

Extracción de F_0 en señales de voz basada en Descomposición Empírica en Modos por Conjuntos Completa con Ruido Adaptativo

M. A. Colominas*

G. Schlotthauer*

M. E. Torres* §

* *Laboratorio de Señales y Dinámicas no Lineales, Universidad Nacional de Entre Ríos*§ *metorres@santafe-conicet.gov.ar*

Abstract— En el presente trabajo se presenta una variación del método de Descomposición Empírica en Modos (EMD). Este método reduce los efectos de mezcla de modos, al estilo de la Descomposición Empírica en Modos por Conjuntos (EEMD), pero con un costo computacional mucho menor y recuperando la propiedad de completitud de EMD. Se introducen aquí resultados preliminares que muestran la capacidad de este nuevo método, llamado Descomposición Empírica en Modos por Conjuntos Completa con Ruido Adaptativo (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise, CEEMDAN) para estimar la frecuencia fundamental en señales de voz normales y patológicas y se comparan con métodos clásicos y el basado en EEMD propuestos anteriormente por dos de los autores, resaltando las ventajas de este nuevo enfoque, en vocales sostenidas.

Keywords— Descomposición Empírica en Modos, Descomposición Empírica en Modos Completa con Ruido Adaptativo, Frecuencia Fundamental

1. INTRODUCCIÓN

El período fundamental T_0 de una señal de voz puede definirse como el tiempo transcurrido entre dos pulsos glóticos sucesivos y la frecuencia fundamental es $F_0 = 1/T_0$ [4]. Siendo F_0 de gran utilidad para un amplio rango de aplicaciones, una estimación confiable de ella se considera aún una tarea difícil. La dificultad aumenta en presencia de ruido o en los casos de voces patológicas. En el habla, las variaciones de F_0 contribuyen a la prosodia, y en idiomas tonales también ayudan a la distinción de categorías segmentales.

Una estimación confiable y precisa de F_0 es esencial para el correcto análisis de las perturbaciones de frecuencia (fenómeno conocido como *jitter*). En el caso de señales de vocales sostenidas, este análisis es un procedimiento habitual en la evaluación clínica de los trastornos de la voz, y en el monitoreo del progreso del paciente durante el tratamiento.

Los algoritmos tradicionales para extraer F_0 toman ventanas de la señal, otorgando en la mayoría de los

casos estimaciones constantes a trozos. Sin embargo, en el análisis de voces patológicas es deseable tener una F_0 suave y precisa. Además, estos métodos suponen como hipótesis que el habla se produce por un sistema lineal y que la señal es localmente estacionaria, dos simplificaciones inapropiadas en el caso de voces patológicas.

Schlotthauer y col. [3] presentaron un nuevo método basado en EEMD para la extracción de F_0 . Trabajaron sobre la hipótesis de que un modo de la descomposición de la señal capturaba la frecuencia fundamental. Para la selección de dicho modo, desarrollaron un algoritmo basado en energías. Evaluaron satisfactoriamente su algoritmo con señales de voz normales y patológicas.

A pesar de presentar la EEMD (*Ensemble Empirical Mode Decomposition*) un gran avance respecto de la EMD (*Empirical Mode Decomposition*), al aliviar notoriamente el problema de la mezcla de modos, posee aún ciertas dificultades [2, 8]. La CEEMDAN (*Complete Ensemble Empirical Mode Decomposition with Adaptive Noise*) resuelve algunas de estas dificultades [5]. El nuevo método resultó satisfactorio cuando se lo evaluó en señales tanto artificiales como reales. En particular en [5] se muestran ejemplos que sugieren que CEEMDAN es un tipo de descomposición que permitiría extraer la frecuencia fundamental de señales de voz y de electrocardiograma (ECG), de manera similar a EEMD, pero con un menor costo computacional. Permite además la reconstrucción exacta de la señal original, como la EMD, característica perdida por el algoritmo de EEMD.

Teniendo en cuenta las dificultades de la EEMD, y el avance que representa la introducción de la CEEMDAN, es que proponemos visitar la experiencia de Schlotthauer y col. y aplicar el nuevo algoritmo de CEEMDAN para la extracción de frecuencia fundamental F_0 en señales de voces normales y patológicas.

