ESTIMACIÓN DE LA FRECUENCIA FUNDAMENTAL DE SEÑALES DE VOZ USANDO TRANSFROMADA WAVELET

RESUMEN
En la estimación de la frecuencia fundamental de señales de voz usando transformada Wavelet es común usar el hecho de que ocurren máximos locales a través de las escalas de descomposición en la vecindad del instante de cierre glótico (*Glottal Closure Instant-GCI*). Dichos métodos se basan en la correlación de las posiciones de los máximos locales para varias escalas de descomposición; pero ello no es tan simple porque existen muchos máximos locales en una señal de voz y, además, las escalas correspondientes a las frecuencias altas son fácilmente afectadas por el ruido. Se propone un método basado en la determinación y correlación de las distancias para cada escala de descomposición, el cual funciona ante perturbaciones de ruido blanco gausiano. Su desempeño se compara respecto a la base de datos *Keele Pitch Database* con el método SIFT(*Simplified Inverse Filtering Tracking*) el cual es un método de estimación de la frecuencia fundamental comúnmente usado en sistemas comerciales.

PALABRAS CLAVES: Frecuencia fundamental, transformada *Wavelet*, selección de la *Wavelet* madre, bioingeniería, voz.

F. ALEXANDER SEPÚLVEDA
Ingeniero Electrónico.
Universidad Nacional de Colombia sede Manizales
franklin@unalmzl.edu.co

GERMÁN CASTELLANOS
Ph.D en telecomunicaciones.
Docente de Planta
Universidad Nacional de Colombia sede Manizales
gcastell@ieee.org

ABSTRACT
We often use the analysis way local maxims, which are present trough the scales of decomposition in the neighbourhood of the Glottal Closure Instant (GCI) for the estimation of the fundamental frequency of speech signal. These methods use the correlation of the local maxima position for various scales of decomposition. This is not simple because there are many local maxims in the speech waveform and, therefore, the scales that correspond to high frequencies are easily affected by noise. A new method is proposed, based on the determination and correlation of distances for each decomposition scale, which works on white noise perturbations. Its achievement is compared respect to the Keele Pitch Database with the Simplified Inverse Filtering Tracking method which is a method commonly used in commercial systems.

KEYWORDS: *Pitch*, *wavelet transform, mother wavelet selection, bioengineering, speech.*

1. INTRODUCCIÓN

El *pitch* o *frecuencia fundamental* F_0 se determina por la velocidad de apertura o cierre de las cuerdas vocales en la laringe durante la fonación de sonidos del tipo sonoro. La estimación de la frecuencia fundamental usando Transformada *Wavelet* (*Wavelet Transform-WT*) ha sido un tema de interés en los últimos años [4],[12],[6],[14]. Su estimación es importante en aplicaciones como la codificación de voz, el desarrollo de sistemas de ayuda a discapacitados (entrenamiento de sordos) [6]. El *pitch* se emplea en la determinación de la entonación y las características emocionales de la voz. Así mismo, sus desviaciones pueden indicar la presencia de desórdenes funcionales y patologías [2].

Las técnicas más comunes para la determinación de la frecuencia fundamental se basan en la propiedad que tiene la WT de producir máximos locales en puntos de singularidad de la señal, tal como se hizo en [8]. Otras técnicas para la estimación de la frecuencia fundamental se basan en el cálculo de la Transformada Wavelet Continua (*Continuos Wavelet Transform-CWT*) usando la función Morlet a modo de Wavelet madre. La frecuencia fundamental aparecerá entonces como una línea horizontal en la representación tiempo-escala. Debido a la *localización limitada* de la WT en el dominio de la frecuencia; la frecuencia fundamental y los formantes aparecerán a modo de bandas esparcidas [1]. En [7],[6], se aprovecha la propiedad de buena resolución en tiempo-frecuencia de la WT para la localización de cambios abruptos que ocurren en los instantes de cierre glótico (*Glottal Closure Instant-GCI*).

En [14] se reporta éxito en la determinación de los GCI, el cual se basa en un algoritmo de programación dinámica. Los algoritmos mostrados en [14] y [12]

Fecha de recepción: 12 Abril de2004
Fecha de aceptación: 16 Abril de 2004

construyen trayectorias de máxima amplitud a través de las escalas de descomposición de la WT, para luego determinar cuales de ellas corresponden a GCI. Aquellas escalas asociadas a las frecuencias altas poseen mayor resolución en el tiempo, por tal motivo se usan aquellos máximos que pertenecen a trayectorias GCI para la determinar de manera más exacta el GCI. Pero debido al hecho de que las escalas asociadas a las altas frecuencias sufren más los efectos del ruido aleatorio, (ver fig. 1), es de esperarse que los algoritmos mostrados en [14] y [12] bajen su rendimiento ante el ruido.

