Reconocimiento de Sonidos Ambientales usando Espectro Beat y Parámetros Estadísticos

Francisco Mondragón de la Luz, Sandra Espinosa Grajeda, Mariko Nakano Miyatake, Héctor Pérez Meana

ESIME Culhuacan, Instituto Politécnico Nacional.
Av. Santa Ana 1000, Col. San Francisco Culhuacan, México, D. F., México, 04430.
fmondragon1200@alumno.ipn.mx, saeg33@gmail.com, {mnakano,hmperezm}@ipn.mx

2014 Published by $\mathcal{DIFU}_{100}ci@$ http://nautilus.uaz.edu.mx/difu100cia

Resumen

El rápido desarrollo de las ciencias forense de las ciencias forenses en las últimas décadas ha dado como resultado la aparición de diversos métodos que permiten el análisis de grabaciones de audio que puede aportar importantes indicios en ciertas investigaciones judiciales, las cuales tradicionalmente se han concentrado en grabaciones de señales de voz. Debido a esto en años recientes ha recibido mucha atención el problema del reconocimiento de sonidos ambientales (ESR) ya que puede contribuir en el desarrollo de las investigaciones judiciales facilitando la recreación de los hechos. En este artículo se presenta un algoritmo para el reconocimiento de ruidos ambientales, empleando diferentes características tanto temporales como una modificación del "espectro beat" el cual permite obtener la periodicidad de la señal en diferentes bandas de la frecuencia. Una vez caracterizadas las señales de audio, los parámetros extraídos se insertan en el clasificador de Batchelor y Wilkins para tomar una decisión. El esquema desarrollado se evaluó utilizando diversos sonidos obteniéndose porcentajes de reconocimiento de superiores al 90 %.

Palabras clave: Espectro beat, reconocimiento de sonidos, Batchelor y Wilkins.

1. Introducción

La capacidad para grabar y almacenar señales de audio se ha incrementado de manera importante en los últimos años gracias al desarrollo de los teléfonos celulares y otros dispositivos de comunicación móviles, las cuales cuentan con grabadoras de audio y video de alta calidad. Esto incrementa la probabilidad de que las autoridades obtengan grabaciones de algún ilícito, ya sea realizada por las víctimas del delito o por terceras personas [1]. Estas grabaciones podrían revelar las conversaciones mantenidas con los delincuentes o dar un indicio del lugar en donde fue cometido el ilícito a partir de un análisis del ruido de fondo. En estas aplicaciones el desarrollo de esquemas que permitan identificar tanto a las personas involucradas como los sonidos ambientales presentes en los lugares donde se lleva a cabo un ilícito puede ser de gran importancia [2]. En lo que se refiere al reconocimiento de personas por medio del habla, se han desarrollado durante las últimas

décadas una amplia variedad de algoritmos; mientras que el interés en el desarrollo de algoritmos para el reconocimiento de sonidos ambientales (ESR), por sus siglas en inglés, se ha incrementado durante las últimas dos décadas debido a sus potenciales aplicaciones.

Los sistemas ESR, además de sus potenciales aplicaciones en la investigación criminalística han mostrado ser de utilidad en otros campos del quehacer humano tales como: la búsqueda eficiente de señales de audio en aplicaciones de etiquetado automático de archivos de audio, basado en descriptores, para la recuperación de archivos de audio [3, 4]. Aplicaciones de navegación robótica la puede ser mejorada con la incorporación de esquemas de ESR [5, 6]. El ESR puede ser también empleado en el monitoreo remoto de lugares cerrados, ya sea para asistir a ancianos que vivan solos en su hogar [7, 8] o en hogares inteligentes [9]. El ESR conjuntamente con sistemas de análisis de video e imágenes también encuentra aplicaciones en sistemas de vigilancia [10, 11]; así como para el reconocimiento y monitoreo de especies animales y de aves a través de sus sonidos distintivos [12].

