Información Tecnológica
Vol. 32(2), 61-78 (2021)
http://dx.doi.org/10.4067/S0718-07642021000200061

Clasificación de sonidos ambientales usando la transformada wavelet continua y redes neuronales convolucionales

Francisco J. Mondragón, Héctor M. Pérez-Meana*, Gustavo Calderón, y Jonathan Jiménez
Escuela Superior de Ingeniería Mecánica y Eléctrica Culhuacan, SEPI, Avenida Santa Ana 1000, San Francisco Culhuacan, Culhuacan CTM V, Coyoacán, 04440 CDMX, México. (Correo-e: fmondragon1200@alumnio.ipn.mx; hmperezm@ipn.mx; jjimeneza@alumno.ipn.mx; gus_auza@hotmail.com)

*Autor a quien debe ser dirigida la correspondencia.

Recibido Sep. 1, 2020; Aceptado Oct. 27, 2020; Versión final Dic. 23, 2020, Publicado Abr. 2021

Resumen

Este artículo propone un esquema en el cual inicialmente se obtiene una representación tiempo-frecuencia usando la transformada wavelet continua (CWT), la cual tiene una resolución logarítmica en el plano de la frecuencia similar a la del sistema auditivo humano. El desarrollo de este tipo de sistemas para la clasificación de sonidos ambientales ha sido un tópico de amplia investigación debido a sus aplicaciones en diversos campos de la ciencia e ingeniería. Al igual que otros esquemas de clasificación, estos se basan en la extracción de parámetros característicos, los cuales se insertan en la etapa de clasificación. La CWT se inserta en una red neuronal profunda, para llevar a cabo el proceso de clasificación. Los resultados obtenidos, usando bases de datos de sonidos ambientales tales como, ESC-50, TUT Acoustic Scene y SONAM-50, demuestran que el esquema propuesto proporciona un funcionamiento superior al de otros esquemas previamente propuestos.

Palabras clave: reconocimiento sonidos ambientales; red neuronal profunda; transformada wavelet continua; espectrograma

Environmental sound recognition using continuous wavelet transform and convolutional neural networks

Abstract

This paper proposes a scheme in which a time-frequency representation is first obtained using the continuous wavelet transform (CWT), which has a logarithmic resolution in the frequency domain, like that of the human ear. The development of these environmental sound classification systems is a topic of extensive research due to its application in several fields of science and engineering. Like other classification schemes, they are based on the extraction of specific parameters that are inserted in the classification stage. The CWT is then inserted into a deep learning neural network to carry out the classification task. The evaluation results obtained using several databases such as ESC-50, TUT Acoustic Scene, and SONAM-50 show that the proposed scheme provides a classification performance that is better than that provided by other previously proposed schemes.

Keywords: environmental sound recognition; deep neural network; continuous wavelet transform; spectrogram

INTRODUCCIÓN

La clasificación de sonidos ambientales (ESC), por sus siglas en Ingles *"Environmental Sound Classification",* es un área de investigación en reciente crecimiento, debido al crucial rol que tienen los sonidos en nuestra interacción con el entorno. Por lo tanto, es fundamental para el éxito de la inteligencia artificial que los robots o las computadoras puedan comprender los sonidos, en forma similar a como lo hacen los humanos. El Desarrollo de tecnologías que ayuden en tareas como, sistemas para el monitoreo de cuartos inteligentes, mejorar la navegación autónoma (Chu et al., 2009), determinación de especies de aves y mamíferos basado en los sonidos que producen (Abber, 2020; Potamitis, 2014; Xie y Zhu, 2019), clasificación de sonidos en ayudas auditivas (Alexandre et al., 2007), indexación y recuperación de contenido multimedia (Tong et al., 2014), entre otras ha sido la motivación para el desarrollo de sistemas de ESC.

Algunos enfoques utilizados en la tarea de ESC se basan en características como la Transformada de Fourier Discreta (DFT), coeficientes cepstrales de las frecuencias Mel (MFCC) (Chu, et al., 2009; Salamon y Bello, 2014), coeficientes cepstrales de frecuencias gammaton (GFCC), características estadísticas y combinación de estas. La representación de los sonidos basada en el sistema de audición humana tiene un gran interés en el desarrollo de características para realizar ESC. Los modelos auditivos se basan en algoritmos matemáticos que intentan imitar el procesamiento de audición humana diseñados a partir de experimentos psicofísicos y fisiológicos.

El análisis wavelet (Martínez et al., 2018; Jiménez et al., 2018) se está convirtiendo en una herramienta matemática habitual en el estudio de señales no estacionarias como lo son la mayoría de los sonidos. Esta herramienta realiza el análisis de una señal usando versiones escaladas y trasladadas de una función base llamada función wavelet madre $\psi(t)$. Hay dos diferentes tipos de transformada wavelet: la Transformada Wavelet Discreta (DWT) y la Transformada Wavelet Continua (CWT). La principal diferencia entre ambas transformadas es la forma en la cual el parámetro de escalamiento es discretizado. La CWT discretiza más fielmente que la DWT. La diferencia es que mientras que en la CWT normalmente se determina alguna base que es una potencia fraccionaria de dos, es decir $2^{j/v}$ con j=1, 2, 3, …, n donde el parámetro v es conocido como el número de voces por octava, ya que para poder incrementar la escala en una octava se necesitan v escalas intermedias, mientras que en la DWT el parámetro de escalamiento siempre se discretiza con potencias enteras de dos, esto es 2^j con j=1 ,2, 3, …, n por lo que el número de voces por octava es siempre 1, por lo tanto la CWT dependiendo del valor de v nos proporciona una mayor resolución en frecuencia de la señal en análisis. Sin embargo, esto también aumenta la cantidad de cálculo requerido. La CWT se puede considerar como un banco de filtros que tienen subbandas de frecuencia espaciadas de manera logarítmica similar al sistema auditivo humano, por lo cual provee una representación tiempo frecuencia con una resolución logarítmica en el eje de la frecuencia, lo que proporciona una mejor representación para la inspección visual, como se muestra en la figura 1.

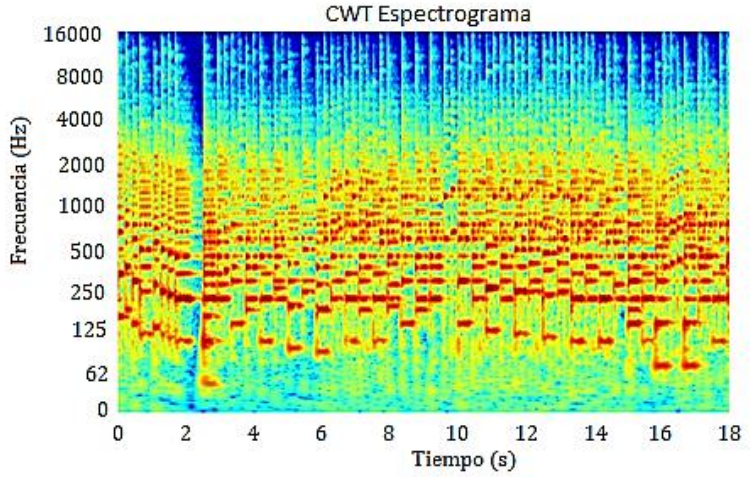


Fig. 1: CWT Espectrograma de una grabación de piano

Recientemente los sistemas de clasificación de sonidos basados ya sea en la DWT (Qian et al., 2017, Waldekar y Saha, 2020), la CWT (Li et al., 2018; 2019) y las redes neuronales de aprendizaje profundo (DNN) han comprobado su utilidad en las tareas de ESC (Piczak, 2015a), donde han mostrado ser útiles para capturar la modulación de energía de los espectrogramas obtenidos de las señales de audio (Salamon y Bello, 2017). En este trabajo se explora el poder de la CWT (Santamaría 2012) para generar espectrogramas representativos para clasificar sonidos ambientales usando redes DNN pre entrenadas. Los experimentos

realizados usando las bases de datos ESC-50, la TUT Acoustic scene (Mesaros et al., 2017) y SONAM-50 muestran que, empleando el espectrograma obtenido con la CWT junto con DNN, se obtiene un desempeño competitivo con con otros sistemas reportados en la literatura al clasificar sonidos y escenas ambientales. El artículo está organizado de la siguiente manera. La teoría relacionada al análisis de la transformada wavelet continua y el método propuesto para calcular la CWT de manera eficiente se presentan en la sección 2. Un resumen de la DNN y la técnica de transferencia de aprendizaje para adaptar las DNN a una nueva tarea de clasificación se exponen en la sección 3. El método propuesto para realizar la tarea de ESC es detallado en la sección 4. La base de datos y los experimentos realizados son descritos en la sección 5. Finalizamos el artículo con las conclusiones en la sección 6.

OTROS ANTECEDENTES

Hay una serie de antecedentes adicionales que es necesario detallar para documentar en mejor forma este trabajo: i) análisis de la transformada wavelet continua; ii) algoritmo propuesto para el cálculo eficiente de la CWT; y iii) redes neuronales convolucionales.

