Codificación perceptual de audio simple para dispositivos de bajo costo

Sergio A. Castells*, Gonzalo D. Sad*, Fernando A. Marengo Rodriguez*†

*Universidad Nacional de Rosario, Rosario, Argentina.
†Universidad Federal de Santa Catarina, Florianópolis, Brasil.
castellssergio@eie.fceia.unr.edu.ar
sad@cifasis-conicet.gov.ar
fmarengo@eie.fceia.unr.edu.ar

Resumen En este documento se propone un nuevo esquema de codificación perceptual de audio basado en técnicas adaptativas. Estos métodos de análisis de señales se basan en la información contenida en dichas técnicas. Aquí se utilizan las siguientes herramientas: 1) descomposición empírica de modos (EMD) y 2) descomposición empírica de modos por conjuntos (EEMD). En comparación con técnicas previas, el esquema de codificación propuesto es simple, ya que descompone la señal de entrada en términos de sus componentes físicamente relevantes, extrae sus extremos locales y posteriormente realiza la codificación Golomb-Rice de los mismos. Se estudia el rendimiento de esta metodología en términos de factor de compresión y calidad perceptual para varias pistas de audio contenidas en el CD SQAM de la Unión Europea de Radiodifusión (EBU). Estos resultados son comparados con los obtenidos utilizando otros codificadores perceptuales de audio.

1. Introducción

Con el objetivo de optimizar el uso de los canales de transmisión de datos, así como la capacidad de almacenamiento, se desarrollaron diferentes esquemas de codificación de audio sin pérdidas (*lossless*) y con pérdidas (*lossy*). Algunos de los comprendidos en este último grupo hacen uso de modelos perceptuales. Los codificadores sin pérdidas permiten reducir el tamaño del archivo de audio sin introducir distorsión. Los archivos de audio contenidos en un CD pueden ser reducidos entre 2 y 6 veces su tamaño original, dependiendo de ciertas características de los datos de entrada, como el rango dinámico y el contenido espectral de los mismos [1]. Como ejemplos de este tipo de codificadores se pueden citar a FLAC [2] y a Monkey's audio [3]. Los codificadores de audio perceptuales como MP3 [4] y OGG Vorbis [5] permiten obtener altos niveles de compresión a cambio de distorsiones no audibles por un oyente promedio, aumentando la complejidad del codificador. Estos codificadores se basan en la transformada de coseno discreta modificada (su sigla en inglés es MDCT). Para ambos tipos de codificación, es crucial que los decodificadores sean de baja complejidad, para permitir que los dispositivos portátiles de bajo costo puedan reproducir en tiempo real los archivos de audio previamente codificados. En este artículo se propone

un nuevo esquema de codificación perceptual basado en técnicas adaptativas, y su rendimiento es cuantitativamente analizado y comparado con esquemas previos usando pistas musicales extraídas del CD SQAM de la Unión Europea de Radiodifusión [6] (EBU por sus siglas en inglés).

Este documento está organizado como sigue: en la sección 2 se describen brevemente las herramientas utilizadas por el método de codificación de audio esbozado en la sección 3. El criterio de selección de los archivos de audio y los resultados de la codificación de los mismos se resumen en las secciones 4 y 5, respectivamente. Finalmente las conclusiones se exponen en la sección 6.

2. Técnicas adaptativas

A diferencia de otros esquemas de codificación, nuestro método se apoya en técnicas de descomposición adaptativa de la señal. Ellas son: 1) descomposición empírica de modos (EMD) [7–9] y 2) descomposición empírica de modos por conjuntos (EEMD) [10]. Estas herramientas permiten descomponer una señal, por más compleja que sea, en un conjunto finito y reducido de funciones AM-FM de banda limitada y media local nula fácilmente representables. A continuación se las describe brevemente.