Inicialmente los algoritmos de reconocimiento de sonidos ambientales fueron una extensión de los paradigmas de empleados en el reconocimiento de voz. Sin embargo estos algoritmos demostraron no ser lo suficiente efectivos en aplicaciones de ESR dadas las características de los sonidos ambientales y al hecho de que los esquemas de reconocimiento de voz a menudo explotan la estructura fonética de las señales de voz; así como los modelos que describen la producción de señales de voz a partir de una señal de excitación y de un modelo del tracto vocal. Mientras que los sonidos ambientales tales como los de un disparo, un rayo, una tormenta o el ruido en la vía pública no tienen, aparentemente, ninguna estructura susceptible de se modelada como en el caso de un fonema.

Con el fin de proponer soluciones a los problemas mencionados anteriormente varios esquemas han sido propuestos durante los últimos años, los cuales se pueden clasificar en esquemas empleando procesamiento basado en tramas. Esquemas empleando procesamiento basado en sub tramas y esquemas empleando procesamiento secuencial. En los esquemas basados en el procesamiento en tramas, las señales de audio para ser clasificadas son primeramente divididas en tramas, usando funciones ventana, generalmente del tipo Hanning o Hamming. Seguidamente se extraen las características relevantes de cada trama las cuales son usadas tanto durante el entrenamiento del sistema como en la etapa de operación del mismo. Finalmente una decisión de clasificación es hecha para cada trama. Cuando se emplea el procesamiento basado en sub tramas, cada trama es segmentada en pequeñas sub tramas, usualmente con un traslape del 50 %. Seguidamente las características extraídas de cada sub trama son concatenadas o promediadas para formar un vector característico el cual es empleado para entrenar un clasificador. Finalmente la salida del clasificador es empleada para tomar la decisión final. Otra posibilidad es entrenar el clasificador para cada una de las sub tramas, tomando entonces una decisión colectiva para cada trama, basada en los resultados obtenidos para cada una de las sub tramas. Finalmente, cuando se emplea el procesamiento secuencial las señales de audio son divididas en pequeñas unidades, típicamente de 20 a 30 ms de duración con un traslape del 50 %. Este método es único cuando se requiere capturar la correlación entre los diferentes segmentos, así como las variaciones de larga duración del sonido bajo análisis. En los tres tipos de esquemas diversos trabajos se han llevado a cabo relativos al análisis de aspectos no estacionarios de los sonidos ambientales, así como el desarrollo de diversos mecanismos para la caracterización de sonidos basados en las propiedades tanto estacionarias como no estacionarias de los sonidos [1, 2]. Estos esquemas intentan maximizar el contenido relacionado con las características temporales y espectrales de la señal bajo análisis.

Diversas características desarrolladas para aplicaciones relativas al reconocimiento de voz y/o música han sido tradicionalmente usadas en técnicas de ESR estacionarias. Estas características están a menudo basadas en propiedades psicoacústicas de los tales como sonoridad, pitch, timbre, etc. Una detallada descripción de características usadas en el procesamiento de audio se puede encontrar en [12]. Las técnicas no estacionarias de ESR emplean características obtenidas por medio de la transformada Wavelet, la representación dispersa y el espectrograma. Entre ellos los métodos basados en Wavelets dan resultados comparables a los métodos estacionarios, mientras que aquellos basados en representación dispersa y el espectrograma, en general, proporcionan en general mejores resultados [13]. En muchas ocasiones se combinan diversos tipos de características para mejorar la precisión de la clasificación.

2. Sistema Propuesto

La Fig. 1 muestra el diagrama a bloques del sistema propuesto el cual consiste de tres etapas, la etapa 1 que lleva a cabo el preprocesamiento de la señal de audio a fin de llevar a cabo la segmentación de las tramas de