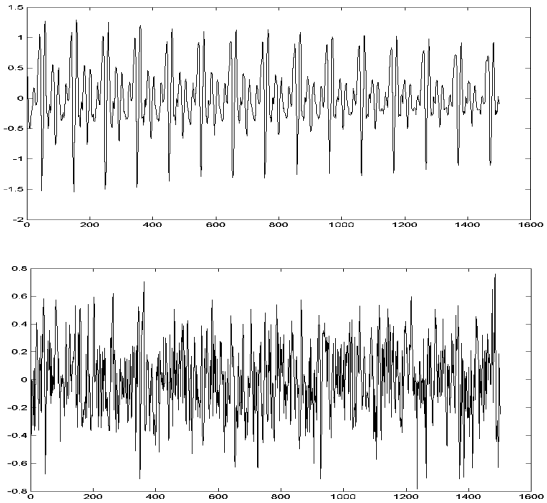

Figura. 1. Escala de mejor resolución en el tiempo, para una señal con ruido y sin ruido respectivamente.

El objetivo del presente trabajo es obtener un método efectivo basado en WT para la estimación de la frecuencia fundamental, que sea sencillo, efectivo y poco sensible a condiciones de ruido [1].

2. ESTIMACIÓN DEL PITCH

2.1 Transformada *Wavelet* Diádica

Existen varios tipos de Transformadas *Wavelet*. Para propósitos de estimación del *pitch* es suficiente tomar sólo unas cuantas escalas de la transformada *Wavelet* continua. Sin embargo, en el cálculo computacional, es preferible el empleo de algoritmos rápidos, en el presente trabajo se usó el algoritmo a huecos o Transformada Wavelet Diádica (DyWT, *Dyadic Wavelet Transform*) [10], el cual está basado en bancos de filtros. El algoritmo se ilustra en la figura 2.

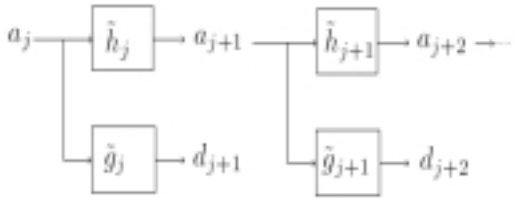

Figura. 2. DyWT: Algortmo a huecos *(Algoritme à Trous)*

En el desarrollo de la DyWT, la escala es muestreada a lo largo de secuencias diádicas de $\{2^j\}$, donde $j \in \mathbf{Z}$, para hacer más rápidos los cálculos numéricos. La DyWT de $f \in L^2$ está definida por:

$$Wf(u,2^j) = \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{2}}\psi\left(\frac{t-u}{2^j}\right)dt = f * \overline{\psi}_{2^j}(u) \qquad (1)$$

siendo

$$\overline{\psi}_{2^j}(u) = \psi_{2^j}(-t) = \frac{1}{\sqrt{2}}\psi\left(\frac{-t}{2^j}\right) \qquad (2)$$

El marco consiste de dilataciones diádicas y traslaciones de la función madre,

$$\psi_{j,n}(t) = \psi(2^j t - 2^{j-J} n) \qquad (3)$$

2.2 Selección de la *Wavelet* madre

En [10] se demuestra que si f es regular y ψ tiene una cantidad p suficiente de momentos de desvanecimiento, definidos por la condición:

$$\int_{-\infty}^{\infty} t^k \psi(t)dt = 0 \qquad para \quad 0 \le k \le p \qquad (4)$$

entonces, los respectivos coeficientes *Wavelet* $\langle f, \psi_{j,n} \rangle$ serán pequeños a escalas finas de 2^j.

Si f presenta una singularidad aislada en el momento t_0, el cual se encuentra dentro del soporte compacto de $\psi_{j,n}(t)$, entonces, el producto $\langle f, \psi_{j,n} \rangle$, en forma general, le corresponderá una amplitud grande. Si ψ tiene un soporte compacto de longitud K, en cada escala 2^j existirán K funciones *Wavelet* ψ, cuyo soporte incluye a t_0. Para minimizar el número de coeficientes de amplitud alta se debe reducir el tamaño del soporte de ψ [10].