2. MATERIALES Y MÉTODOS

En esta sección se describen brevemente el método de EEMD, el nuevo algoritmo de descomposición empírica en modos, el algoritmo basado en energías de selección del modo que captura F_0 , el método para calcular la frecuencia instantánea y las señales con las que se estudió su desempeño.

2.1. Descomposición Empírica en Modos por Conjuntos

La EMD [2] descompone una señal en un número habitualmente pequeño de Funciones de Modo Intrínsecas (*Intrinsic Mode Functions, IMFs*) o, simplemente, modos. Para ser considerada una IMF, una señal debe satisfacer dos condiciones: (i) la cantidad de puntos extremos (máximos y mínimos) y la cantidad de cruces por cero deben ser iguales o diferir a lo sumo en uno; y (ii) la media local, definida como el valor medio de las envolventes superior e inferior, debe ser nula para todo punto.

La EEMD [8] define a los modos «verdaderos» (aquí simbolizados como \overline{IMF}) como el promedio de las correspondientes IMFs obtenidas con EMD sobre un conjunto de señales generadas al sumar diferentes realizaciones de ruido blanco de varianza finita a la señal original (señal de interés) $x[n]$. El algoritmo de EEMD puede describirse de la siguiente manera:

1. generar $x^{(i)}[n] = x[n] + w^{(i)}[n]$, donde $w^{(i)}[n]$ ($i = 1, \dots, I$) es una realización de ruido blanco gaussiano,
2. descomponer completamente cada $x^{(i)}[n]$ ($i = 1, \dots, I$) mediante EMD, obteniendo sus modos $IMF_k^{(i)}[n]$, donde $k = 1, \dots, K$ indica el modo, de la siguiente manera
 - a) inicializar los índices $k = 1, j = 1$, y hacer $r_1^{(i)}[n] = x^{(i)}[n]$,
 - b) encontrar todos los puntos extremos de $r_k^{(i,j)}[n]$,
 - c) interpolar entre los mínimos (máximos), para obtener la envolvente $e_{min}^{(i,j)}[n]$ ($e_{max}^{(i,j)}[n]$),
 - d) calcular la media local $m_k^{(i,j)}[n] = (e_{min}^{(i,j)}[n] + e_{max}^{(i,j)}[n]) / 2$,
 - e) extraer la j -ésima candidata a IMF $d_k^{(i,j)}[n] = r_k^{(i,j)}[n] - m_k^{(i,j)}[n]$,
 - f) verificar si $d_k^{(i,j)}[n]$ es una IMF:
 - si lo es, hacer $r_k^{(i,j+1)}[n] = d_k^{(i,j)}[n]$, $j = j + 1$, e ir al paso ii,
 - si no lo es, hacer $IMF_k^{(i)}[n] = d_k^{(i,j)}[n]$, $k = k + 1, j = 1$ y calcular el residuo $r_k^{(i)}[n] = x^{(i)}[n] - \sum_{l=1}^{k-1} IMF_l^{(i)}[n]$,
 - g) continuar con el proceso hasta alcanzar un criterio de detención predefinido.
3. asignar \overline{IMF}_k como el k -ésimo modo de $x[n]$, obtenido al promediar los correspondientes $IMF_k^{(i)}$: $\overline{IMF}_k[n] = \frac{1}{I} \sum_{i=1}^I IMF_k^{(i)}[n]$.

La extracción de cada $IMF_k^{(i)}$ se realiza mediante un proceso que requiere $j = 1, \dots, J_k^{(i)}$ iteraciones, denominado *sifting*.

2.2. Descomposición Empírica en Modos por Conjuntos Completa con Ruido Adaptativo

Nótese que en la EEMD, cada $x^{(i)}[n]$ se descompone independientemente de las otras realizaciones y para cada una de ellas se obtiene un residuo $r_k^{(i)}[n] = r_{k-1}^{(i)}[n] - IMF_k^{(i)}[n]$.

A continuación, describiremos el método de CEEMDAN recientemente presentado en [5]. En lo que sigue, los modos obtenidos con él se indicarán como \widetilde{IMF}_k .