Análisis de la transformada wavelet continua

La transformada wavelet descompone una señal unidimensional dada en una combinación de funciones (wavelets), las cuales se obtienen mediante la dilatación y traslación de una función llamada wavelet madre. Esto permite obtener una representación bidimensional, tiempo-frecuencia, a partir de una señal unidimensional en el tiempo. Las funciones wavelet madre $\psi(t)$, deben satisfacer la llamada condición de admisibilidad demostrada por Sadowsky (Sadowsky, 1996; Najmi y Sadowski, 1996), que es necesaria para que la transformada wavelet inversa exista, esto es:

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega \; < \; \infty, \tag{1}$$

donde ω es la frecuencia angular y $\Psi(\omega)$ es la transformada de Fourier de la función wavelet madre. Por consiguiente, la condición de admisibilidad implica que la función wavelet madre no tiene componentes de corriente directa (CD) es decir

$$\Psi(0) \; = \; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi(t) = 0. \tag{2}$$

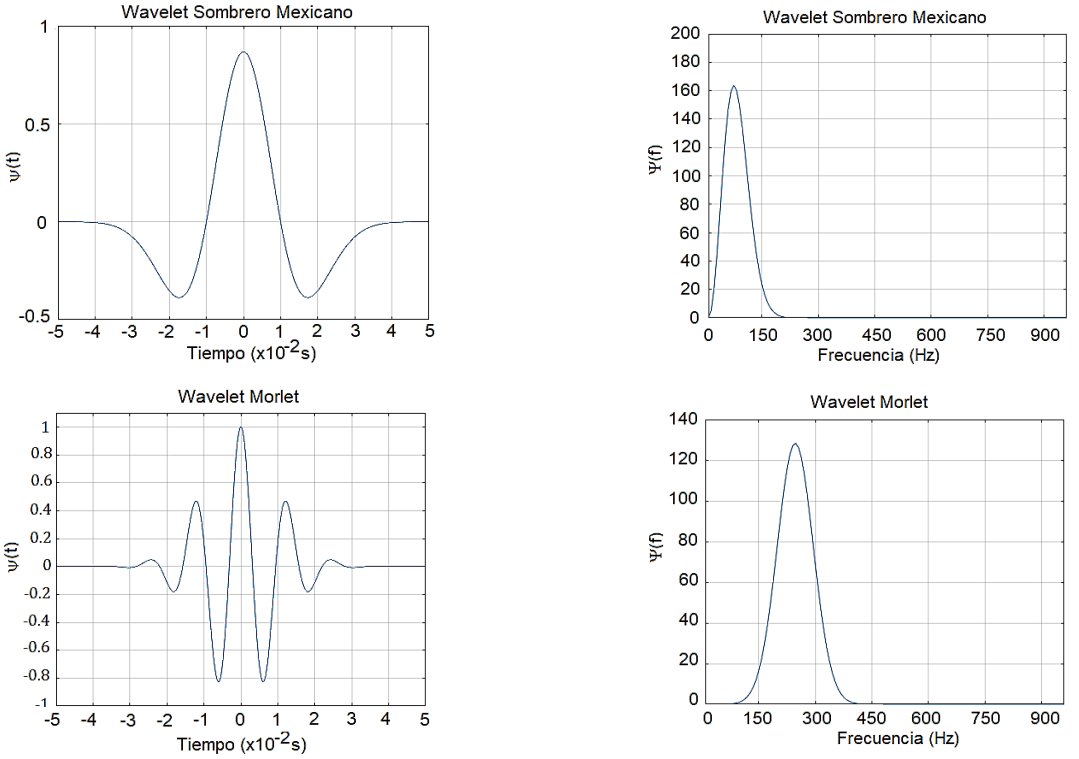


Fig. 2: Funciones wavelet madre, las gráficas de la izquierda están en el dominio del tiempo y las de la derecha son su correspondencia en el dominio de la frecuencia

Así, la condición de admisibilidad implica que la integral de la función wavelet madre es cero. Para que esto suceda, la función debe contener fluctuaciones, eso significa que debe tener suficiente área negativa para anular la positiva. Ciertamente, esto es lo que significa que no hay componentes de DC. Existen varias funciones madre que cumplen la condición de admisibilidad, utilizadas en el análisis de señales entre las cuales se encuentran; la wavelet de Morlet, la wavelet de sombrero mexicano entre otras, las cuales se ilustran en la figura 2. Asumiendo que la función wavelet madre $\psi(t)$ cumple la condición de admisibilidad, la CWT de una señal s(t) se obtiene al modificar la función wavelet madre a partir de dos factores, la dilatación a>0 y la traslación b, esto es:

$$W(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} s(t)\psi^*\left(\frac{t-b}{a}\right) dt, \quad (3)$$

donde * es la conjugación compleja. Por lo tanto, la CWT es la correlación de la señal s(t) con $\psi_{ab}(t)$. En el dominio del tiempo para un factor a>1 la wavelet se expande; mientras que para a<1 se contrae, en frecuencia este factor tiene un comportamiento inverso, es decir un reescalamiento por un factor "a" en el dominio del tiempo conlleva a reescalar por 1/a en el dominio de la frecuencia. Este comportamiento para la wavelet "sombrero mexicano" se puede ver en la figura 3.

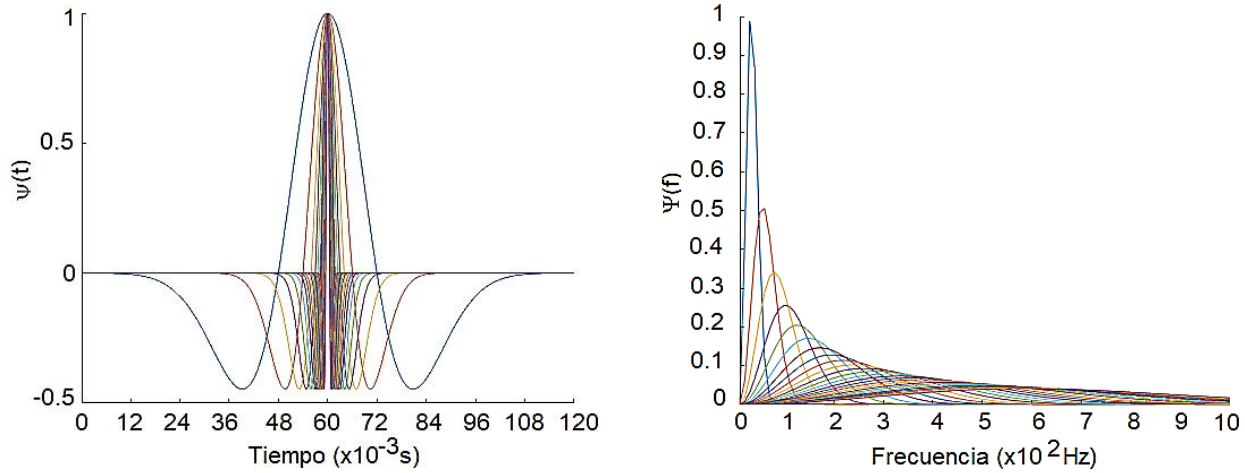


Fig. 3: Funciones wavelet sombrero mexicano en tiempo (izquierda) y frecuencia (derecha) con 25 diferentes valores del factor de dilatación a=1/m; m=0, 1, 2, … ,25.

Selección de escalas

Una contribución fundamental del trabajo de Morlet fue demostrar que el muestreo natural del factor de dilatación "a" es logarítmico, es decir potencias de base 2, (Najmi y Sadowsky, 1997). Debido a que la escala musical también está basada en potencias de base dos y el rango de escalas es dividido en términos de octavas y voces, se adopta la terminología musical para describir el comportamiento del factor de dilatación. Además, la escala musical se definió por la forma en que el sistema auditivo humano percibe las señales acústicas. El número de octavas circunscribe el ancho de banda de frecuencias que se analizan, mientras que el número de voces determina las muestras entre cada octava. Con respecto a la selección de "a" para calcular la CWT, siendo el número de octavas N_0 y el número de voces por octava N_v. La progresión de los índices de octava y voces son respectivamente:

$$i_o = 0,1,2,\ldots,N_o - 1, \quad (4)$$

$$i_v = 0,1,2,\ldots,N_v - 1. \quad (5)$$

Entonces las escalas para una octava i_0 y voz i_v específicamente, viene dada por:

$$a = 2^{(i_o + i_v/N_v)}. \quad (6)$$

Los progresos de las escalas están referenciados a un índice, de la siguiente manera.

$$i_a = 0,1,2,\ldots,N_o N_v. \quad (7)$$

La selección de octavas y escalas depende de la frecuencia de muestreo o frecuencia de Nyquist del audio a analizar, la cual es una opción confiable porque las señales generalmente se muestrean de tal manera que la frecuencia de Nyquist es al menos el doble de frecuencia más alta contenida en la señal. La relación del factor "a" y la frecuencia se ilustran en la figura 4, donde se puede observar la relación logarítmica de las frecuencias en análisis y el parámetro de dilatación "a".

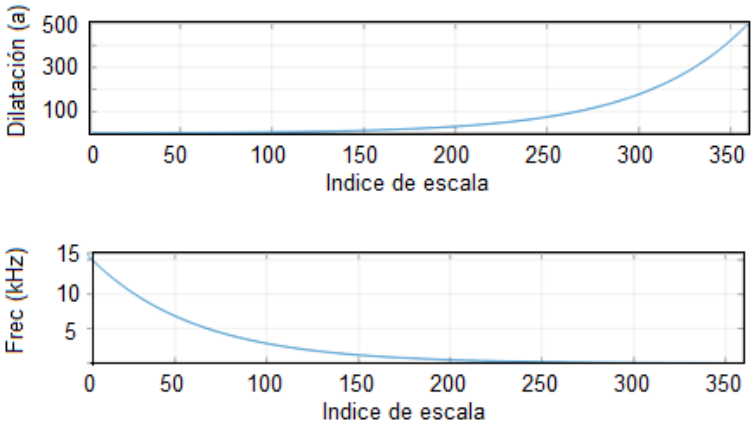


Fig. 4: Relación del factor de escalamiento y la frecuencia para señales de audio con frecuencia de muestro de 32 KHz.

Selección de función wavelet madre

Para llevar a cabo la selección de la función wavelet madre, se deben tomar en consideración las siguientes características: a) Funciones wavelets ortogonales o no ortogonales. En el análisis usando wavelets ortogonales, el número de variaciones en cada escala es proporcional a la dilatación de la función wavelet en esa escala. Esto es útil cuando se busca una representación más compacta de la señal. Por su parte, un análisis con wavelets no ortogonales es altamente redundante a grandes escalas, dando así un espectro altamente correlacionado. Esto es útil para el análisis donde se esperan variaciones continuas y suaves en la amplitud de la función wavelet (Torrence y Compo, 1998). b) Funciones wavelets complejas o reales. Una función wavelet compleja proporcionará información sobre la amplitud y fase de la señal bajo análisis, por lo cual es mejor para capturar el comportamiento oscilatorio, mientras que una función wavelet real devuelve solo una componente por lo tanto puede usarse para aislar picos y discontinuidades (Torrence y Compo, 1998). c) Anchura de la función. La resolución de una función wavelet está determinada por el equilibrio entre el ancho en el espacio de tiempo, contra el ancho en el espacio de Fourier. Una función estrecha (en el tiempo) tendrá una buena resolución de tiempo, pero una resolución de baja frecuencia. Mientras que una función amplia tendrá una resolución pobre en tiempo, pero una buena resolución de frecuencia (Torrence y Compo, 1998). d) Forma. La función wavelet tiene que reflejar las características presentes en las señales. Así, para series de tiempo con saltos o pasos agudos se debería elegir una función tipo vagón como la de Harr; mientras que para señales de tiempo que varían suavemente sería recomendable una función suavizada similares a una onda sinusoidal amortiguada (Torrence y Compo, 1998).

