizaciones interiores frecuentemente son voces. Por otra parte, hasta 24 localizaciones correspondientes a paradas de metro tienen medias cepstrales similares a localizaciones exteriores o vehículos. Se puede hipotetizar que esto se deba a la diversidad de tamaños y configuraciones posibles de las paradas de metro. Con todo, los resultados muestran que 312 de las 514 localizaciones analizadas (60,70%) se agrupan correctamente por tipo de escena.

Tabla 2 – Resumen de la composición de los agrupamientos en los que hay un tipo de escena dominante. La numeración de los agrupamientos es la misma que en la gráfica inferior derecha de la Figura 2. El tamaño de cada agrupamiento indica el número de ubicaciones incluidas en él. El número de localizaciones correspondientes al tipo dominante se indica entre paréntesis.

#	Tamaño	Tipo de escena dominante	Otras escenas
1	15	Exteriores (13): 11 parque, 2 plaza	
13	46	Exteriores (31): 2 parque, 7 plaza, 22 calle con tráfico	9 parada de metro
17	70	Exteriores (43): 5 parque, 12 plaza, 8 calle con tráfico, 18 calle peatonal	11 parada de metro
2	20	Vehículo (14): 6 bus, 5 metro, 3 tranvía	4 parada de metro
3	26	Vehículo (25): 12 bus, 4 metro, 9 tranvía	
4	11	Vehículo (11): 9 bus, 2 tranvía	
5	29	Vehículo (25): 10 bus, 3 metro, 12 tranvía	
10	29	Vehículo (27): 10 bus, 4 metro, 13 tranvía	
11	22	Vehículo (20): 11 bus, 2 metro, 7 tranvía	
14	21	Vehículo (18): 1 bus, 13 metro, 4 tranvía	
20	23	Vehículo (17): 7 bus, 3 metro, 7 tranvía	
6	32	Interiores (18) : 12 aeropuerto, 4 parada de metro, 2 centro comercial	9 calle peatonal
7	61	Interiores (38) :14 aeropuerto, 6 parada de metro, 18 centro comercial	16 calle peatonal
9	12	Interiores (12): 6 aeropuerto, 1 parada de metro, 5 centro comercial	

Con el fin de realizar un análisis cualitativo de las diferencias entre los 14 agrupamientos enumerados en la Tabla 2 se han realizado las siguientes operaciones. En primer lugar, para cada agrupamiento se obtuvo el valor de $\widehat{\mathcal{H}}(q)$ promedio de todas sus localizaciones, tal como se indicó anteriormente. Dado que estos valores son estimaciones de los cepstra de los correspondientes canales acústicos, a partir de ellos se generó una estimación de la respuesta en frecuencia media asociada a cada agrupamiento aplicando la DFT a cada $\widehat{\mathcal{H}}(q)$ y calculando la exponencial del resultado. Esto dio lugar a 14 respuestas en frecuencia estimadas, una correspondiente a cada agrupamiento. Con el fin de compararlas se calculó después la diferencia de cada una de ellas con respecto a la media de las 14. Estas diferencias aparecen representadas en la Figura 3.

En las gráficas de la Figura 3 se puede apreciar que a altas frecuencias, por encima de 4000 Hz, todas las estimaciones de canales acústicos son similares. Las principales diferencias entre ellos se ubican en dos bandas de frecuencia: por debajo de 70 Hz, y entre 100 Hz y 2000 Hz. Es bien conocido que cuando un canal acústico se encuentra rodeado de grandes superficies planas como las paredes, la respuesta de este canal se caracteriza por la aparición de modos que causan resonancias espectrales cuya densidad es función creciente de la frecuencia. Igualmente, se sabe que los modos sólo aparecen por encima de una cierta frecuencia, la frecuencia propia del modo más bajo, y que esta frecuencia umbral es función inversa de las dimensiones del espacio encerrado entre las superficies planas [11]. La práctica ausencia de modos en parques y plazas provoca que la respuesta en frecuencia asociada a las localizaciones del agrupamiento 1 sea mucho mayor a bajas frecuencias que en otras localizaciones. Para los agrupamientos en los que aparecen más calles, agrupamientos 13 y 17, la tendencia se mantiene, si bien al ser las

calles espacios más cerrados que parques y plazas la diferencia en esta banda de frecuencias con otros agrupamientos es menor.