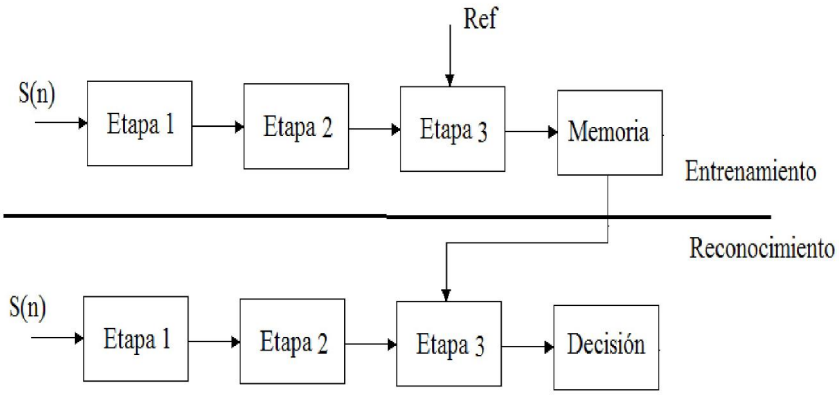

Figura 1. Sistema propuesto para el reconocimiento de sonidos ambientales

audio. Una vez procesada la señal de audio, esta se inserta en la etapa 2 la cual lleva a cabo la extracción de características. Seguidamente las características estimadas se insertan en la etapa de reconocimiento la cual se entrena de manera que se minimice algún criterio de la diferencia entre la respuesta real y la deseada. Una vez que el sistema se ha entenado los parámetros óptimos del clasificador se guardan en memoria para ser usados durante la operación de reconocimiento, en donde los parámetros del sistema se extraen de memoria y se insertan en el clasificador para llevar a cabo el reconocimiento.

2.1. Diseño del sistema

Cualquier sistema de reconocimiento de patrones depende fuertemente del funcionamiento de la etapa de extracción de características. A continuación se describen los esquemas de extracción de características empleadas en el sistema propuesto. Una de los parámetros que se proponen para caracterizar las señales de audio es el espectro beat, el cual es una medida de la periodicidad acústica de las diversas bandas de frecuencia de la señal bajo análisis, la cual se puede estimar a partir del espectrograma de la misma [13]. Con este fin, dado un archivo de audio $x(n)$, primero se calcula su espectro de frecuencia $X(f)$, empleando una ventana de Hamming con 50 % de traslape. A continuación, se obtiene la magnitud espectrograma $V(f)$, tomando el valor absoluto de los elementos de $X(f)$, después de desechar la parte simétrica. Seguidamente se calcula el promedio de la autocorrelación en una cierta banda de frecuencias del espectrograma, $V^2(f)$, a fin de obtener la matriz B. Finalmente la auto correlación acústica promedio b de $x(n)$ se normaliza con respecto al valor $b(0)$. Este proceso se describe por medio de las ecuaciones siguientes:

$$B(i,j) = \frac{1}{m-j+1} \sum_{k=1}^{m-j+1} V^2(i,j)V^2(i,k+j-1) \quad (1)$$

$$b(j) = \frac{1}{n} \sum_{i=1}^{n} B(i,j) \; entonces \; b(j) = \frac{b(j)}{b(1)} \quad (2)$$

Donde $i = 1, 2, ..., n$ denota la $i\ esima$ componente de frecuencias en escala mel, $n = N/2 + 1, ..., N$, el número total de componentes de frecuencias y m denota el número total de tramas en la señal de audio. Otro parámetro muy que se emplea frecuentemente para llevar acabo la caracterización de las señales de audio es el ancho de banda de la señal de entrada, el cual está dado por:

$$\sqrt{\frac{\int_0^{w_0}(w-w_0)^2|F(w)|^2dw}{\int_0^{w_0}|F(w)|^2dw}} \quad (3)$$

El ancho de banda representa una característica de frecuencia que ha demostrado ser muy eficiente en muchos sistemas empleados para el reconocimiento de audio [15, 19, 20, 21].

El centroide, el cual es una medida de brillo espectral también ha mostrado ser de utilidad para caracterizar sonidos, se puede estimar como:

$$C = \frac{\sum_1^N fM[f]}{\sum_1^N M[f]} \quad (4)$$

La duración espectral (SF), la cual representa el valor promedio de la variación de espectro entre dos tramas adyacentes de una señal de audio, es otro parámetro importante para la caracterización de sonidos la cual está dada por:

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta)]^2 \quad (5)$$

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(n) w(nL-m) e^{\frac{-j2\pi km}{L}} \right| \quad (6)$$