Algoritmo utilizado para el cálculo eficiente de la CWT

La CWT tiene un alto costo computacional comparado con otros procesos para obtener una representación tiempo-frecuencia de la señal, como la transformada de Fourier de corto tiempo (STFT), por lo cual en los últimos años se han propuesto algoritmos para aproximar el cálculo de la CWT de maneras más eficiente. La CWT usa el hecho de que la parte dominante de la respuesta en frecuencia de la r-ésima componente se encuentra centrada alrededor de la frecuencia central de la misma, lo cual es común en muchos tipos de funciones wavelet. Así la CWT de una señal de entrada x(t) está dada por:

$$W_r(t) = \int_{-\infty}^{\infty} x(\tau)\psi_r^*(t - \tau)d\tau. \quad (8)$$

La ecuación (8) se puede calcular usando la Transformada Inversa de Fourier,

$$W_r(t) = \int_{-\infty}^{\infty} X(2\pi f)\Psi_r^*(2\pi f)e^{j2\pi ft}\, df, \quad (9)$$

Donde x(2πf) es la transformada de Fourier de la señal y $\Psi(2\pi f)$ la transformada de Fourier de la función wavelet. Para simplificar el cálculo de $W_r(t)$, la ecuación (9) se puede calcular usando la transformada discreta de Fourier. Así, tomando en cuenta que las señales involucradas son limitadas en banda se obtiene:

$$W_r(n) = \sum_{k=0}^{T-1} X(k)\Psi_r^*(k)e^{2\pi nk/T}. \quad (10)$$

Suponga que $\psi_r(t)$ tiene valores significativos únicamente en el intervalo B≤k≤B+D y T-B-D≤T-D, como se muestra en la figura 5(b). Así la ecuación (10) se puede calcular como:

$$W_r(n) = \sum_{k=B}^{B+D-1} X(k)\,\Psi_r^*(k)\,e^{2\pi nk/T} + \sum_{k=T-(B+D)}^{T-B-1} X(k)\Psi_r^*(k)e^{2\pi nk/T}. \quad (11)$$

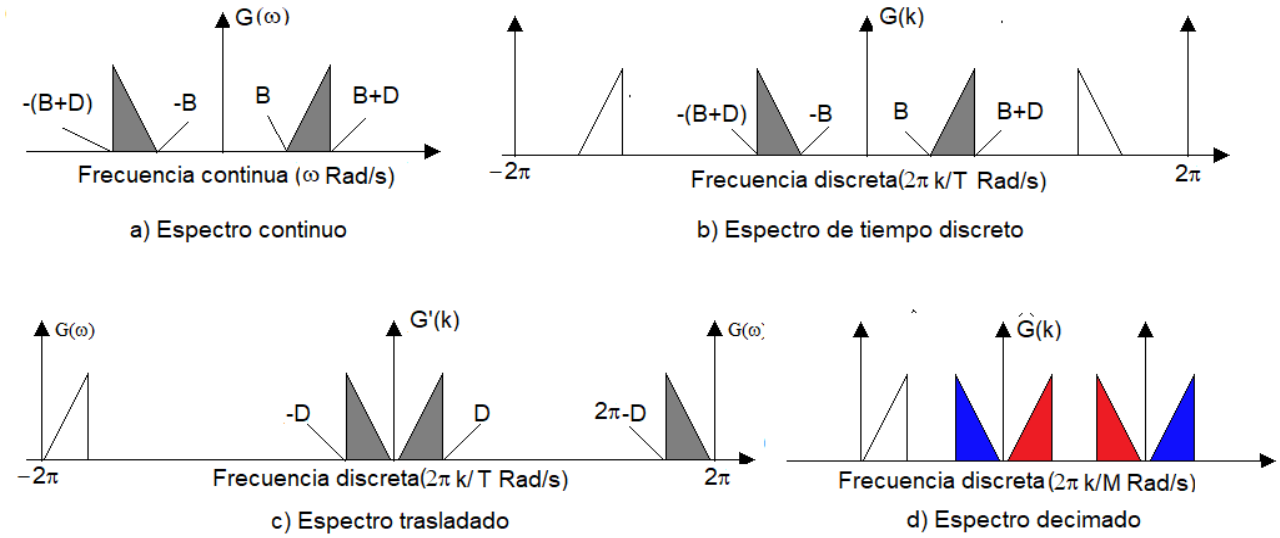


Fig. 5: Tratamiento del espectro de la señal.

Por simplicidad de notación hagamos $Y_r(k)=X(k)Y_r^*(k)$, a la vez haciendo el cambio de variables k=m+B en la primera sumatoria y k=m+T-(B+D) en la segunda sumatoria, de la ecuación (11) se obtiene:

$$W_r(n) = \sum_{m=0}^{D-1} Y_r(m+B)\,e^{j2\pi n(m+B)/T} + \sum_{m=0}^{D-1} Y_r(k)\,e^{j2\pi n(m+T-B-D)/T}. \quad (12)$$

Así, la señal trasladada estará dada por:

$$W_r(n) = e^{j2\pi nB/T}\sum_{m=0}^{D-1} Y_r(m+B)\,e^{j2\pi nm/T} + e^{-j2\pi nB/T}\sum_{m=0}^{D-1} Y_r(k)\,e^{j2\pi n(m+T-D)/T}. \quad (13)$$

Esto se muestra en la Figura 5(c). De lo anterior se desprende que el rango de frecuencias se puede reducir de 0≤k≤T a 0≤k≤M, como se muestra en la Figura 5(c) dado que en el rango de frecuencias de D≤k≤T-D el valor de las frecuencias es cero. Así, reduciendo la frecuencia de muestro de T a M, donde M<T, lo cual equivale a un proceso de decimación (Figura 5(d)), se puede obtener una reducción importante en la complejidad computacional al calcular la IFFT. Para llevar a cabo este proceso se define:

$$\hat{G}_r(k) = \begin{cases} Y_r(k+B), & k = 1,2,\ldots,D-1, \\ 0, & k = D,D-1,\ldots,M-D-1, \\ Y_r(k+T-B-M), & k = M-D,M-D+1\ldots,M. \end{cases} \quad (14)$$

Así, finalmente la r-ésima componente de la transformada wavelet de la señal de entrada x(n) está dada por:

$$W_r(n) = \sum_{k=0}^{M-1} \hat{G}_r(k)\,e^{2\pi nm/M}. \quad (15)$$

Finalmente se tiene que la $W_r(n)$ puede obtenerse mediante la transformada inversa de Fourier con M puntos en lugar de usar T puntos. Así dado que M<<T el número de operaciones se reduce de manera significativa. El algoritmo implementado por (Nakamura y Kameoka, 2014) toma en cuenta lo anterior y el hecho que la FFT se calcula más rápido para señales con longitudes de potencia de base 2, para acelerar el computo de la CWT, a partir de la ecuación (15), primero encuentra el inicio N y fin M de cada función wavelet dilatada (Figura 6), seguidamente se busca una longitud, L, con base en las potencias de 2 que sean mayores o iguales a M-N, ecuación (16), y finalmente se calcula la CWT para esta longitud usando la ecuación (17).

$$L = 2^n, \text{ con } (2^n \geq M - N) \text{ y } n = 0,1,2,\ldots,\infty, \tag{16}$$

$$W_r(n) = \sum_{k=0}^{L-1} \hat{G}_r(k)\, e^{2\pi nm/L}\,. \tag{17}$$

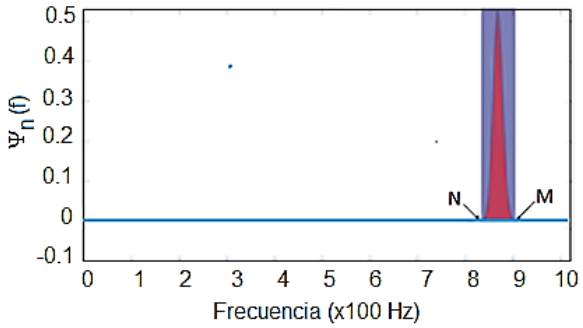


Fig. 6: Localización de inicio N y fin M de cada filtro.

Nakamura y Kameoka (2014) utilizan la función wavelet madre log-normal (18), definida en el dominio de su trasformada de Fourier, para obtener una aproximación rápida de la CWT.

$$\Psi(\omega) = \begin{cases} e^{\left(-\frac{(\log \omega)^2}{4\,\sigma^2}\right)} & si \quad x(\omega > 0), \\ 0 & si \quad x(\omega < 0). \end{cases} \tag{18}$$

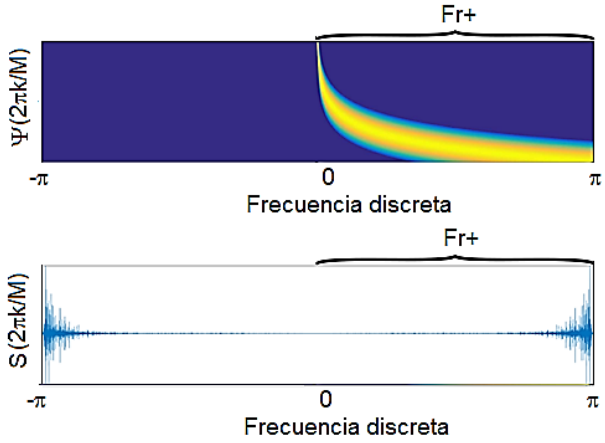


Fig. 7: Transformada de Fourier de la función madre wavelet a diferentes escalas y una señal respectivamente

Finalmente, proponemos un algoritmo para obtener una aproximación rápida de la CWT, a partir de observar la transformada de Fourier de una señal de audio y de una función wavelet madre dilatada a diferentes escalas (figura 7), cuando se calcula el espectro de una señal, el resultado del contenido espectral tiene frecuencias positivas y negativas, donde básicamente el comportamiento es como de un espejo entre las frecuencias positivas y negativas. Esto se debe a la propiedad de las secuencias reales de la FFT la cual establece que G(k)=G(N-k)=G-k), así con el fin de reducir las operaciones y por lo tanto el tiempo de cómputo, se usan solamente las frecuencias positivas, $\omega \geq 0$ tanto para la señal transformada $S(\omega)$ como para los filtros calculados $\Psi_r(a\omega)$, y se aplica el filtro solo para valores donde $\Psi_r(a\omega)$, es diferente de 0. Esto es la ecuación (17), se puede calcular eficientemente usando la IFT (Inverse Fourier Transform) como sigue,

$$W_r(n) = \sum_{k=0}^{D-1} X(k)\, \Psi_r^*(k)\, e^{2\pi nk/D} \tag{19}$$

Redes Neuronales Convolucionales

Recientemente clasificadores basados en técnicas de aprendizaje profundo (Fierro, 2019), en particular las DNN han mostrado ser útiles para las tareas de ESC (Piczak, 2015a; Salamon and Bello, 2017; Cakic et al., 2017). La motivación proviene de modelos de visión por computadora, donde las DNN han sido utilizadas exitosamente transfiriendo el aprendizaje de un dominio a otro, es decir de una tarea a otra (Cakir et al., 2017; Jiang et al., 2015). El enfoque de la transferencia de aprendizaje en las tareas de ESC hasta el momento está casi inexplorado. En este planteamiento se toma una red previamente entrenada para aprender una nueva tarea tomando ventaja que ya aprendió representaciones características para otro tipo de señales. Las redes neuronales profundas están conformadas por millones de pesos que conectan las diversas capas de neuronas de estas redes. Estos pesos se ajustan durante el proceso de entrenamiento a partir de la salida obtenida y la respuesta deseada.