2.1. Descomposición empírica de modos

En el método EMD, se extraen progresivamente detalles de los datos de entrada, comenzando con los de resolución temporal más fina hasta llegar al de resolución temporal más gruesa, mediante un proceso de tamizado. Intuitivamente, los datos de entrada $x(t)$ pueden verse como una suma de funciones detalle $d_k(t)$ oscilatorias de media local nula. Cada función detalle se extrae como sigue:

1) Se detectan los máximos (mínimos) locales de la señal de entrada y luego se los interpolan entre sí, dando como resultado la envolvente superior (inferior) $e_{\max}(t)$ $[e_{\min}(t)]$.
2) Se calcula la media local $m_1(t) = (e_{\max}(t) + e_{\min}(t))/2$ y se determina el detalle de primer orden según: $d_1(t) = x(t) - m_1(t)$.
3) Se calcula el primer residuo $r_1(t) = x(t) - d_1(t)$ y se lo usa como entrada para volver a repetir los pasos 1) y 2). La señal de salida es el detalle de segundo orden $d_2(t)$, así como el residuo de segundo orden es $r_2(t) = r_1(t) - d_2(t)$.
4) El proceso de tamizado continúa iterando entre los pasos 1) y 3) hasta que el residuo $r_K(t) = r_{K-1}(t) - d_K(t)$ no posee más extremos locales y, por lo tanto no hay más detalles para extraer. En este punto, la señal de entrada fue completamente descompuesta y puede ser reconstruida según:

$$x(t) = \sum_{k=1}^{K} d_k(t) + r_K(t). \tag{1}$$

donde $d_k(t)$ es conocida como función de modo intrínseco o IMF por su sigla en inglés y $r_K(t)$ es el residuo final, también denotado $r(t)$ por simplicidad.

Hay que resaltar que cada función detalle al final del paso 2) puede no tener media nula, en cuyo caso se debe iterar desde el paso 1) para sustraer dicha media. Este proceso de iteración se lleva a cabo hasta que la media correspondiente es suficientemente pequeña [7], [9]. A fin de minimizar la cantidad de iteraciones, en [11] se sugiere un límite máximo de sólo 10 iteraciones. Ese límite es el empleado en el codificador propuesto.

Ya que cada IMF depende únicamente de la señal de entrada, el algoritmo EMD se caracteriza por ser completamente adaptativo y siempre arroja una pequeña cantidad de IMF que pueden ser descriptas como señales AM-FM, es decir, $d_k(t) = a_k(t)\cos[\theta_k(t)]$, donde $a_k(t)$ es la amplitud instantánea y $\frac{1}{2\pi}\frac{\mathrm{d}\theta_k(t)}{\mathrm{d}t}$ es la frecuencia instantánea de la k-ésima IMF $d_k(t)$ [8].

El método EMD se ha utilizado extensivamente para el análisis robusto de datos [8], y también para compresión de datos en 2D [12] y 1D [13], incluyendo señales de audio [14–16]. El presente artículo es una extensión optimizada del método presentado en [16], que fue desarrollado paralelamente a [14]. Cabe destacar que el presente códec contiene varias contribuciones originales que serán detalladas en el transcurso del documento. La más destacable es la inclusión del método EEMD.

2.2. Descomposición empírica de modos por conjuntos

Esta técnica consiste en efectuar múltiples aplicaciones del algoritmo EMD a la señal de entrada contaminada con diferentes realizaciones de ruido blanco gaussiano de potencia finita, con el propósito de añadir contenido espectral en la secuencia de entrada. De este modo, en cada realización EMD trabaja de forma más similar a un banco de filtros diádicos [17,18] y se obtiene un conjunto de IMF más concentradas en determinadas bandas espectrales. (Estrictamente hablando, EMD funciona de esa forma sólo para pequeñas clases de señales de banda ancha, ya que las IMF resultantes no están usualmente concentradas en una octava como se espera. Sin embargo, el método EEMD permite obtener, después de cada realización, un conjunto de IMF mejor concentradas espectralmente que el obtenido via EMD). Finalmente las IMF homólogas son promediadas luego de L realizaciones, dando como resultado un conjunto de IMF promedio $d_k(t)$ $(k = 1, 2, K)$ dado por

$$d_k(t) = \frac{1}{L}\sum_{l=1}^{L} d_{k,l}(t). \tag{2}$$

Un inconveniente de este método es que no satisface íntegramente la ecuación (1) de completitud. Sin embargo, este problema se minimiza reduciendo la potencia del ruido blanco gaussiano añadido [19], especialmente para señales cuyo espectro está más concentrado en bajas frecuencias [10]. Una ventaja importante de EEMD sobre EMD es que cada IMF resultante no contiene información

respecto a dos o más fenómenos físicos diferentes, también conocido en la literatura como *mezcla de modos*. Esta es una consecuencia de la adición de ruido blanco gaussiano, que completa la información espectral en las bandas de menor energía. Esta ventaja es utilizada por el codificador aquí propuesto y es una contribución original respecto a [15, 16].