Donde $x(n)$ es la señal de audio de entrada discreta, $w(m)$ de la ventana función; L es la longitud de la ventana, k denota la $k\ esima$ componente de frecuencia, δ es un valor muy pequeño para evitar el desbordamiento de cálculo y n es el número total de muestras en una trama de audio [18]. El número de cruces por cero se define como el número de cambios de signo que experimenta la señal de audio en cada trama. Esto proporciona de manera sencilla una medida de la frecuencia fundamental de la señal. El número de cruces por cero esta dado por:

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1) - sgn[x(m)]| \quad (7)$$

Donde sgn es una función de signo, $x(m)$ es la señal de audio y $m = 1, ..., N$. El ZCR es un buen discriminador entre las señales de voz y las señales de ruido ambiente. Debido a esto, muchos sistemas [14, 15, 16, 17, 18] han utilizado el ZCR para la clasificación de señales de audio.

2.2. Etapa de Clasificación

El clasificador empleado para llevar a cabo el reconocimiento de las señales de audio es el esquema de Batchelor & Wilkins (B & W), el cual a diferencia del esquema K means, es un método de agrupamiento en el cuál el número de clases es desconocido de antemano. Este es un clasificador de muy eficiente y de baja complejidad, aunque su comportamiento esta sesgado por el orden de la presentación de los patrones. El clasificador B & W se muestra en la Tabla 1.

Tabla 1. Algoritmo de Batchelor & Wilkins

Algoritmo
Primer agrupamiento: Patrón escogido al azar
Segundo agrupamiento: Patrón más alejado del primer agrupamiento
Mientras se creen nuevos agrupamientos obtener el patrón cuya distancia con los agrupamientos existentes sea máxima
Si la distancia obtenida es mayor que una fracción de la distancia media entre los agrupamientos, crear un nuevo agrupamiento con el patrón seleccionado
Asignar cada patrón a su agrupamiento más cercano

3. Evaluación Experimental

Con el fin de evaluar el desempeño del sistema propuesto con distintos tipos de sonidos, se creó una base de datos consistente una mezcla de señales de voz con diversas clases de ruidos ambientales tales como bosque, selva, gato, perro, caballo, disparo, bebé, así como niños jugando, con lo que se generó una base de datos de 8 clases con tres archivos de audio ambiental por cada clase. Seguidamente se extrajeron las partes no vocalizadas de la señal bajo análisis, a fin de recuperar los sonidos ambientales para su posterior caracterización y reconocimiento mediante el algoritmo de B & W.

Tabla 2. Resultados experementales empleando los parametros temporales y el espectro beat

Sonido	PT %	BSC %	b1 %	b2 %	b3 %	b3 %
Bosque	0	100	100	100	100	100
Selva	100	100	100	100	100	100
Gato	0	100	33.3	66.6	66.6	33.3
Perro	0	100	66.6	100	100	100
Caballo	100	100	100	100	100	100
Disparo	66.6	100	100	100	100	100
Bebe	33.3	100	66.6	33.3	66.6	33.3
Niños	33.3	100	100	100	33.3	100

Inicialmente, los sonidos ambientales se caracterizaron por medio del centroide, ancho de banda, duración espectral, taza de cruces por cero, etc., tomando tramas de 30 ms con un traslape de 50 %, cuando esto es necesario. Estas características se concatenaron para llevar a cabo el proceso de reconocimiento. Seguidamente se procedió a evaluar el desempeño del sistema propuesto, cuando los sonidos ambientales se caracterizaron empleando el “espectro beat”. Con esta finalidad, se calculó el espectro beat de los sonidos bajo análisis para 8 bandas de diferentes frecuencias en escala mel, tomando tramas de 10 ms con el cual se construyó el beat espectrograma para poder visualizar en cuál de las bandas del espectro beat existe una periodicidad característica para cada uno de los sonidos. De estos resultados se observa que, empleando las primeras cuatro bandas del espectro beat se puede caracterizar los audios ambientales debido a que estas muestran una periodicidad más marcada para cada una de estas bandas. Seguidamente en la figura 2 se pueden observar las gráficas del espectro beat de las cuatro bandas seleccionadas para las 8 clases donde se observa una clara caracterización para los sonidos ambientales de