Transferencia del aprendizaje

Existen varias redes de aprendizaje profundo con un rendimiento de vanguardia (a veces tan bueno o incluso mejor que el rendimiento humano) que se han desarrollado y probado en áreas como la visión por computadora y el procesamiento del lenguaje natural (NLP). En la mayoría de los casos, los analistas comparten los detalles de estas redes para que otros las usen. Estos modelos pre entrenados forman la base del aprendizaje de transferencia en el contexto del aprendizaje profundo. Los modelos de aprendizaje profundo son arquitecturas que aprenden diferentes características a través de sus diversas capas. Estas capas se conectan a una última capa de clasificación/regresión (generalmente una capa completamente conectada, en el caso del aprendizaje supervisado) para obtener el resultado final.

Revisemos las dos estrategias más populares para aplicar la transferencia de aprendizaje en las DNN. a) Modelos pre entrenados usados como extractores de características: La arquitectura en capas permite utilizar una red pre entrenada sin la capa de clasificación/regresión como un extractor de características para otras tareas. b) Ajuste fino de modelos pre entrenados: en esta técnica reemplazamos la capa final (clasificación/ regresión) para adaptarla al número de clases que se desea clasificar y reentrenamos selectivamente algunas de las capas anteriores. Se ha observado que las capas iniciales capturan características genéricas, mientras que las posteriores se centran en las características específicas. Las DNN son arquitecturas altamente configurables con varios hiper parámetros. Teniendo en cuenta esto, podemos congelar ciertas capas, esto es fijar los pesos, mientras se reentrenan y se ajustan el resto de las capas, adecuando los parámetros, para realizar a una nueva tarea. Las redes pre entrenadas generalmente consisten en millones de parámetros o pesos que se optimizaron durante la etapa de entrenamiento para llevar a cabo una gran variedad de tareas. Debido a que existe una amplia variedad de redes pre entrenadas, se llevó a cabo un extenso análisis para determinar la red que permita una mejor clasificación de los espectrogramas CWT, para lo cual se empleó el mecanismo de transferencia de aprendizaje. Debido a la baja cantidad de archivos de audio disponibles se usó el enfoque de ajuste fino para llevar a cabo la adaptación de las redes mostradas en la Tabla 2, las cuales han mostrado su eficacia en la clasificación de imágenes.

METODO PROPUESTO

El sistema propuesto para la clasificación de sonidos ambientales e identificación de senarios acústicos se muestra en la figura 8.

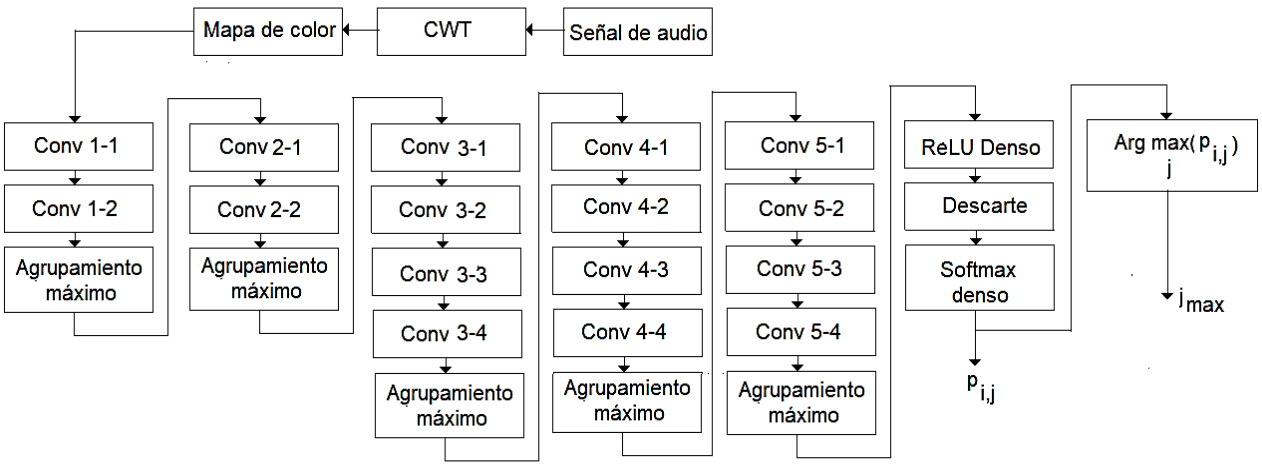


Fig. 8. Sistema propuesto para la clasificación de sonidos y escenarios ambientales

Inicialmente la CWT se emplea para obtener una representación tiempo frecuencia, de la señal de audio bajo análisis, con una resolución logarítmica similar a como percibe los sonidos el sistema auditivo humano. Seguidamente dado que la imagen del espectrograma CWT se emplea para caracterizar el audio de entrada, esta se representa usando el mapa de color que proporcione una mejor caracterización del espectrograma CWT. Finalmente, la imagen del espectrograma CWT se inserta en una red DNN, usando el enfoque de transferencia de aprendizaje para su entrenamiento. Durante la etapa de reconocimiento se inserta el espectrograma CWT en la DNN para clasificar la señal de entrada obteniéndose los valores softmax, $p_{j,i}$, de cada clase. Finalmente, si se desea clasificar el evento acústico, los valores de $p_{j,i}$, se insertan en un módulo que estima el argumento, j, correspondiente al valor máximo de $p_{j,i}$, que representa el evento acústico detectado como se muestra en la figura 8.

Cálculo rápido de la CWT

Como mencionamos anteriormente, lo primero para implementar la CWT es seleccionar de función wavelet madre apropiada para el análisis de la señal, que en nuestro caso son audios, para lo cual implementamos la aproximación rápida de CWT (19) para las funciones wavelet madre de Morlet, Mexican Hat, Meyer, Paul, Shannon, dadas en la Tabla 1, y la log-normal, dada por (18), que son algunas de las funciones wavelet madre más usadas en el análisis de señales no estacionarias, como lo son los sonidos.

Tabla 1: Funciones wavelet madre

Nombre	*Función* $\psi(t)$	
Morlet	$\psi(t) = \pi^{1/4} \cdot e^{i\omega_0 t} \cdot e^{t^2/2}$	(20)
Paul	$\psi(t) = \frac{2^m i^m m!}{\sqrt{\pi(2m)!}} \cdot (1-it)^{-(m+1)}$	(21)
Shannon	$\psi(t) = sinc(t) \cdot e^{-i2\pi t}$	(22)
Sombrero mexicano	$\psi(t) = 2/\sqrt[4]{\pi} \cdot (1-t^2) \cdot e^{-t^2/2}$	(23)
Meyer	$\psi_{mey}(t) = \psi_1 + \psi_2$ $\psi_1 = \frac{\frac{4}{3\pi}(t-0.5)\cos\left[\frac{2\pi}{3}(t-0.5)\right] - \frac{1}{\pi}\sin\left[\frac{4\pi}{3}(t-0.5)\right]}{(t-0.5) - \frac{16}{9}(t-0.5)^3}$ $\psi_2 = \frac{\frac{8}{3\pi}(t-0.5)\cos\left[\frac{8\pi}{3}(t-0.5)\right] - \frac{1}{\pi}\sin\left[\frac{4\pi}{3}(t-0.5)\right]}{(t-0.5) - \frac{64}{9}(t-0.5)^3}$	(24)

En la función Morlet (20) ω_0 es la frecuencia no dimensional, este valor depende de la frecuencia de muestreo de la señal que se desea analizar, nosotros usamos ω_0=200 para obtener el CWT espectrograma mostrado en la figura 9(b). En la función Paul (21) m es el orden de esta función que también depende de la frecuencia de muestreo, aquí nosotros empleamos m=85, para obtener el CWT espectrograma mostrado en la figura 9(f) y en la función madre Log-normal (18) ω es la frecuencia angular y σ es la desviación estándar. Aquí utilizamos σ=0.02 para obtener el CWT espectrograma mostrado en la figura 9(a).

Se decidió usar cinco segundos de una grabación de piano debido a que este instrumento de cuerdas percutidas produce tonos con frecuencias fundamentales que se acompañan con sobre tonos. Así, teniendo esto en cuenta se busca encontrar la mejor representación a partir de las funciones wavelet madre presentadas en la Tabla 1. El audio usado es una grabación monoaural con 16 bits/muestra en formato WAV con una frecuencia de muestreo de 32 kHz. Con esta frecuencia de muestreo el CWT espectrograma obtenido contiene 9 octavas y para obtener una adecuada resolución a la vista se determinó colocar 20 voces entre cada octava. También se decidió usar el mapa de color Jet debido a que tiene una gran variación de tono. Los espectrogramas obtenidos para cada función wavelet madre se muestran en la figura 9.

Lo que se espera observar en un espectrograma de una grabación de piano, es la presencia de la frecuencia fundamental y las resonantes en múltiplos de esta frecuencia de la cuerda percutida. Al observar los CWT espectrogramas de la figura 9, podemos inferir que las funciones wavelet madre que muestran el comportamiento esperado son las de Morlet y Log-normal. Las otras funciones muestran más correlación con las componentes de las frecuencias vecinas, lo cual no permite apreciar de manera adecuada la frecuencia fundamental y sus resonantes presentes en la señal analizada. Debido a que la función Log-normal es la que muestra con mejor detalle la ubicación de los tonos y la presencia de sus resonantes, se decidió usar esta función para obtener la más adecuada representación del CWT espectrograma de nuestra señal al proseguir con la clasificación.