La idea general del CEEMDAN es la siguiente: dado $x[n]$ se generan las $x^{(i)}[n]$ y se calcula el primer modo $IMF_1[n] = \overline{IMF}_1[n]$ como en EEMD. A continuación se calcula un único primer residuo, independiente de las realizaciones de ruido:

$$r_1[n] = x[n] - \widetilde{IMF}_1[n]. \quad (1)$$

Luego, sobre un conjunto de $r_1[n]$ más diferentes realizaciones de un ruido particular, se calcula para cada una de ellas el primer modo mediante EMD. Se define \widetilde{IMF}_2 como el promedio de estos primeros modos. El próximo residuo está dado por: $r_2[n] = r_1[n] - \widetilde{IMF}_2[n]$. Este procedimiento se repite hasta que se alcanza un criterio de detención. En este trabajo, utilizamos el criterio de Rilling [6]

El siguiente algoritmo detalla el método CEEMDAN aquí propuesto. Definamos el operador $E_j(\cdot)$ que produce el j -ésimo modo obtenido mediante EMD. Sea $w^{(i)}$ una realización de ruido blanco con media cero y varianza unitaria. Entonces:

1. para cada $i : 1, \dots, I$ descomponer cada realización $x[n] + \varepsilon_0 w^{(i)}[n]$ mediante EMD, hasta obtener el primer modo y calcular

$$\widetilde{IMF}_1[n] = \frac{1}{I} \sum_{i=1}^I IMF_1^{(i)}[n] = \overline{IMF}_1[n],$$

2. en la primera etapa ($k = 1$) calcular el primer residuo como en la Ec. (1):

$$r_1[n] = x[n] - \widetilde{IMF}_1[n],$$

3. descomponer $r_1[n] + \varepsilon_1 E_1(w^{(i)}[n])$, $i = 1, \dots, I$, con EMD hasta obtener el primer modo y definir el segundo modo de CEEMDAN como:

$$\widetilde{IMF}_2[n] = \frac{1}{I} \sum_{i=1}^I E_1 \left(r_1[n] + \varepsilon_1 E_1(w^{(i)}[n]) \right),$$

4. para $k = 2, \dots, K$ calcular el k -ésimo residuo:

$$r_k[n] = r_{(k-1)}[n] - \widetilde{IMF}_k[n], \quad (2)$$

5. descomponer $r_k[n] + \varepsilon_k E_k(w^{(i)}[n])$, $i = 1, \dots, I$, con EMD hasta obtener el primer modo y definir

el $(k + 1)$ -ésimo modo de CEEMDAN como:

$$\widetilde{IMF}_{(k+1)}[n] = \frac{1}{I} \sum_{i=1}^I E_1 \left(r_k[n] + \varepsilon_k E_k(w^{(i)}[n]) \right), \quad (3)$$

6. ir al paso 4 para el próximo k .

Iterar los pasos 4 a 6 hasta que el residuo obtenido no admita descomposición con EMD.

Por construcción de CEEMDAN, el residuo final satisface:

$$R[n] = x[n] - \sum_{k=1}^K \widetilde{IMF}_k, \quad (4)$$

siendo K el número total de modos. Por lo tanto, la señal de interés $x[n]$ puede ser expresada como:

$$x[n] = \sum_{k=1}^K \widetilde{IMF}_k + R[n]. \quad (5)$$

Esto garantiza que la descomposición propuesta es completa y proporciona una reconstrucción exacta de los datos originales.

Puede apreciarse que en el caso del método aquí propuesto, los coeficientes ε_i permiten seleccionar la SNR en cada etapa. Usamos aquí cien realizaciones y la misma SNR para todas las etapas. En todas las implementaciones utilizamos la *toolbox* de EMD disponible en: <http://perso.ens-lyon.fr/patrick.flandrin/emd.html>.