Finalmente, es importante mencionar que el algoritmo EEMD puede ser optimizado agregando ruido de baja potencia. Esta opción es recomendada en [10] para señales más concentradas en las bajas frecuencias, lo cual se cumple para muchos tipos de señales de audio. Un beneficio adicional para esto es que se requiere un bajo número de realizaciones, lo que permite incrementar la velocidad del algoritmo EEMD.

3. Códec propuesto

El conjunto codificador/decodificador propuesto se explica en las siguientes subsecciones.

3.1. Codificador

La figura 1 muestra el diagrama de bloques del codificador propuesto. El funcionamiento general del mismo se explica a continuación.

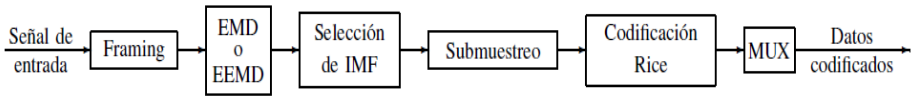


Figura 1. Diagrama de bloques del codificador propuesto.

El codificador toma como entrada un archivo de audio (en este caso en formato WAVE [.wav]) y detecta su tamaño, la tasa de muestreo, la cantidad de canales y el número de bits por muestra. Luego secciona el archivo en cuadros o *frames* de longitud fija (4096 muestras para el presente caso).

A cada frame se le aplica el algoritmo EMD o EEMD, obteniendo como resultado una cantidad determinada de funciones IMF.

Por otra parte, no todas las IMF son representativas de la señal, es decir, algunas aportan muy poca información [20]. Por ello, con el fin de seleccionar las más relevantes, el conjunto obtenido es sometido a un proceso de filtrado compuesto por dos etapas. Esto se ilustra en el diagrama de la figura 2 y constituye uno de los aportes originales del codificador en relación a [14, 16].

La primera etapa de filtrado se apoya en la hipótesis de que las IMF relevantes deben tener una fuerte correlación con la señal analizada, mientras que las IMF irrelevantes poseerán una correlación débil [20]. De este modo, siendo μ_i el coeficiente de correlación entre la i-ésima IMF y la secuencia de entrada, y η un

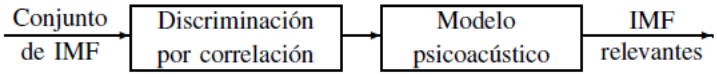


Figura 2. Detección de IMF psicoacústicamente relevantes.

factor de relación (por ejemplo $\eta = 10$), se establece un umbral $\lambda = \max(\mu_i)/\eta$ y se determina que serán conservadas sólo aquellas IMF cuyo coeficiente μ_i sea mayor o igual que λ.

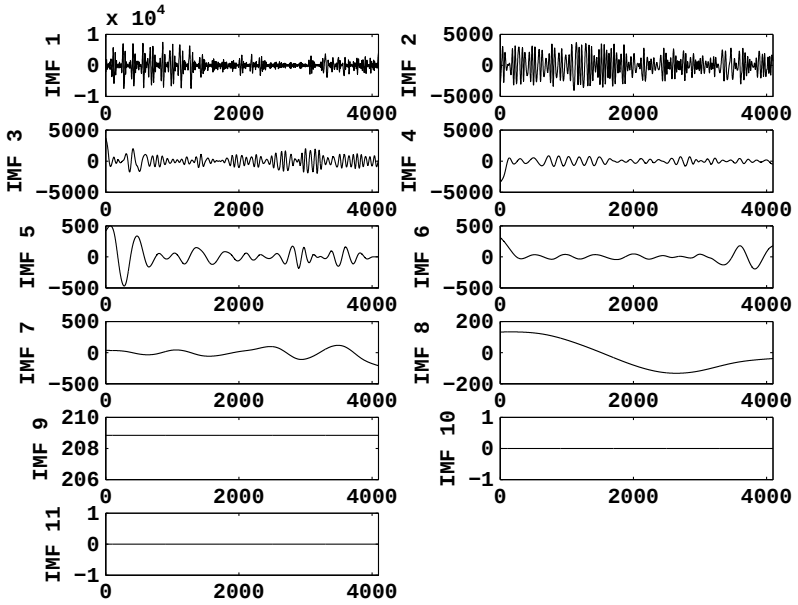
El paso siguiente consiste en aplicar un modelo psicoacústico de enmascaramiento espectral [21], a fin de detectar información que, si bien puede ser representativa a nivel visual, es irrelevante para el oído humano promedio, y por lo tanto su eliminación favorece a la compresión. Para detectar componentes enmascaradas en cada banda crítica se aplica la función de *spreading* ISO/IEC MPEG Psychoaoustic Model 2 (ver [22], pág. 187). Luego, se aplica otra curva de enmascaramiento espectral contemplando el conjunto de todas las bandas críticas de las IMF sobrevivientes (ver [22], pág. 192). En este caso, se suman las intensidades de las curvas de enmascaramiento de cada banda interviniente elevada al factor empírico $\alpha = 0{,}33$.