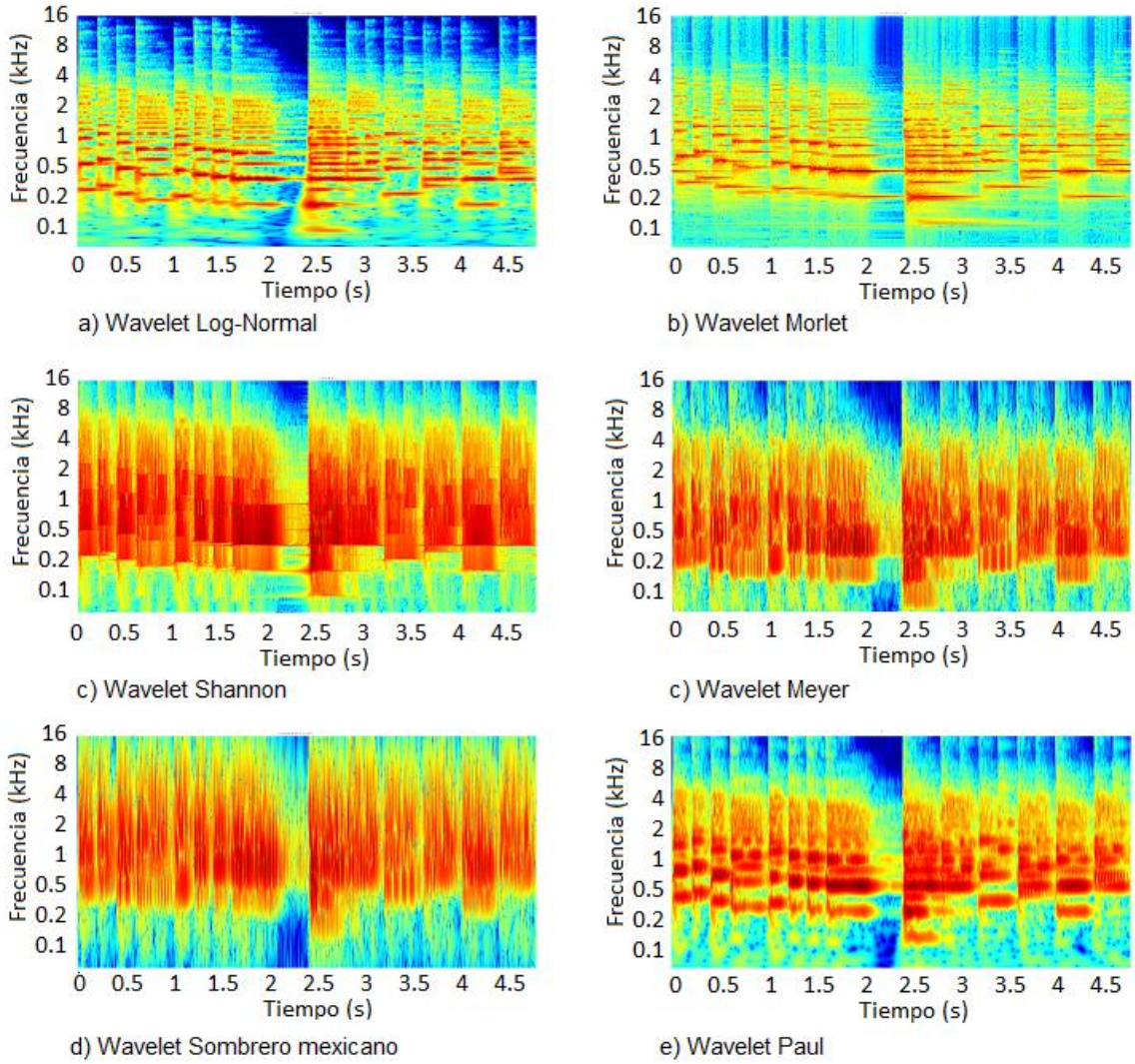


Fig. 9: CWT espectrograma implementado con las funciones madre wavelet Log-normal, Morlet, Shannon, Meyer, Sombrero mexicano y Paul.

RESULTADOS Y DISCUSIÓN

Con el fin de determinar los parámetros óptimos del sistema se llevaron a cabo diversos experimentos que permitieron determinar la CWT, así como la red DNN adecuadas para la ESC. Para esta finalidad se procedió a calcular el espectrograma CWT obtenido de los audios de las bases de datos de sonidos ambientales, los cuales se usaron para realizar la tarea de ESC usando como clasificador una DNN pre entrenada. Las pruebas se llevaron a cabo en una computadora que cuenta con un procesador Intel(R) Core (TM) i5-9400 CPU @2.9 GHz, 32 GB de memoria RAM y el software MATLAB® 2019b.

Bases de Datos

Las bases de datos empleadas para llevar a cabo la evaluación del sistema propuesto son: a) ESC-50 (Piczak, 2015b), la cual contiene 2000 archivos de audio de sonidos ambientales con 5 segundos de duración cada uno, en formato WAV, con frecuencia de muestreo de 44,100 Hz. Está dividida en 50 clases con 40 muestras por cada clase. Esta base de datos incluye sonidos tales como vocalizaciones de animales, sonidos provocados por el agua, paisajes sonoros naturales, sonidos no vocalizados emitidos por humanos, sonidos domésticos y sonidos urbanos. b) La segunda base de datos desarrollada por nosotros, llamada SONAM-50, contiene 1200 grabaciones de diferentes longitudes de tiempo, en formato WAV, con frecuencia de muestreo de 32,000 Hz. Está basada en sonidos usados en el estudio de similitud y categorización de sonidos ambientales reportado por Gygi (Gygi et al., 2007). Esta base de datos tiene 50 clases de sonidos ambientales tales como: sonidos producidos por maquinas, sonidos de varias condiciones climáticas, sonidos humanos no vocalizados, vocalizaciones de animales y sonidos generados por actividades humanas. Cada uno con 24 muestras por cada clase. Como preparación para el procesamiento de estas bases de datos se aplicó una compuerta de ruido, usando como umbral -30 dB para la máxima amplitud de la grabación antes de saturar y

se normalizaron las grabaciones a un nivel de -3 dB del nivel máximo en la grabación. c) La tercera base usada, la TUT Acoustic Scene (Mesaros A. et al. 2017) la cual consiste en grabaciones de varios escenarios acústicos con distintas locaciones de grabación, para cada locación se realizaron grabaciones de 3 a 5 minutos y después fueron divididos en archivos de 10 segundos de duración, estos tienen una frecuencia de muestreo de 44100 Hz, 24 bits de resolución y estereofónica. Contiene 15 distintos escenarios acústicos con 312 muestras por cada escenario, estos escenarios los conforma grabaciones hechas en algún transporte (autobús, automóvil, tren, tranvía), en el interior de un recinto (cafetería, tienda, casa, biblioteca, estación de metro, oficina) y en el exterior (ciudad, camino forestal, playa, residencial, parque).

Ajuste fino de DNN pre entrenadas

Para encontrar la DNN que mejor se ajuste a la tarea de ESC y tener un punto de comparación con otros trabajos (Xie y Zho, 2019; Demir et al., 2018; Kumar y col, 2018; Abdoli et al. 2019; Piczak, 2015a), usamos los espectrogramas CWT de la base de datos ESC-50 para realizar el ajuste fino con las DNN. Después de aplicar la compuerta de ruido a los archivos de la base ESC-50, obtenemos audios con diferentes duraciones; por lo cual para obtener la misma resolución de todos los espectrogramas CWT se repitieron segmentos de los sonidos menores a 5s. Con el fin de producir una representación a color de los espectrogramas CWT se usó el mapa de color Viridis que ha sido usado en otros trabajos de clasificación de sonidos (Noda et al., 2015; Demir et al., 2018). Las opciones de entrenamiento utilizadas en esta tarea son: Descenso de gradiente estocástico con impulso "*Stochastic Gradient Descent with Momentum"* (SGDM) con un impulso igual a 0.9, una tasa de aprendizaje inicial igual a 3×10^{-4}, un tamaño de lote igual a 6, una frecuencia de validación de 3 y 30 épocas. La base de datos se dividió aleatoriamente en el 80% para entrenamiento, 20% para validación y sin usar la técnica de aumento de datos.

Tabla 2: Ajuste fino de redes pre-entrenadas.

Red	*Profundidad*	*Tamaño de la entrada*	*Capas pre-entrenadas*	*Precisión*
Alexnet	25	227 X 227	6	71.10
VGG16	41	224 X 224	4	70.17
VGG19	47	224 X 224	14	72.17
Squeezenet	68	227 X 227	14	66.17
Googlenet	144	224 X 224	6	64.67
Incepction3	315	229 X 229	12	66.50
Resnet18	71	224 X 224	12	69.00
Resnet50	177	224 X 224	12	69.33
Resnet101	347	224 X 224	12	69.33
Xception	170	229 X 229	15	60.17
Inceptionrenetv2	824	229 X 229	5	64.17

El resultado de realizar el ajuste fino para la adaptación de las redes neuronales a la tarea de ESC, usando la base de datos ESC-50 se muestra en la Tabla 2. En esta se presenta la profundidad de cada red es decir el número de capas de la DNN, el tamaño de entrada son las dimensiones de las imágenes para las cuales fue diseñada la red, el número de capas convolucionales cuyos parámetros se mantuvieron fijos lo que permitió obtener el mejor resultado de clasificación y el porcentaje de precisión para clasificar los sonidos ambientales contenidos en ESC-50 de cada red. En la Tabla 2 podemos observar que la red DNN VGG19, es la que muestra el mejor desempeño al clasificar los espectrogramas CWT por lo cual es la seleccionada para llevar a más detalle el ajuste fino del sistema.