En el caso en que f tenga pocas singularidades aisladas y sea suficientemente suave entre dichas singularidades, es preferible el uso de *Wavelets* madre con bastantes momentos de desvanecimiento, con el fin de obtener la mayor cantidad de coeficientes $\langle f, \psi_{j,n} \rangle$ de valor cercano a cero. En cambio, si la densidad de singularidades por unidad de tiempo se incrementa, sería recomendable reducir el tamaño del soporte, aunque su costo sería la reducción de los momentos de desvanecimiento.

[1] Este trabajo es financiado por COLCIENCIAS según contrato 11191412867.

Respecto al compromiso entre una menor longitud del soporte compacto contra una mayor cantidad de los momentos de desvanecimiento, las *Wavelets* que mejor se desempeñan son las pertenecientes a la familia *Daubechies* y las del tipo S*pline* [10].

En [9] se compara el desempeño relativo de las *Wavelets* de fase lineal y las *Wavelets* de fase mínima para la detección de eventos del pitch usando un algoritmo de detección de eventos basado en la DyWT. Empleada para detectar el cierre glótico. En dicho trabajo se reporta que las *Wavelets* de la familia *Spline* entregan mejores resultados. Usando tal resultado en [8] los autores crean un algoritmo para la determinación de la frecuencia fundamental, para probar dicho algoritmo se usaron señales sintéticas. Al llevarlo a la práctica su desempeño desmejora notoriamente, y aún más ante condiciones de ruido [6].

Dentro de la familia de las *Spline*, en calidad de la mejor *Wavelet* madre se escoge aquella que entregue la mayor cantidad de coeficientes cercanos a cero de tal forma, que sólo unos pocos sean de valor grande respecto de los demás. Con tal propósito, se usa la medida de variabilidad de la energía propuesta en [3], tomando el valor resultante al aplicar la función de entropía de Shannon calculada en cada escala de descomposición. La suma total de estos valores, por todas las escalas de descomposición, corresponde a la función de costo final. De tal manera, que la selección final de la *Wavelet* madre recae sobre la función que tenga el menor valor de la función de costo, estimada de la forma:

$$C = \min_{k} \sum_{\lambda=1}^{J} C_{k,\lambda} \tag{5}$$

donde k corresponde a la k-ésima ondita madre a probar, y $C_{k,.}$ está dado por

$$C_{k,\lambda} = -\sum_{m=1}^{N} \frac{\left|\langle f, \psi_{m,\lambda} \rangle\right|^2}{\left\|B^{\lambda}\right\|^2} \log_e \frac{\left|\langle f, \psi_{m,\lambda} \rangle\right|^2}{\left\|B^{\lambda}\right\|^2} \tag{6}$$

2.2 Algoritmo de estimación del *pitch*

Para el algoritmo basado en la correlación de distancias, se puso especial empeño en la selección de los máximos locales ya que tal etapa es vital para dicho algoritmo.

1. *Determinación de los máximos locales*: Es la primera etapa del sistema y consiste en determinar el máximo valor para cada ventana de análisis. El tamaño de la ventana N_m se escoge asumiendo que la mayor frecuencia del *pitch* que se pueda encontrar será de 500 Hz, lo cual nos da un mínimo período de *pitch*, con lo que se garantiza que para cada ventana existirá a lo sumo un máximo que corresponda a un GCI.
2. *Corrección de los máximos locales*: Se desarrolla en dos etapas:
   - Si se encuentra que dos máximos locales están separados por una distancia menor a N_m, entonces se descarta el menor de ellos.
   - En éste trabajo se encontró que generalmente entre dos máximos que corresponden a GCI se encuentra un tercer máximo, pero de menor altura respecto a la altura de sus vecinos. Particularmente se descartaron aquellos máximos cuya altura era menor a un determinado porcentaje F del valor medio de sus vecinos.
3. *Distancia entre máximos*: El paso siguiente consiste en determinar las distancias entre los máximos para cada escala. Debido a que máximos locales erróneos no pueden eliminarse en un 100%, entonces se aplica un filtro mediana de tamaño N_f para eliminar aquellos valores anómalos del vector de distancias. Se toman J escalas de descomposición.
4. *Selección de escalas*: de los J vectores de distancias se toman aquellos p vectores de distancias que poseen la menor desviación estándar respecto a su valor medio. De esa forma se logra que el sistema sea robusto ante el ruido. Cuando el ruido existe, los máximos locales se desordenan en aquellas bandas que corresponden a frecuencias mayores, lo que hace es no tener en cuenta dichas escalas.
5. *Promediado*: el resultado final se obtiene al promediar los valores de estimación del *pitch* entregados para cada escala.