En la figura 3(a) se muestra el conjunto de IMF asociado a un cuadro de 4096 muestras y en la figura 3(b) se exponen las IMF relevantes luego de aplicar el proceso de selección correspondiente. Como se observa, sólo 4 de las 11 IMF iniciales sobreviven al proceso de discriminación de funciones relevantes. Gracias a este paso, se descartó un gran volumen de información considerada irrelevante, y esto es un gran aporte para la compresión de los datos.

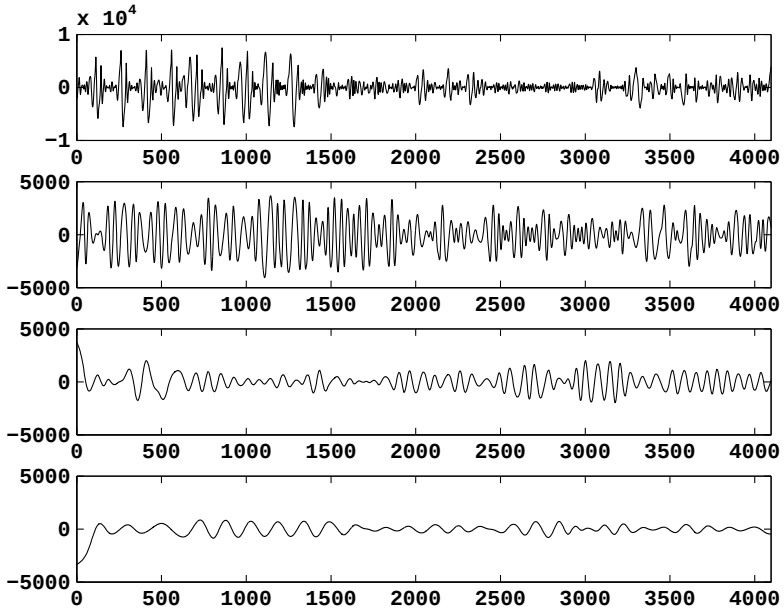
Las IMF resultantes del paso previo son representadas mediante sus correspondientes extremos locales. Debido a que la descomposición EMD es sensible a errores en los bordes, se minimiza ese efecto ignorando la información de las IMF que está hasta un 10 % de cada borde del cuadro actual. Luego, se solapa la información de cada cuadro con sus homólogos adyacentes en un 10 % al principio y al final del mismo. Esta etapa es el bloque de submuestreo de la figura 1 y los resultados del proceso se ilustran en la figura 4.

La interpolación de los extremos locales (por ejemplo via splines cúbicas) permite reconstruir cada IMF con un bajo error [14, 16]. Las abscisas n_i y ordenadas P_i de los extremos locales son codificadas en forma separada. Para las abscisas se deja el primer valor n_1 tal como está y luego se calculan las diferencias entre abscisas adyacentes $\delta_i = n_{i+1} - n_i$. Cabe aclarar que este proceso no se contempló en [15]. Como resultado del mismo, se obtiene un conjunto de números más pequeños formado como sigue.

n_1	δ_1	δ_2	δ_3	...	δ_i	...	δ_n



(a)



(b)

Figura 3. (a) Conjunto de IMF resultante de descomponer una señal de audio via EMD. (b) IMF sobrevivientes al proceso de filtrado.