Selección del mapa de color

Debido a que no hay evidencia que muestre el impacto del mapa de color aplicado a los espectrogramas CWT para realizar la tarea de ESC, se evaluó el comportamiento del sistema cambiando el mapa de color a los espectrogramas CWT obtenidos usando 10 voces por octava de la base de datos ESC-50 antes de insertarlos en la DNN. Los resultados obtenidos, reportados en la Tabla 3, muestran que los mapas de color que proporcionan mayor porcentaje de precisión en la clasificación son: Jet, Viridis, Gray, Magma e Inferno. La evaluación del sistema se llevó a cabo usando el algoritmo SGMD con un valor de impulso igual a 0.9, una tasa de aprendizaje inicial igual a 3×10^{-4}, tamaño de lote de 6, frecuencia de validación igual a 3 y 30 épocas, mezclando los datos de entrenamiento y validación de manera aleatoria. La tasa de entrenamiento se

actualiza cada dos épocas multiplicándola por un factor de 0.9. Se emplea un factor de regularización 0.005 y el mapa de color Viridis, para representar el espectrograma. La base de datos, la cual consiste en 50 sonidos ambientales diferentes con 40 muestras de cada uno, se dividió aleatoriamente en 80% para entrenamiento y 20% para validación sin usar el procedimiento de aumento de datos.

Tabla 3: Precisión de clasificación para diferentes mapas de color.

Mapa de color	*Precisión*
Viridis	76.16
Magma	73.67
Inferno	73.67
Fake parula	53.83
Jet	76.83
HSV	53.83
Hot	53.83
Cool	68.17
Summer	73.67
Autumn	71.50
Winter	71.50
Gray	76.67
Bone	71.67
Copper	72.83
Pink	71.00

Análisis de la frecuencia de muestreo y duración de la señal en ESC

Otro factor importante que debe ser analizado es la frecuencia de muestreo y la duración de la señal de entrada. La base de datos ESC-50 contiene sonidos impulsivos y/o repetitivos, por ejemplo: abrir una lata de soda, fuegos artificiales, ladridos, truenos, etc., y continuos como lo son el ruido producido por las turbinas de un avión, el grillar de grillos, un motor trabajando, el tic tac de un reloj, que al extraer los silencios de los audios tienen duraciones que van desde 0.5 segundos a decenas de segundos. Por lo tanto, para observar el impacto que tiene en la clasificación la duración de la señal bajo análisis, se usaron diferentes longitudes para clasificar sonidos ambientales. Para unificar la duración de los sonidos, las señales con menor tiempo se repiten hasta alcanzar la duración requerida en el análisis, de igual manera se reduce la frecuencia de muestreo de los archivos de audio para las distintas duraciones de los sonidos. Los resultados de este experimento se presentan en la Tabla 4, siendo las opciones de entrenamiento utilizadas las mismas del experimento anterior.

Tabla 4: Precisión en la clasificación de ESC usando distintas frecuencias de muestreo y tiempo de la señal.

Frecuencia de muestreo	*5s*	*2.5s*	*1.5s*	*1s*	*0.5s*
8 KHz	69.50	67.50	63.00	60.17	50.33
16 KHz	72.50	69.17	69.17	64.33	54.83
22.05 KHz	75.67	75.33	76.17	70.67	60.67
32 KHz	76.67	77.33	74.5	70.33	64.17
44.1 KHz	76.83	75.00	76.17	58.17	58.17

Análisis del número de voces por octava

La selección del número de voces para el análisis de señales de audio al usar la CWT es algo de gran importancia. A este respecto existen recomendaciones para el uso de entre 8 y 12 voces por cada octava (Gersem et al., 1977), para obtener una adecuada resolución a la vista humana y no tener un elevado costo de cómputo. Para analizar el efecto que tiene la selección del número de voces en la clasificación de los CWT espectrogramas de sonidos ambientales, se realizó una prueba cambiando el número de voces en potencias enteras de base dos, desde 1 hasta 32. Las condiciones del entrenamiento utilizadas son las mismas del experimento anterior y la duración máxima para los sonidos fue de 2.5 segundos. Los resultados obtenidos al identificar los 50 sonidos ambientales contenidos en la base de datos ESC-50, se pueden observar en la Tabla 5.

Tabla 5: Precisión de clasificación para diferentes números de voces.

Numero de voces por octava	*Precisión de clasificación*
1	53.83
2	59.67
4	70.67
8	74.50
16	77.33
32	79.67

Precisión proporcionada por el sistema propuesto.

Una vez determinados los parámetros óptimos del sistema propuesto, se procedió a evaluar su capacidad de reconocimiento. Para evaluar el funcionamiento del sistema propuesto basado en la CWT y DNN, se emplearon las bases de datos ESC-50 y SONAM-50, las cuales contienen 50 sonidos como se muestra en la Tabla 6. Para llevar a cabo la evaluación, los CWT espectrogramas se calcularon usando la función wavelet madre Log-normal, las señales se muestrearon a 32kHz, con 32 voces por octava y el mapa de color Viridis.

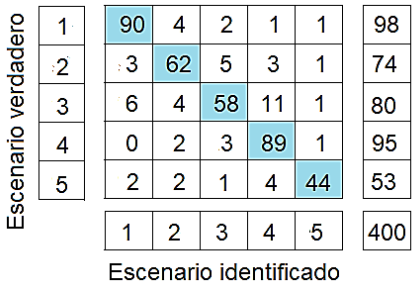
Tabla 6: Clases de la base de datos ESC-50 y SONAM-50.

clase	*ESC-50*	*SONAM-50*	*clase*	*ESC-50*	*SONAM-50*
1	abrir lata	aceleración de automóvil	26	moto-sierra	limpia parabrisas
2	alarma	aplaudir	27	Clic de mouse	lluvia
3	aplaudir	arpa	28	olas	moto-sierra
4	aspiradora	andar de automóvil	29	oveja	olas
5	avión	ave	30	pájaros	oveja
6	bebe	avión	31	pasos	pasos
7	beber	balón	32	perro	perro
8	bocina de auto	bebe	33	puerco	pelota de ping pong
9	campana	bocina de auto	34	puerta de madera	platos rompiéndose
10	serrucho	bolos	35	rana	puerta
11	chisporroteo	burbujas	36	reloj	relinchar
12	cuervo	campana	37	respirar	reloj
13	estornudar	chapoteo	38	risas	risas
14	fuegos artificiales	cierre	39	ronquidos	silbato
15	gallina	cristal	40	cepillar dientes	sirena
16	gallo	disparo	41	sirena	tambor
17	gato	estornudo	42	teclear	teclado de computadora
18	golpe de puerta	fósforo	43	inodoro	teclear máquina
19	gotas de agua	gallo	44	toser	teléfono
20	grillos	galope	45	tren	tijeras
21	helicóptero	gato	46	trueno	toser
22	insectos volando	gotas de hielo	47	vaca	tren
23	lavadora	hacha	48	verter agua	trueno
24	lluvia	helicóptero	49	vidrios rompiéndose	vaca
25	motor	inodoro	50	viento	verter agua

La capacidad de clasificación del método propuesto se evaluó en dos formas diferentes. En la primera, el sistema propuesto es requerido a clasificar diferentes eventos acústicos (Piczak, et al., 2015a; Sailor et al 2017), para lo cual los archivos contenidos en la base de datos ESC-50 se agruparon en 5 eventos diferentes: a) sonidos de animales, b) paisajes sonoros y sonidos de agua, c) sonidos humanos no vocalizados, d) sonidos de interiores, e) sonidos exteriores y ruidos urbanos (Sailor et al., 2017; Kumar et al., 2018). Para la evaluación, la base de datos se dividió en 80% de los datos para entrenar y 20% para para evaluar. Los resultados obtenidos se muestran en la Tabla 7, mientras que la matriz de confusión obtenida por el método propuesto, junto con una descripción de los sonidos en cada evento se muestran en la figura 10. Los resultados obtenidos muestran que el sistema propuesto presenta una clasificación superior a la proporcionada por otros esquemas de clasificación de eventos acústicos, incluidos en la base de datos ESC-50, empleando redes neuronales profundas y muy cercana al funcionamiento proporcionado por el esquema propuesto por Sailor (Sailor et al, 2017) usando Convolutional Restricted Boltzman Machine (ConvRBM). Los resultados obtenidos muestran además que el sistema propuesto presenta mejores resultados que otros esquemas basados en transformadas wavelet y redes neuronales profundas presentados como parte de los eventos “Detection and Classification of Acoustic Scenes and Events (DCASE)” (Mesaros et al., 2017; Mun et al. 2017).

Tabla 7: Comparación de la precisión de clasificación del CWT espectrograma con otros esquemas reportados en la literatura, para la base de datos ESC-50.

Referencia	*Método*	*Precisión*
(Piczak, 2015a)	Mel Espectrograma + CNN	64.50
(Demir et al., 2018)	Deep Spectrum (VGG-M, VGG-D) + SVM	74.60
(Sailor et al., 2017)	ConvRBM + CNN	78.45
(Sailor et al., 2017)	FBEs + ConvRBM + BANK	86.50
(Kumar et al., 2018)	Mel Espectrograma + CNN	83.50
(Piczak, 2015b)	Oído	81.30
Propuesto	CWT espectrograma + CNN	85.75



a) Matriz de confusión

	Evento	Sonidos en cada evento
1	Sonidos animales	12, 15, 16, 17, 20, 22, 29, 30, 32, 33, 35, 47
2	Paisajes sonoros y sonidos de agua	14, 19, 24, 28, 43, 46, 48
3	Sonidos humanos no vocalizados	6, 7, 11, 13, 38, 37, 39, 44
4	Sonidos interiores	1, 2, 3, 4, 10, 18, 23, 27, 31, 34, 35, 40, 42, 49
5	Sonidos exteriores	5, 8, 9, 21, 25, 26, 41, 45, 50

b) Tipo de eventos acústicos relacionados a la Tabla 6.

Fig. 10. Matriz de confusión obtenida usando el sistema propuesto cuando se requiere identificar los eventos mostrado en la Tabla 6.

La capacidad de clasificación del método propuesto se evaluó en dos formas diferentes. En la primera, el sistema propuesto es requerido a clasificar diferentes eventos acústicos (Kumar et al., 2018; Piczak, et al., 2015a; Sailor et al 2017; Kumar et al., 2018), para lo cual los archivos contenidos en la base de datos ESC-50 se agruparon en 5 eventos diferentes: a) sonidos de animales, b) paisajes sonoros y sonidos de agua, c) sonidos humanos no vocalizados, d) sonidos de interiores, e) sonidos exteriores y ruidos urbanos (Sailor et al., 2017; Kumar et al., 2018). Para la evaluación, la base de datos se dividió en 80% de los datos para entrenar y 20% para para evaluar. Los resultados obtenidos se muestran en la Tabla 7, mientras que la matriz de confusión obtenida por el método propuesto, junto con una descripción de los sonidos en cada evento se muestran en la figura 10.