2.3. Transformada de Hilbert

Una vez que se obtienen los modos, es necesario un método para separar la amplitud y la frecuencia instantáneas. El uso de la Transformada de Hilbert (TH) está ampliamente aceptado para llevar a cabo esta tarea. Dada una señal $x(t)$, se define su TH como

$$y(t) = \mathcal{H} \{x(t)\} = \frac{1}{\pi} (\text{VP}) \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau, \quad (6)$$

donde VP indica el valor principal de Cauchy. De esta manera podemos definir la señal analítica $z_x(t)$, en la que $x(t)$ constituye la parte real, de la forma

$$z_x(t) = x(t) + i y(t) = A(t)e^{i\theta(t)}, \quad (7)$$

en la cual

$$A(t) = \sqrt{x^2(t) + y^2(t)} \quad (8)$$

y

$$\theta(t) = \arctan \left(\frac{y(t)}{x(t)} \right). \quad (9)$$

Aquí $A(t)$ es la amplitud instantánea y $\theta(t)$ la función de fase. Entonces la frecuencia instantánea (en radianes por unidad de tiempo) es simplemente

$$\omega(t) = \frac{d\theta(t)}{dt}. \quad (10)$$

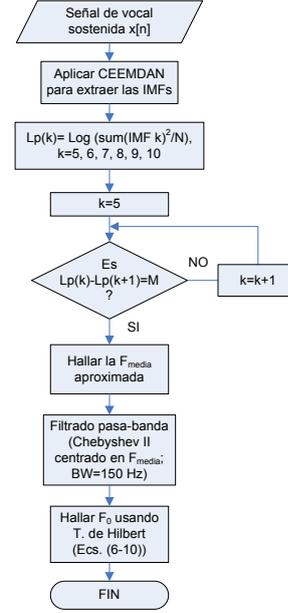


Figura 1: Diagrama de flujos del algoritmo.

2.4. Selección del modo que captura F_0

En los casos en los que no se posee información acerca de la F_0 media de la señal, resulta necesario un método para decidir cuál modo captura F_0 . Schlotthauer y col. presentaron un método basado en energías, el cual fue corroborado estadísticamente en [3]. Al aplicar su método a las descomposiciones EEMD de las señales de vocal sostenida, observaron una notable caída entre la energía del modo que contiene a F_0 y el modo siguiente. De esta forma, se cuenta con un indicador objetivo sobre el modo mediante el cual estimaremos F_0 .

Una vez hallado el modo, con el objetivo de eliminar componentes espurios de frecuencia, se le aplica un filtrado pasa-banda Chebyshev tipo II. La frecuencia central de este filtro corresponde a la frecuencia donde se da el valor máximo en la densidad espectral de potencia. Como se mostró en [3], esta frecuencia es una buena aproximación a la media de F_0 . Seleccionamos un ancho de banda de 150 Hz. A modo de resumen, presentamos en la Fig. 1 un diagrama de flujos con el algoritmo completo, donde el valor para el modo inicial ($k = 5$) se eligió como en [3] y $M = \max [L_p(k) - L_p(k + 1)]$, $k = 5 \dots K - 1$.

2.5. Datos

En el presente trabajo se analizan señales de voces tanto normales como patológicas. Se utilizaron señales de la base de datos desarrollada por el Massachusetts Eye and Ear Infirmary (MEEI)[1], la cual consta de muestras de 710 sujetos. Están incluidas muestras de fonación de vocal /a/ sostenida de pacientes con trastornos orgánicos, neurálgicos, traumáticos y psicogénicos de la voz, así como 53 voces normales.

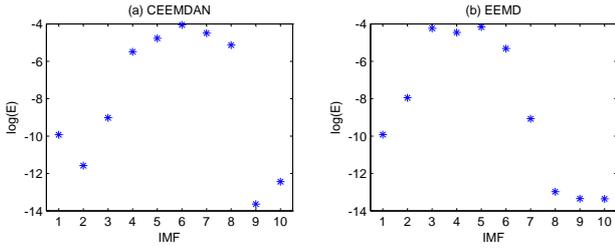


Figura 2: Logaritmo de la energía de los modos 1 a 10 de las descomposiciones de la señal JTH1NAL mediante: (a) CEEMDAN (\widehat{IMF}) y (b) EEMD (\widehat{IMF}). Desvío estándar del ruido agregado $\varepsilon = 0,2$. Cantidad de realizaciones $I = 100$.