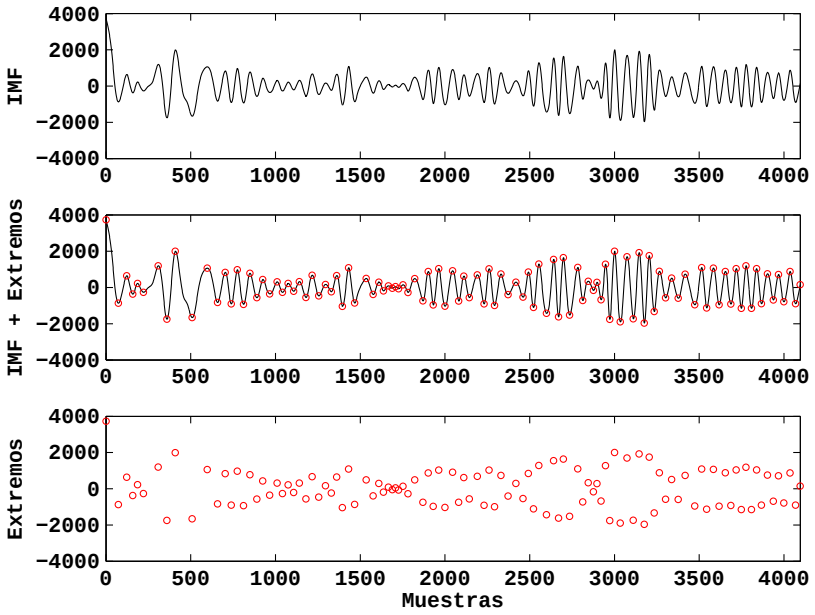


Figura 4. IMF y sus extremos locales para un cuadro específico.

Por el lado de las ordenadas, se aprovecha la propiedad de que los extremos locales adyacentes alternan de signo debido al comportamiento oscilatorio de las IMF. Es decir, si el primer extremo es positivo, el segundo es negativo y así sucesivamente. Por lo tanto en primera instancia determina el signo del primer extremo sgn. Luego, se representan en valor absoluto todas las ordenadas P_i y se les sustrae su mediana m, obteniéndose valores $p_i = |P_i| - m$ más cercanos a cero y conformándose un bloque de datos como el siguiente:

sgn	m	p_1	p_2	p_3	...	p_i	...	p_n

Este modo de representación de ordenadas también es una contribución original del presente codificador. Los datos pertenecientes tanto al bloque de las abscisas como al de las ordenadas son codificados en formato Golomb-Rice [23] y finalmente multiplexados.

3.2. Decodificador

El diagrama de bloques del decodificador se muestra en la figura 5 y funciona de la siguiente forma. El archivo de audio codificado es demultiplexado y luego

tanto el bloque de datos correspondiente a las abscisas, como el correspondiente a las ordenadas son decodificados via Golomb-Rice. Las abscisas son recuperadas a través del siguiente proceso iterativo:

$$\begin{aligned} n_2 &= n_1 + \delta_1, \\ n_3 &= n_2 + \delta_2, \\ &\vdots \\ n_k &= n_{k-1} + \delta_{k-1}. \end{aligned}$$

Las ordenadas son extraídas según las siguientes operaciones:

$$\begin{aligned} P_k &= sgn \times (m + p_k) && \text{si } k \text{ es impar.} \\ P_k &= -sgn \times (m + p_k) && \text{si } k \text{ es par.} \end{aligned}$$

Los extremos locales resultantes se interpolan mediante polinomios de interpolación Hermite cúbico por tramos (o PCHIP por su sigla en inglés). De este modo, se reconstruyen las IMF relevantes, las cuales son posteriormente sumadas para obtener la señal decodificada.

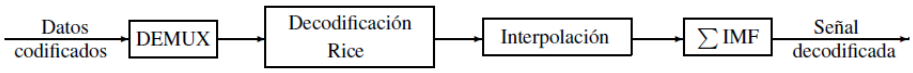


Figura 5. Diagrama de bloques del decodificador propuesto.