Los resultados obtenidos muestran que el sistema propuesto presenta una clasificación superior a la proporcionada por otros esquemas de clasificación de eventos acústicos, incluidos en la base de datos ESC-50, empleando redes neuronales profundas y muy cercana al funcionamiento proporcionado por el esquema propuesto por Sailor (Sailor et al, 2017) usando Convolutional Restricted Boltzman Machine (ConvRBM). Los resultados obtenidos muestran además que el sistema propuesto presenta mejores resultados que otros esquemas basados en transformadas wavelet y redes neuronales profundas presentados como parte de los eventos “Detection and Classification of Acoustic Scenes and Events (DCASE)” 2017 (Mesaros et al.; Mun et al., 2017).

Seguidamente se llevó a cabo la evaluación del sistema propuesto, cuando este es requerido a identificar cada uno de los 50 sonidos contenidos tanto en la base de datos ESC-50 y SONAM-50, usando los mismos parámetros que en la evaluación anterior. Aquí, nuevamente, ambas bases datos se dividieron en 80% para entrenar y 20% para prueba. Bajo estas condiciones los resultados obtenidos son 80% de precisión usando la base de datos ESC-50 y 87.6% usando la base SONAM-50. El porcentaje de precisión en la clasificación obtenida demuestra que el sistema propuesto es una herramienta competitiva para la caracterización no solamente de eventos acústicos, sino de sonidos ambientales individuales. Las matrices de confusión obtenidas al clasificar las bases de datos ESC-50 y SONAM-50 se muestran en las figuras 11 y 12 respectivamente.

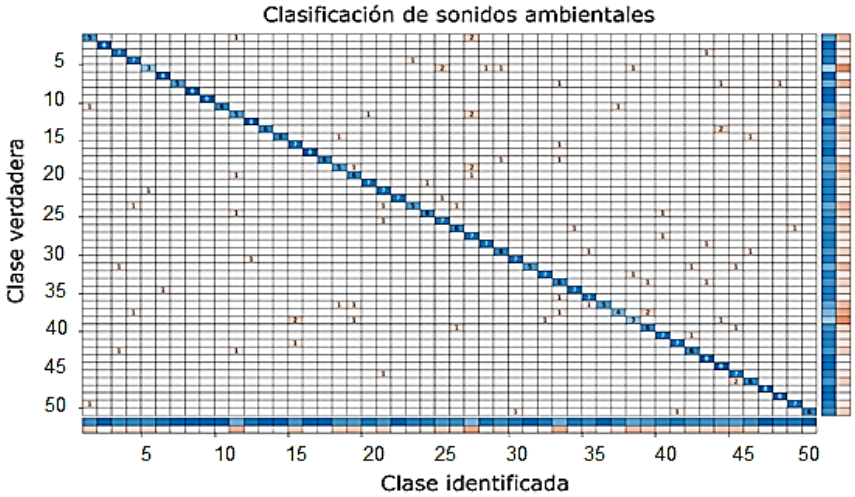


Fig. 11: Matriz de confusión al clasificar la base de datos ESC-50.

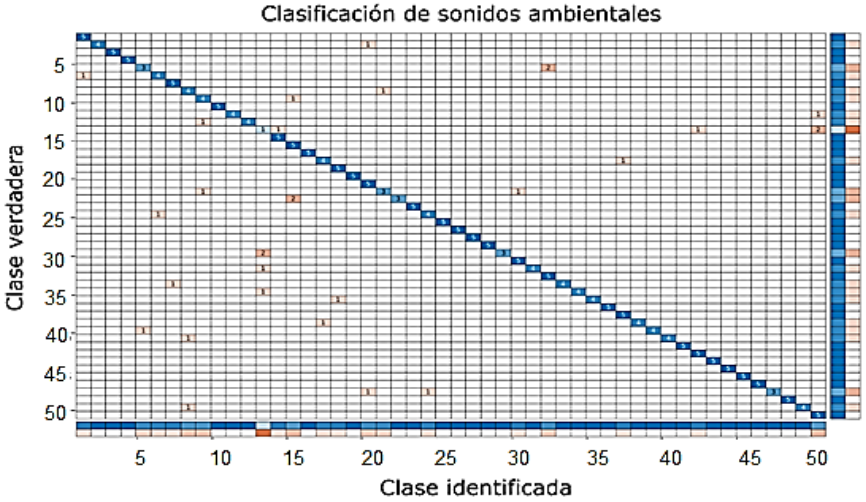


Fig. 12: Matriz de confusión al clasificar la base de datos SONAM-50.

También, se evaluó el sistema propuesto cuando se requiere llevar a cabo la clasificación de escenarios acústicos o paisajes sonoros los cuales contienen diversos sonidos (por ejemplo, personas hablando, el trinar de los pájaros, perros ladrando, etc.) a diferentes niveles, dependiendo de la localización del punto de grabación, por lo que los escenarios acústicos se considera todos los sonidos presentes en dada escena y un evento acústico se consideran solo fuentes individuales (Mesaros A. et al. 2017, Waldekar y Saha, 2020; Abeber, 2020)). Para esta evaluación se utilizó la base de datos TUT Acoustic scene, que utiliza el 70% de la base de datos para entrenamiento y 30% de validación, los CWT espectrogramas se calcularon usando la función wavelet madre Log-normal, no se modifica la frecuencia de muestreo de las señales 44100 Hz, con 32 voces por octava, se usa como longitud del audio 2.5 segundos y el mapa de color Viridis. La evaluación del sistema se llevó a cabo usando el algoritmo SGMD con un valor de impulso igual a 0.9, una tasa de aprendizaje inicial igual a 3×10^{-4}, tamaño de lote de 6, frecuencia de validación igual a 30 y 30 épocas, mezclando los datos de entrenamiento y validación de manera aleatoria. La tasa de entrenamiento se actualiza cada dos épocas multiplicándola por un factor de 0.9. Se emplea un factor de regularización 0.005 y el mapa de color Viridis, para representar el espectrograma. La Tabla 8 se muestran los resultados obtenidos.

Tabla 8: Precisión de clasificación de escenarios acústicos (TUT Acoustic scene 2017).

Escenario acústico	*Sistema propuesto Precisión*	*Escenario acústico*	*Sistema propuesto Precisión*	*Escenario acústico*	*Sistema propuesto Precisión*
Autobús	93.6	Casa	97.9	Playa	97.9
Automóvil	92.6	Ciudad	92.6	Residencial	70.2
Biblioteca	87.2	Estación metro	81.9	Tienda	93.6
Cafetería	96.8	Oficina	95.7	Tranvía	94.7
Camino forestal	93.6	Parque	84	Tren	90.4
				Promedio	90.85

Tomando en cuenta que el sistema propuesto usa 2.5s para caracterizar, se propone una técnica de aumento de datos generando 7 espectrogramas CWT por cada grabación de 10s. Es decir, espectrogramas de 2.5s con un traslape entre ellos de 1.25s, como se ilustra en la figura 13.

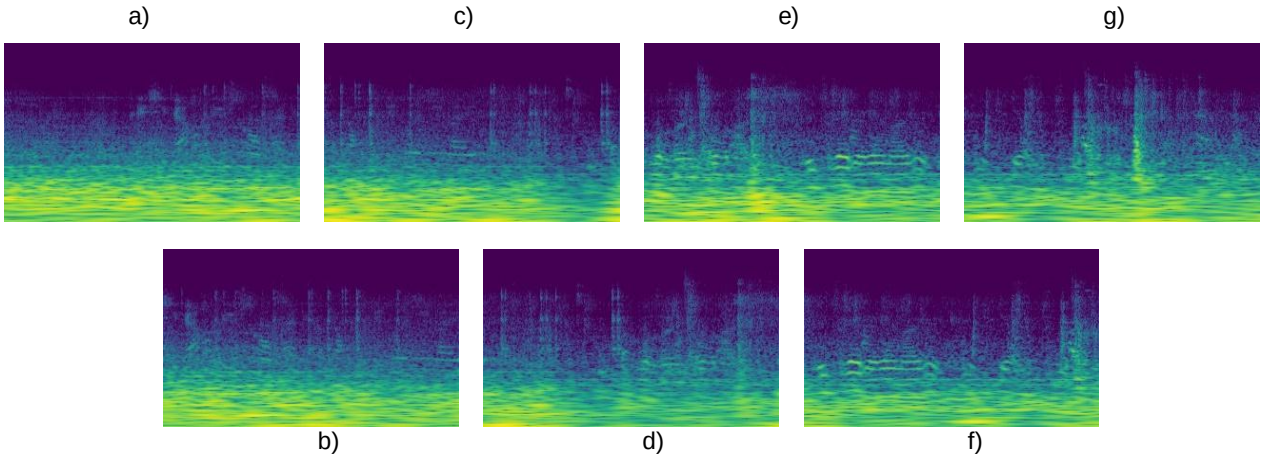


Fig. 13: Segmentación del archivo a_076_40_50.wav. a) 0 – 2.5 s, b) 1.25 – 3.75 s, c) 2.5 – 5 s, d) 3.75 – 6.25, e) 5 – 7.5 s, f) 6.25 – 8.75 s, g) 7.5 – 10 s.

De esta manera obtenemos 2184 muestras para cada clase, para realizar el entrenamiento con estas muestras usamos el 80% para implementar el sistema y 20% para probar, a la vez dividimos la parte de implementación en 70% para entrenar y 30% para validar. Las opciones de entrenamiento son las mismas que en el experimento anterior, excepto por el tamaño de lote y la frecuencia de validación, que se usaron 42 y 210 respectivamente. Además, se implementó un sistema de votación por cada segmento, es decir se toma como predicción final, la clase más seleccionada entre los 7 segmentos, los resultados y comparación con previos trabajos se muestran en la Tabla 9.