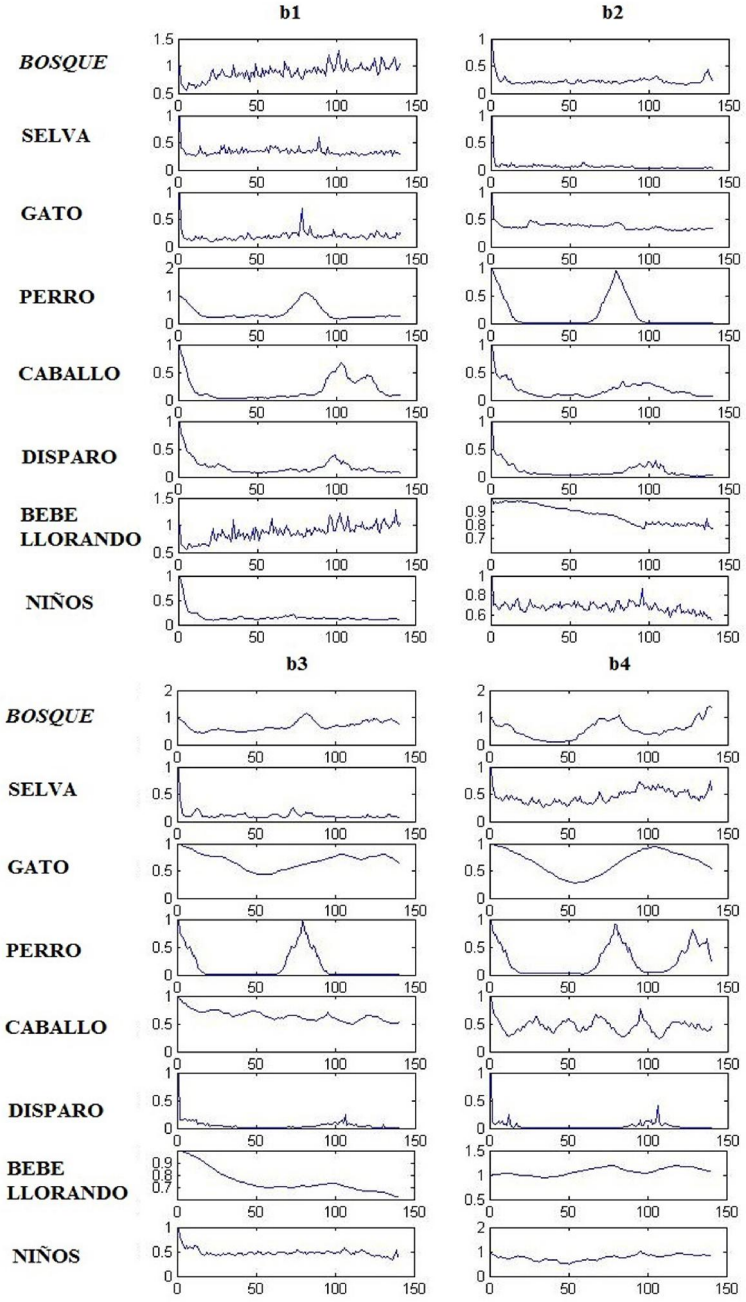

Figura 2. Espectro de PL obtenido de un depósito sobre substrato de silicio (muestra G3)

cada clase. Para evaluar el desempeño al caracterizar sonidos ambientales del espectro beat en las bandas seleccionadas así como al concatenar los vectores característicos de las cuatro bandas se entrenó con estos vectores el clasificador B & W, obteniendo los resultados que se pueden ver en la tabla 2, junto con los resultados obtenidos empleando una concatenación de las otras características descritas en la sección 2.

4. Conclusiones

De los resultados obtenidos empleando el sistema propuesto cuando la extracción de características se realiza empleando el espectro beat y las características temporales usualmente usadas tales como centroide, ancho de banda, duración espectral, taza de cruces por cero, se puede observar que el espectro beat muestra ser un método más eficiente para llevar a cabo la caracterización de diversos sonidos ambientales, especialmente cuando se usa un vector característico concatenando los vectores del espectro beat de las cuatro bandas seleccionadas. Esto sugiere que el espectro beat es una característica muy útil para el reconocimiento de sonidos ambientales.