La presencia de propagación modal también proporciona una explicación razonable al hecho de que la respuesta en frecuencia promedio en localizaciones interiores sea en general de mayor magnitud que la de exteriores en la banda entre 100 Hz y 2000 Hz. Asimismo, la similitud geométrica que puede existir entre algunas calles y paradas de metro puede explicar las similitudes entre las respuestas de los agrupamientos 7 y 17. Si bien uno corresponde mayoritariamente a localizaciones exteriores y el otro a interiores, la mayor confusión entre ellos se da precisamente en los casos de calles y paradas de metro.

En el caso de los vehículos, sus dimensiones típicamente más reducidas que las de edificios y otras estructuras urbanas provocan que la aparición de modos tenga lugar a frecuencias más altas, por lo que la respuesta en frecuencia a bajas y medias frecuencias, por debajo de unos 2000 Hz, es normalmente de menor magnitud que en los otros dos tipos de localizaciones. La variabilidad y los picos que aparecen en estos casos entre 50 Hz y 200 Hz es posible que sean debidos a ruidos de rodadura [12].

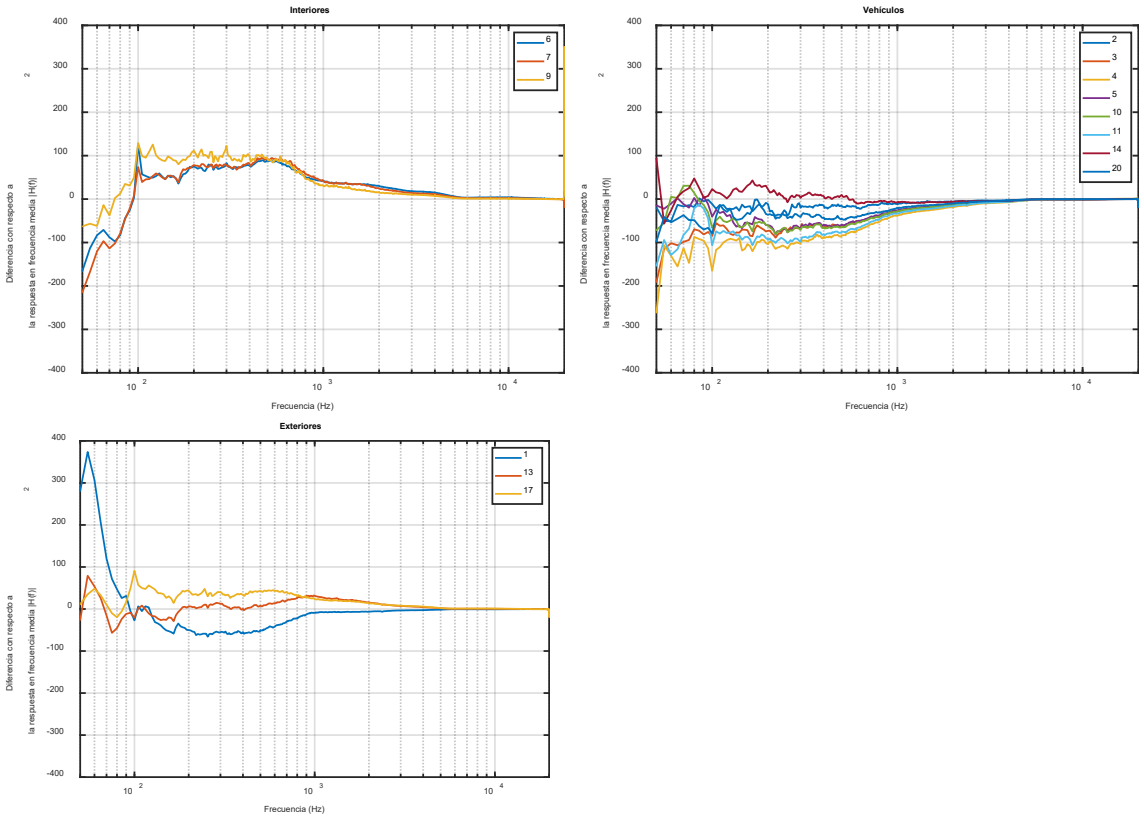

Figura 3 – Diferencia entre el espectro medio de todos los agrupamientos y el espectro central de cada uno para interiores (gráfica superior izquierda), vehículos (superior derecha), y exteriores (inferior).

5. CONCLUSIONES

El análisis de grabaciones mostrado en esta comunicación indica que la media cepstral de las grabaciones proporciona información útil para la clasificación de escenas acústicas, a la vez que interpretable desde un punto de vista acústico. Cuando el cálculo de la media cepstral se realiza sobre grabaciones suficientemente largas que incluyen un abanico amplio de fuentes acústicas se puede asumir que esta media cepstral es una estimación de la respuesta del canal acústico. De hecho, el estudio realizado sobre grabaciones realizadas en 514 localizaciones repartidas