Cabe destacar que el codificador y sobre todo el decodificador aquí propuestos son muy simples. La simplicidad del decodificador es crucial para permitir que dispositivos portátiles de bajo costo reproduzcan en tiempo real los archivos de audio codificados.

4. Análisis del rendimiento

El esquema de codificación de audio basado en EMD/EEMD fue evaluado con archivos de audio WAVE extraídos del CD SQAM de la Unión Europea de Radiodifusión (EBU) [6], con el objeto de analizar una variedad de señales de audio de acuerdo a los parámetros recomendados en [24]:

- *Transitorios* (pre-eco sensitivo, manchas de ruido en el dominio temporal),
- *Estructura tonal* (ruido sensitivo, aspereza),
- *Habla natural* (distorsión sensitiva),
- *Sonido complejo* (hace incapié en el dispositivo bajo prueba),
- *Señales de banda ancha* (hace incapié en el dispositivo bajo prueba, pérdida de altas frecuencias, modulación de ruido de alta frecuencia).

Entonces, los archivos de audio seleccionados del CD EBU-SQAM fueron:

- Castañuelas (pista 27),
- Clarinete (pista 16),
- Voz Femenina (pista 49),
- Soprano (pista 44),
- Lira (pista 35).

Cabe destacar que es de interés evaluar sólo un conjunto reducido de archivos de audio con características simples y bien definidas. En futuros trabajos se analizará el desempeño del codec con sonidos más complejos como piezas musicales completas de rock, pop, folklore, tango, etc. Por otra parte, en un futuro se propone contemplar otros parámetros aparte de los aquí contemplados para evaluar el presente codec.

Cada archivo de audio codificado fue evaluado en términos del factor de compresión, que es la relación entre los tamaños de los archivos de entrada y de salida. La calidad de los datos codificados fue medida a través del grado de diferencia objetiva (ODG) [24], utilizando el software disponible en www.hoertech.de → products → downloads. Los resultados se compararon con los obtenidos usando los esquemas de codificación perceptual OGG Vorbis [5] y MP3 [4].

5. Resultados

Los valores numéricos correspondientes al factor de compresión y ODG de los archivos de audio procesados se muestran en la tabla 1. Para el algoritmo EEMD se llevaron a cabo pruebas con 100 realizaciones y ruido blanco gaussiano de diferentes potencias porcentuales σ. Se eligió en forma heurísica el valor de σ que optimiza la compresión para cada archivo de audio analizado. En la tabla 1 se indican los valores que generaron los mejores resultados.

Para cada archivo de audio, la compresión obtenida con EMD supera por poco a la obtenida vía EEMD, aunque con peor fidelidad, dado que el ODG es más negativo (ver tabla 1). Esta mejora en la fidelidad ilustra una ventaja del método EEMD por proveer funciones IMF con mejor concentración espectral, es decir, con menor mezcla de modos.

Finalmente se observa que la fidelidad de los archivos de audio codificados con el método propuesto es superada por los otros codificadores. Este punto está actualmente bajo estudio para ser optimizado.

6. Conclusiones

El esquema de codificación de audio EMD/EEMD fue presentado y evaluado con archivos de audio provistos por la Unión Europea de Radiodifusión y comparado con otro algoritmos de codificación existentes. Este codificador es simple

		EMD	EEMD	MP3 VBR	OGG VBR	MP3 64k
Castañuelas (1)	FC	4,39	4,21	4,56	7,05	**7,95**
	ODG	-2,97	-2,24	-2,16	-2,56	-2,87
Clarinete (1)	FC	**11,48**	7,92	4,33	7,26	7,35
	ODG	-3,12	-2,04	-1,02	-2,13	-2,22
Voz	FC	4,98	4,62	3,51	5,62	**6,24**
Femenina (2)	ODG	-2,99	-2,09	-1,56	-2,15	-2,32
Soprano (2)	FC	**8,67**	7,82	3,82	7,23	7,36
	ODG	-3,11	-2,31	-1,10	-2,27	-2,16
Lira (1)	FC	**12,32**	11,88	2,75	5,94	4,95
	ODG	-3,02	-2,50	-0,97	-2,11	-2,05

Nota 1: $\sigma = 0,01$ en EEMD.
Nota 2: $\sigma = 0,05$ en EEMD.