Referencias

[1] A. Neustein and H. Partil. "Forensic Speaker Recognition Law Enforcement and Counter Terrorism". 2013. Springer, 2012

[2] S. Ikram and H. Malik. "Digital Audio Forensics using Background Noise". *In: Proc. of International Conference on Multimedia and Expo,* pp. 106–110, IEEE Press (2010).

[3] 3T. Virtanen and M. Helén. "Probabilistic model based similarity measures for audio query by example". *In: Workshop of Applications of Signal Processing to Audio and Acoustics*, pp. 82–85, IEEE Press (2007).

[4] S.Duan, J. Zhang, P. Roe and M. Towsey, "A survey of tagging techniques for music, speech and environmental sound". Artificial Intelligence Review, pp. 1–25 (2012).

[5] S. Chu, S. Narayanan, J. Kuo and M. Mataric M. J., "Where am I? Scene recognition for mobile robots using audio features". *In: Proc. of Multimedia and Expo*, pp. 885–888, IEEE Press (2006).

[6] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. Okuno. "Environmental sound recognition for robot audition using Matching Pursuit". *In Modern Approaches in Applied Intelligence*, pp. 1–10 Springer (2011).

[7] J. Chen, H. Kam, J. Zhang, N. Liu and L. Shue. "Bathroom activity monitoring based on sound". *In Pervasive Computing*, pp. 47–61, Springer(2005).

[8] M. Vacher, F. Portet, A. Fleury and N. Noury "Challenges in the processing of audio channels for ambient assisted living". *In Proc. of e Health Networking Applications and Services (Healthcom)*, pp. 330–337, IEEE Press (2010).

[9] Wang J. J. C., H. Lee, J. Wang and C. Lin. "Robust environmental sound recognition for home automation". IEEE Trans. Autom. Sci. Eng. 5, pp. 25–31, (2008).

[10] M. Cristani, M. Bicego and V. Murino. "Audio visual event recognition in surveillance video sequences". IEEE Trans. Multimedia, 9(2), pp. 257–267, (2007).

[11] R. Sitte and L. Willets. "Non speech environmental sound identification for surveillance using self organizing maps". *In Proc. Int. Conf. on: Signal Processing, Pattern Recognition, and Applications*,pp. 281–286, ACTA Press (2007).

[12] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder. "Features for content based audio retrieval". *Advances in computers*, 78, pp. 71–150 (2010).

[13] J. Foote and S. S. Uchihashi. "The beat spectrum: A new approach to rhythm analysis". *In: Proc. of Int. Conf. Multimedia and Expo*, pp. 881–884, IEEE Press (2001).

[14] J. Saunders. "Real time Discrimination of Broadcast Speech/Music". *Proc. of ICASSP96, Vol. II*,pp. 993–996, Atlanta, May, 1996.

[15] E. Scheirer and M. Slaney. "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator". *In: Proc. of ICASSP 97*, pp 1331–1334 IEEE Press.1997.

[16] T. Zhang, C. C. J. Kuo. "Heuristic Approach for Generic Audio Data Segmentation and Annotation". *Proc. of ACM Multimedia 99*, pp. 67–76, 1999.

[17] S. Srinivasan, D. Petkovic and D. Ponceleon. "Towards robust features for classifying audio in the CueVideo System". *In: Proc. int. conf. on Multimedia*, pp.393–400.ACM (1999).

[18] L. Lu,H. Jiang and H. Zhang. "A Robust Audio Classification and Segmentation Method". *In: Proc, of int. conf. on Multimedia*, pp. 203–211, ACM (2001).

[19] S. Li. "Content based classification and retrieval of audio using the nearest feature line method". *IEEE Transactions on Speech and Audio Processing*, 8 (5) 619–625 (2000).

[20] E. Wold, T. Blum, D. Keislar and J. Wheaton. "Content based classification, search and retrieval of audio". *IEEE Multimedia Magazine*, 3(3): 27–36, (1996).

[21] Z. Liu, J. Huang, Y. Wang and T. Chen. "Audio feature extraction and analysis for scene classification". *In: Workshop on Multimedia Signal Processing*, IEEE (1997).