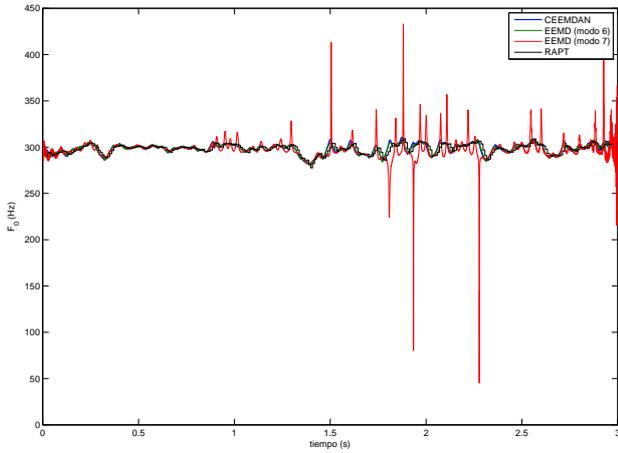


Figura 3: F_0 de la señal normal de voz JTH1NAL.

3. RESULTADOS Y DISCUSIÓN

Como primer ejemplo, aplicamos el algoritmo de la Fig. 1 a la señal de voz normal JTH1NAL, usada también en [3]. En el trabajo citado, se usó una cantidad de realizaciones $I = 5000$. Esto fue necesario, dado que la señal reconstruida a partir de los modos obtenidos por EEMD mantiene un ruido remanente y se debe aumentar la cantidad de realizaciones para reducirlo [8]. Como para la CEEMDAN el ruido remanente es nulo, es posible reducir drásticamente la cantidad de realizaciones I . En este ejemplo usamos $I = 100$ y comparamos con el algoritmo aplicado a una descomposición EEMD con los mismo parámetros. En ambos casos se utilizó $\varepsilon = 0,2$. Para comparar nuestra estimación con los métodos tradicionales, estimamos F_0 mediante el método RAPT (del inglés *Robust Algorithm for Pitch Tracking*), que tiene un excelente desempeño [7]. Implementamos dicho método con la *toolbox* VOICEBOX, disponible en: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.

En la Fig. 2 se muestran las gráficas del logaritmo de la energía de los diez primeros modos, obtenidos mediante CEEMDAN (Fig. 2.a) y EEMD (Fig. 2.b). Mientras que para CEEMDAN se puede apreciar una clara caída de $\log(E)$ luego del modo 8, indicándolo

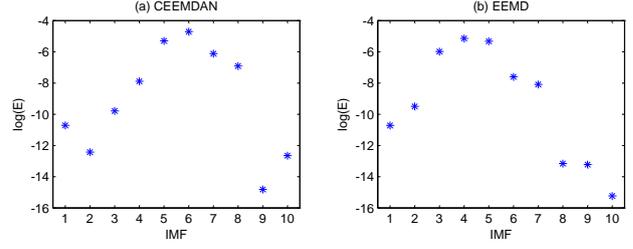


Figura 4: Logaritmo de la energía de los modos 1 a 10 de las descomposiciones de la señal JEG29AN mediante: (a) CEEMDAN (\widehat{IMF}) y (b) EEMD (\widehat{IMF}). Desvío estándar del ruido agregado $\varepsilon = 0,2$. Cantidad de realizaciones $I = 100$.

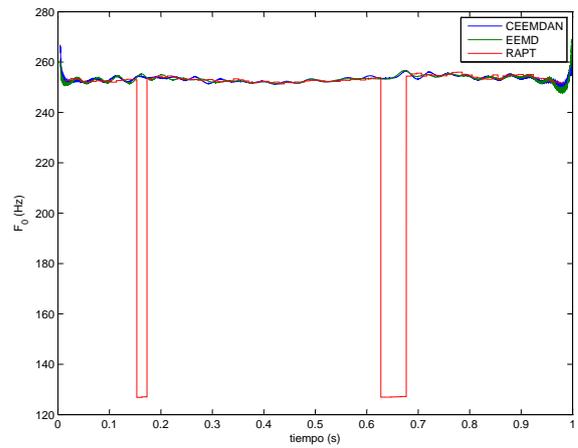


Figura 5: F_0 de la señal de voz patológica JEG29AN.

como portador de F_0 , para EEMD la caída de $\log(E)$ entre los modos 6 y 7 y entre los modos 7 y 8 tienen valores muy similares, con lo que resulta difícil decidir en el cual se encuentra F_0 . Un análisis visual de \widehat{IMF}_6 e \widehat{IMF}_7 revela que ambos poseen considerable energía en la zona del espectro correspondiente a F_0 , evidenciando mezcla de modos.