Tabla 1. Factor de compresión y ODG para diferentes archivos de audio procesados con diferentes algoritmos de codificación.

y provee alta compresión. Nuevas mejoras relacionadas con la fidelidad y la velocidad de procesamiento se encuentran en desarrollo. Una ventaja importante es que se necesita un decodificador muy simple independientemente del método utilizado para codificar (EMD o EEMD). Esto es crucial para ser utilizado en dispositivos portátiles de bajo costo.

Referencias

[1] F. A. Marengo-Rodriguez, E. A. Roveri, J. M. Rodríguez-Guerrero, and M. Treffiló, "Análisis comparativo de codificadores de audio sin pérdidas y una herramienta gráfica para su selección y predicción de su desempeño," *Mecánica Computacional*, vol. 30, no. 41, Acoustics and Mechanical Vibrations (B), pp. 3167–3186, 2011.

[2] xiph.org Foundation, "FLAC - Free lossless audio codec," 2013.

[3] M. Ashland, "Monkey's audio," 2013.

[4] ISO/IEC, "ISO/IEC 11172-3:1993 - Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio," Padrão, 1993.

[5] xiph.org Foundation, "Vorbis I specification," 2012.

[6] EBU, "Sound Quality Assessment Material, Recordings for subjective tests - Users'Handbook for the EBU-SQAM Compact Disc," European Broadcasting Union, Tech. Rep. 3253, 2008.

[7] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. of the Royal Soc. of London*, vol. 454, no. 1971, pp. 903–995, 1998.

[8] S. S. N.E. Huang, *The Hilbert-Huang Transform and Its Applications (Interdisciplinary Mathematical Sciences)*. World Scientific Publishing Company, 2005.

[9] G. Rilling, P. Flandrin, and P. Gonçalves, "On empirical mode decomposition and its algorithms," in *Proc. of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03*, Grado (Italy), 2003.

[10] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[11] N. E. Huang, M.-L. C. Wu, S. R. Long, S. S. Shen, W. Qu, P. Gloersen, and K. L. Fan, "A confidence limit for the empirical mode decomposition and hilbert spectral analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[12] A. Linderhed, "2-D empirical mode decompositions - in the spirit of image compression," in *Proceeding of SPIE,Wavelet and Independent Component Analysis Applications IXI*, vol. 4738, Orlando (USA), 2002, pp. 1–8.

[13] C. Ho, "Empirical mode decomposition based novel data compression algorithm for wireless data transmission in machine health monitoring," Ph.D. dissertation, City University of Hong Kong, 2009.

[14] K. Khaldi, A. Boudraa, M. Turki, I. Samaali, and T. Chonavel, "Audio encoding based on the empirical mode decomposition," in *EUSIPCO 09*, 2009.

[15] K. Khaldi, A. Boudraa, B. Torresani, and T. Chonavel, "HHT - based audio coding," in *Signal, image and video processing, vol. 7, no. 2*, 2013.

[16] F. A. Marengo-Rodriguez and F. Miyara, "Representación de señales de audio con descomposición empírica de modos y submuestreo adaptativo," in *Primeras Jornadas Regionales de Acústica*, no. A056R. Rosario (Argentina): Asociación de Acústicos Argentinos, 2009.

[17] P. Flandrin, G. Rilling, and P. Gonçalves, "Empirical mode decomposition as a filter bank," *IEEE, Signal Processing Letters*, vol. 11, no. 2, pp. 112–114, 2004.

[18] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proc. of the Royal Society of London*, vol. 460, 2004.

[19] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise." in *ICASSP*. IEEE, 2011, pp. 4144–4147.

[20] Z. K. Peng, P. W. Tse, and F. L. Chu, "A comparison study of improved hilbert-huang transform and wavelet transform: Application to fault diagnosis for rolling bearing," *Mechanical Systems and Signal Processing*, vol. 19, no. 5, pp. 974–988, 2005.

[21] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models (Springer Series in Information Sciences)*. Springer, 2007.

[22] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*, ser. Kluwer international series in engineering and computer science. Power electronics and power systems. Springer US, 2003.

[23] D. Salomon, *Data Compression.: The Complete Reference.* Springer, 2004.

[24] ITU-R Recommendation BS 1387-1, *Method for objective measurements of perceived audio quality*, Std., 2001.