Tabla 9: Resultados y comparación de la precisión de clasificación del CWT espectrograma con otros esquemas reportados, para la base de datos TUT Acoustic scene 2017

Escenario acústico	*Sistema propuesto precisión validación*	*Sistema propuesto precisión prueba*	*Sistema propuesto precisión prueba votación*	*(Mum et al. 2017)*	*(Weiping et al. 2017)*	*(Park et al. 2017)*
Autobús	93.6	94.7	100	87.2	93.91	93.8
Automóvil	92.6	81.7	82.3	88.5	97.76	93.2
Biblioteca	87.2	98.2	100	96.2	87.82	77.8
Cafetería	96.8	70	71	87.2	66.03	73.1
Camino forestal	93.6	96.1	98.4	94.9	98.8	97.1
Casa	97.9	82.2	87.1	89.7	91.16	82.5
Ciudad	92.6	96.6	100	98.7	88.46	86.3
Estación metro	81.9	93.1	96.8	84.6	99.689	89.2
Oficina	95.7	87	91.9	96.2	98.08	91
Parque	84.0	62	64.5	71.8	78.21	73.1
Playa	97.9	97.5	100	71.8	85.58	88.5
Residencial	70.2	71	73	87.2	88.78	70.5
Tienda	93.6	97	98.4	79.5	95.19	84.6
Tranvía	94.7	97.9	100	91	92.95	82.7
Tren	90.4	75.5	75.8	82.1	86.22	70.1
Promedio	90.5	86.69	89.26	87.1	89.86	83.6

Los resultados obtenidos muestran que el sistema propuesto presenta resultados competentes con otros esquemas basados en transformadas wavelet y redes neuronales profundas presentados como parte de diversos eventos tale como el IEEE-ICASSP, “Detection and Classification of Acoustic Scenes and Events (DCASE)”, entre otros.

CONCLUSIONES

Este artículo propone un algoritmo para la clasificación o reconocimiento de eventos acústicos o sonidos ambientales basado en la transformada wavelet continua y redes neuronales profundas. Los resultados experimentales, (figura 9) muestran las funciones wavelet madre log-normal y Morlet, proporcionan los CWT espectrogramas más apropiados para caracterizar las señales de audio. Se evaluaron además diversas estructuras de redes DNN y de la Tabla 2 se puede observar que, de las redes evaluadas, la red VGG-19 es la más adecuada para realizar la tarea de ESC. Así mismo, dado que el mapa de color de los espectrogramas tiene un rol importante, se llevó a cabo una evaluación de varios de ellos, obteniéndose que los mapas de color Viridis, Jet y Gray podrían ser adecuados para crear los CWT espectrogramas usado para clasificar los sonidos. Así mismo, se encontró que una duración de 2.5s parece adecuada para llevar a cabo el reconocimiento, al igual que una frecuencia de muestreo de aproximadamente 22kHz. Finalmente se analizó el número de voces por octava que proporcionan el mejor funcionamiento del sistema propuesto, encontrándose que el incremento del número de voces por octava impacta de manera importante en la precisión al implementar ESC. Una vez determinados los parámetros del sistema se evaluó su funcionamiento cuando se requiere detectar tanto eventos acústicos, como sonidos ambientales. Los resultados experimentales obtenidos muestran que cuando el algoritmo propuesto se emplea para reconocer eventos acústicos, éste presenta una tasa de reconocimiento del 85.75% usando la base de datos ESC-50, 89.26% usando las bases de datos ESC-50 y 90.85% usando la base de datos TUT Acoustic scene, respectivamente, la cual es superior al reconocimiento proporcionado por otros esquemas previamente reportados en la literatura, basados en la transformada wavelet y redes DNN; además de ser muy cercano a otro esquema basado en máquinas de Boltzman restringidas (RBM), como se muestra en la Tabla 7. Finalmente se observa que, el sistema propuesto presenta una tasa de reconocimiento de 80% cuando se emplea para el reconocimiento de sonidos ambientales usando la base de datos ESC-50 y 87.6% de reconocimiento cuando se emplea la base de datos SONAM-50.

REFERENCIAS

Abeber J., A Review of Deep Learning Based Methods for Acoustic Scene Classification, https://doi.org/10.3390/app10062020, Applied Science 10 (2020)

Alexandre, E., Cuadra, L., Rosa, M. y Lopez-Ferreras, F., *Feature Selection for Sound Classification in Hearing Aids Through Restricted Search Driven by Genetic Algorithms*, https://doi.org/10.1109/TASL.2007.905139, IEEE Trans. on Audio, Speech and Lang. Process 15 (8), 2249–2256. (2007)

Abdoli S. Cardinal P. y Lameiras K., *End-to-end Environmental Sound Classification using 1D Convolutional Neural Networks*, https://doi.org/10.1016/j.eswa.2019.06.040, Expert Systems and Applications, 136, 252-263 (2019)

Cakir E., Parascandolo G. Heittola T., Huttenen H. y otro autor más, *Convolutional Neural Networks for Polyphonic Sound Event Detection*, https://doi.org/10.1109/TASLP.2017.2690275, IEEE/ACM Trans. on Audio Speech and Language Processing, 25(6), 1291-1303 (2017)

Chu S., Narayanan, S. y Kuo, C., *Environmental Sound Recognition with Time Frequency Audio Features*, https://doi.org/10.1109/TASLP.2009.2017438, IEEE Trans. on Audio Speech and Language Processing, 17(6), 1142-1158 (2009)

Demir, F., Sengur, A., Lu, H., Amiriparian, S. y dos autores mas, *Compact Bilinear Deep Features for Environmental Sound Recognition*, https://doi.org/10.1109/IDAP.2018.8620779, International Conference on Artificial Intelligence and Data Processing (2018)

Fierro A., Nakano M., Yanai K. y Perez H.; *Redes Convolucionales Siamesas y Tripletas para la Recuperación de Imágenes Similares en Contenido*, https://doi.org/10.4067/S0718-07642019000600243, Inf. Tecnológica 30 (6), 243-254 (2019)

Gersem, P. D., Moor, B. D. y Moonen, M., *Applications of the Continuous Wavelet Transform in the Processing of Musical Signals*, https://doi.org/10.1109/ICDSP.1997.628411, Proceedings of 13th International Conference on Digital Signal Processing 2, 563–566 (1977)

Gygi, B., Kidd, G. R. y Watson, C., *Similarity and Categorization of Environmental Sound*, https://doi.org/10.3758/BF03193921, Perception and Psychophysics 69 (6), 839–855 (2007)

Jiang Q, Chang F. y Sheng B., *Bearing Fault Classification Based on Convolutional Neural Networks in Noise Environmental*, https:**//**doi.org/10.1109/ACCESS.2019.2919126 IEEE Access 7, 69795-69807 (2015)

Jiménez G.. Rivas E. y Aparício L., *Compresión de Señales Electrocardiográficas Fetales Mediante la Transformada Wavelet Packet*, https://doi.org/10.4067/S0718-07642018000300145, Inf. Tecnológica 29 (3), 145-154 (2018)

Kumar, A., Khadkevich, M. y Fügen, C*., Knowledge Transfer from Weakly Labelled Audio using Convolution Neural Network for Sound Event and Scenes*, https://doi.org/10.1109/ICASSP.2018.8462200, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 326– 330 (2018)

Martínez A., Compeán I, Fosado R. y Ávila R., *Codificación Esteganográfica usando la Transformada de Onditas Haar Discreta Multi-resolución*, https://doi.org/10.4067/S0718-07642018000400317, Inf. Tecnológica 29 (4), 317-328 (2018)

Mesaros A., Heittola T., Diment A., Elizalde B., y cuatro autores mas*, DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System,* Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), 85–92 (2017)

Mun S., Park S., Han D K. y Ko H*. Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection using SVM Hyperplane*, Detection and Classification of Acoustic Scenes and Events (DCASE), (2017).

Najmi, A., Sadowsky, J., *The Continuous Wavelet Transform and Variable Resolution Time-Frequency Analysis*, Johns Hopkins Appl. Technical Digest, 18 (1), 134–140 (1997)

Nakamura, T., Kameoka, H*., Fast Signal Reconstruction from Magnitude Spectrogram of Continuous Wavelet Transform Based on Spectogram Consistency*, Proc. of the 17th Int. Conference on Digital Audio Effects, DAFx1-14 (2014)

Noda K., Yamaguchi Y., Nakadai K., Okuno H., y otro autor mas, *Audio Visual Speech Recognition using Deep Learning,* https://doi-org/10.1007/s10489-014-0629-7, Applied Intelligence, 42(4), 722-737 (2015)

Park S., Mun S., Lee Y. y Ko H., *Acoustic Scene Classification Based on Convolutional Neural Network Using Double Image Features*., Detection and Classification of Acoustic Scenes and Events DCASE) (2017)

Piczak, K. J., *Environmental Sound Classification with Convolutional Neural Networks,* https://doi.org/10.1109/MLSP.2015.7324337, 25th Int. Workshop on Machine Learning for Signal Processing 1–6 (2015a)

Piczak, K. J*., ESC: Dataset for Environmental Sound Classification*, https://doi.org/10.1145/2733373.2806390, Proceedings of the ACM International Conference on Multimedia (2015b)

Potamitis I., *Automatic Classification for Taxon-rich Community Recorded in the Wild*, https://doi.org/10.137/Journal.pone.00936, PLos One, 9(5) (2014)

Sadowsky, J.; *Investigation of Signal Characteristics using the Continuous Wavelet Transform*, Johns Hopkins Appl. Technical Digest, 17 (3), 258–269 (1996)

Sailor, H., Agrawal, D. y Patil, H., *Unsupervised Filter Bank Learning using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification*, http://doi.10.21437/interspeech.2017-831, Interspeech**,** 3102–3111 (2017)

Salamon, J. y Bello, J. P., *Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification*, https://doi.org/10.1109/LSP.2017.2657381, IEEE Signal Process. Letters, 24 (3), 279–283 (2017)

Santamaría F., Cortés C. y Román F., *Uso de la Transformada de Ondeletas (Wavelet Transform) en la Reducción de Ruidos en las Señales de Campo Eléctrico Producidas por Rayos***,** http:**//**doi.org/10.4067/S0718-07642012000100008*,* Inf. Tecnol*ógica*, 2012, 23(1), 65-78. (2012)

Tong, W., Yang, Y., Jiang, L., Yu, S.-I. y cinco autores más, *E-lamp: Integration of Innovative Ideas for Multimedia Event Detection*, https://doi.org/10.1007/s00138-013-0529-6, Machine Vision and Applications, 25 (1), 5–15 (2014)

Torrence, C. y Compo, G. P., *A Practical Guide to Wavelet Analysis*, http://doi.org/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2, Bulletin of the American meteorological society 79 (1), 61–78 (1998)

Waldecar S. y Saha G., *Analysis and Classification of Acoustic Scenes with Wavelet Transform-Based Mel-Scaled Features*, https://doi.org/10.1007/s11042-019-08279-5, Multimedia Tools and Applications, 79, 7911-7926 (2020)

Xie J. y Zhu M*., Handcrafted Features and Late Fusion with Deep Learning for Bird Sound Classification,* https://doi.prg/10.1016/j.ecoinf.2019.05.007, Ecological and Informatics 52 74-81 (2019)