Puede apreciarse en la Fig. 3 las estimaciones de F_0 con el método RAPT (negro), y con el algoritmo propuesto mediante \widehat{IMF}_6 (verde), \widehat{IMF}_7 (rojo) y \widehat{IMF}_8 (azul). Se observa que la estimación mediante RAPT es satisfactoria, lo que es habitual en voces normales. La estimación a través de \widehat{IMF}_6 es correcta durante el primer cuarto de la señal, pero luego comienza un comportamiento no satisfactorio, con súbitos aumentos y descensos de la frecuencia fundamental, no presentes en la señal original. Las estimaciones a través de \widehat{IMF}_7 y \widehat{IMF}_8 poseen el comportamiento esperado.

Presentamos como segundo ejemplo una señal de voz patológica. La señal JEG29AN, usada también en [3], corresponde a un paciente que padece disfonía espasmódica de aducción. En la Fig. 4 se muestran las gráficas del logaritmo de la energía los diez primeros modos, obtenidos mediante CEEMDAN (Fig. 4.a) y EEMD (Fig. 4.b). En este caso, para ambas descom-

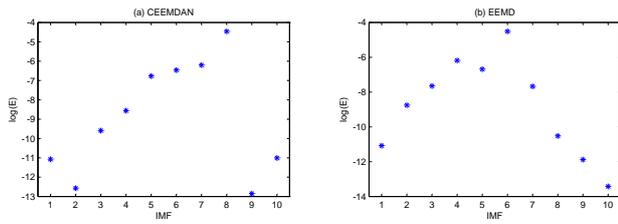


Figura 6: Logaritmo de la energía de los modos 1 a 10 de las descomposiciones de la señal LBA24AN mediante: (a) CEEMDAN (\widetilde{IMF}) y (b) EEMD (\widetilde{IMF}). Desvío estándar del ruido agregado $\varepsilon = 0,2$. Cantidad de realizaciones $I = 100$.

posiciones se ve una clara caída de $\log(E)$ para el modo que captura a F_0 . En el caso de CEEMDAN, esto se da para \widetilde{IMF}_8 , mientras que para EEMD F_0 aparece en \widetilde{IMF}_7 .

En la Fig. 5 se presentan las estimaciones de F_0 de la señal JEG24AN con el método RAPT (rojo), y con el algoritmo propuesto mediante \widetilde{IMF}_7 (verde) y \widetilde{IMF}_8 (azul). Puede verse cómo el método tradicional falla en dos ocasiones, mientras el algoritmo propuesto entrega un estimación suave y precisa para toda la duración de la señal.

Como tercer ejemplo, se analiza la señal patológica LBA24AN. En la Fig. 6 pueden verse las gráficas de los logaritmos de las energías de los diez primeros modos para ambas descomposiciones. En el caso de CEEMDAN, se aprecia con claridad la caída de $\log(E)$ luego del modo 8, señalándolo como portador de F_0 . Por otro lado, los modos de la descomposición mediante EEMD no permiten distinguir claramente cuál de ellos contiene a F_0 ; tal es así que la diferencia entre $\log(E)$ de los modos 6 y 7 es similar a la diferencia entre los modos 7 y 8, evidenciando nuevamente mezcla de modos.

Pueden verse en la Fig. 7 las estimaciones de F_0 de la señal LBA24AN. En negro se muestra la estimación con el método RAPT, en verde y rojo las estimaciones con el algoritmo propuesto mediante \widetilde{IMF}_6 e \widetilde{IMF}_7 respectivamente, y en azul la estimación a través de \widetilde{IMF}_8 . El método RAPT falla en tres ocasiones y la estimación obtenida a través de \widetilde{IMF}_7 muestra variaciones importantes en su amplitud. Como contrapartida, las estimaciones a través de \widetilde{IMF}_6 e \widetilde{IMF}_8 resultan satisfactorias.