entre 11 ciudades europeas muestra que algunos rasgos de estas estimaciones de canal se pueden interpretar de manera coherente con algunos principios fundamentales de la acústica arquitectónica. Se puede concluir, por lo tanto, que la media cepstral de una grabación sonora permite, aun con limitaciones, separar el efecto del canal acústico del de las fuentes y, consecuentemente, se puede utilizar como descriptor del entorno físico en que se han realizado tales grabaciones.

AGRADECIMIENTOS

Este trabajo ha sido apoyado por la Universidad Politécnica de Madrid a través del Programa Propio de I+D+I, en concreto la de contratos predoctorales.

REFERENCIAS

[1] Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbley, M.D. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Mag.*, 32(3), 2015, 16–34.

[2] DCASE Community. *DCASE Events*. 2019. [En línea] http://dcase.community/events

[3] Savioja, L.; Huopaniemi, J.; Lokki, T.; Väänänen, R. Creating interactive virtual acoustic environments. *Journal Audio Eng. Soc.*, 47(9), 1999, 675–705.

[4] Westphal, M. The use of cepstral means in conversational speech recognition. *Proc. of Eurospeech-1997*, 1997, 1143–1146.

[5] Van der Maaten, L.; G. Hinton, G. Visualizing data using t-SNE. *Journal Machine Learning Res.*, 9, 2008, 2579–2605.

[6] Nabney, I. *NETLAB: algorithms for pattern recognition*. Springer Science & Business Media, 2002.

[7] Mesaros, A.; Heittola, T.; Virtanen, T. A multidevice dataset for urban acoustic scene classification,. *Proc. of DCASE2018*, 2018, 9–13. [En línea] https://arxiv.org/abs/1807.09840

[8] Noll, A.M. Cepstrum pitch determination. *Journal Acoust. Soc. Amer.*, 41(2), 1967, 293–309.

[9] Childers, D.G.; Skinner, D.P.; Kemerait, R.C. The cepstrum: A guide to processing, *Proc. IEEE*, 65(10), 1977, 1428–1443.

[10] Oppenheim, A.V.; Schafer, R.W.; Buck, J.R. *Discrete-Time Signal Processing*. Prentice-Hall, 1989.

[11] Vigran, T.E. *Building Acoustics*. Taylor & Francis, 2008.

[12] Hills, E.; Mace, B. R.; Ferguson, N.S. Acoustic response variability in automotive vehicles. *Journal Sound, Vib.*, 321(1-2), 2009, 286–304.