4. CONCLUSIONES

En el presente trabajo hemos presentado una variación del método de Descomposición Empírica en Modos (EMD) utilizando ruido adaptivo CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) el que recupera su propiedad de completitud, reduciendo las mezclas de modos al estilo EEMD, pero con una cantidad de realizaciones mucho menor que este último. Se introducen aquí re-

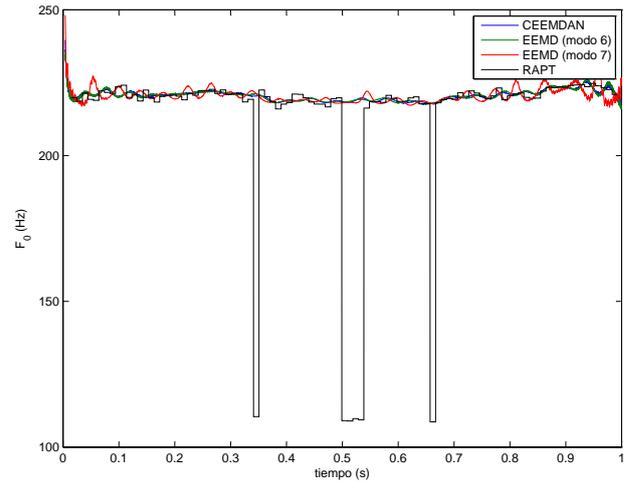


Figura 7: F_0 de la señal de voz patológica LBA24AN.

sultados preliminares que muestran la capacidad del CEEMDAN para estimar la frecuencia fundamental en señales de voz normales y patológicas, a partir de vocales sostenidas.

Los resultados muestran que con tan sólo 100 realizaciones, el algoritmo de estimación de F_0 basado en CEEMDAN permite obtener buenos resultados, representando una ganancia en términos computacionales en algunos casos. Las estimaciones obtenidas mediante los métodos de EEMD y CEEMDAN resultan satisfactorias y comparables a las que se obtienen con el método RAPT. Sin embargo, se observa que el método basado en EEMD puede dar estimaciones erróneas de F_0 tanto para voces normales como patológicas, en aquellos casos en los que persisten las mezclas de modos. Por otra parte, cuando en EEMD se utiliza poca cantidad de realizaciones la reconstrucción por modos mantiene un nivel elevado de ruido.

Como trabajos futuros se prevé realizar un estudio que permita automatizar la selección del modo inicial ($k = 5$ en nuestros ejemplos) y la aplicación del algoritmo a toda la base de datos para obtener resultados estadísticos sobre su desempeño.

REFERENCIAS

- [1] Kay Elemetrics Corp. *Disordered voice database 1.03*. Massachusetts Eye and Ear Infirmary, Voice and Speech Lab, Boston, 1994.
- [2] N.E. Huang et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A*, 454:903–995, 1998.
- [3] Schlotthauer G., Torres M. E., and Rufiner H. A new algorithm for instantaneous F_0 speech extraction based on ensemble empirical mode decomposition. In *17th European Signal Processing Conference 2009*, Glasgow, Scotland, United Kingdom, 2009.

- [4] W.J. Hess. *Springer Handbook of Speech Processing*, chapter Pitch and Voicing Determination of Speech with an Extension Toward Music Signals, pages 181–208. Springer, 2008.
- [5] Torres M.E., Colominas M.A., Schlotthauer G., and Flandrin P. A complete ensemble empirical mode decomposition with adaptive noise. In *Proc. 36th Int. Conf. on Acoustics, Speech and Signal Processing ICASSP (May 22-27, Prague, Czech Republic)*, pages 4144–4147, 2011.
- [6] G. Rilling, P. Flandrin, and P. Gonçalvès. On empirical mode decomposition and its algorithms. In *Proc IEEE-EURASIP Workshop NSIP-03*, Grado, Italia, 2003.
- [7] D. Talkin. *Speech Coding and Synthesis*, chapter A robust algorithm for pitch tracking (RAPT), pages 121–173. Elsevier Science, November 1995.
- [8] Z. Wu and N. E. Huang. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(1):1–41, 2009.