3. DISEÑO EXPERIMENTAL

Para el valor de F (valor usado para la corrección de los máximos locales en el paso 2 del algoritmo de estimación del *pitch*) se escogió de tal forma que fuese el 80%. El tamaño de filtro mediana escogido fue de 3, empíricamente se observó que dicho orden brindaba buenos resultados.

¿Como puede llevarse a cabo la evaluación de un sistema de estimación de la frecuencia fundamental ? Uno de lo métodos más utilizados para ello es la comparación del nuevo de estimación con otro método de calidad ya comprobada. El detector basado en el *cepstrum* se ha utilizado ampliamente para tal fin [6]. Otra posible solución es la utilización de instrumentos de determinación del período del *pitch*, pero éste camino no siempre resulta sencillo debido a la resistencia de los locutores a usar dichos instrumentos aparte de las modificaciones que pudiesen ser introducidas por dichos instrumentos en la forma de hablar. Una tercera solución es la determinación del período del *pitch* a partir de la forma de onda de la señal entregada por el laringógrafo, dichas señales conforman nuestra base de datos para el establecimiento de una referencia. A continuación se usó un sistema semi-automático (su resultado se revisa manualmente) para la marcación de los instantes de cierre

glótico. Para usar dicha información en la evaluación de sistemas de estimación del período del *pitch*, a cada instante de cierre glótico se le asocia un valor del periodo del *pitch* equivalente a la distancia entre dicho instante y el anterior [6]. A partir de dichas muestras se genera el contorno del *pitch* de referencia.

- *Sujetos*: Los datos usados para el diseño experimental provienen de la base de datos *Keele Pitch Database*, del *Centre of Cognitive Neuroscience, The University of Liverpool*. Los datos corresponden a las salidas del Laringógrafo y de voz tomadas simultáneamente para un texto balanceado fonéticamente, el cual fue leído por 10 hablantes, 5 mujeres (f1,...,f5) y 5 hombres (m1,…,m5) [13].
- En dicha base de datos, tanto la señal de voz como la del laringógrafo fueron tomadas a una frecuencia de muestreo de 22 *kHz*. De dicha base de datos se usó un algoritmo de clasificación sonora/sorda para extraer las porciones señal a la cual se le extraería la frecuencia fundamental. El algoritmo de segmentación es una combinación el método usado en [2] y el método usado en [15]. A la señal proveniente del laringógrafo se le aplicó un algoritmo de marcación de instantes de *GCI*, se tomaron aquellos segmentos para los cuales se detectaron los *GCI* en un 100%. A modo de segunda prueba, a los segmentos se les agregó ruido blanco gaussiano.

- Para evaluar la potencialidad del algoritmo propuesto se usa el esquema utilizado en [2] para la estimación de la frecuencia fundamental; el cual corresponde a una variante del algoritmo *SIFT (Simplified Inverse Filtering racking)*. El SIFT busca la periodicidad de una señal estimada por filtrado inverso, y es uno de los métodos más comúnmente usados en equipos comerciales [6]. Los métodos de extracción del período fundamental por filtrado inverso tiene sus ventajas sobre los métodos basados en la autocorrelación y análisis *cepstral*. Para la comparación se usa la medida del error cuadrático medio. Para poder realizar la comparación mediante el error cuadrático medio se usa la interpolación por *splines,* de la cual se aprovecha la propiedad de que pasa por todos los puntos que se desean interpolar.

3. RESULTADOS

En la figura 4 aparecen los contornos para una de los segmentos de la hablante 1(f1), de la base de datos [15]. Las señales de voz se muestran en la figura 3. Se puede apreciar que el algoritmo funciona ante condiciones de ruido.

En la tabla 1 se presentan las medidas de error cuadrático medio para cada hablante, respecto a la señal de referencia, los contornos están en unidades de Hertz. Aparece el error cuadrático medio para los hablantes, a cuyas señales adicionalmente se les agregó ruido blanco gaussiano de varianza equivalente al 50% de la varianza de la señal de voz.

4. CONCLUSIONES

Se presentó un algoritmo de estimación de la frecuencia fundamental usando WT, basado en la capacidad que posee dicha herramienta de entregar máximos locales ante singularidades producidas por los GCI en señales de voz del tipo sonoro. En general presentó mejor comportamiento que el método usado en [2], el cual corresponde a una variante del SIFT. El algoritmo funciona ante condiciones de ruido blanco gaussiano respecto a las bases de datos usadas, además no reportó errores dobles ni errores mitad.

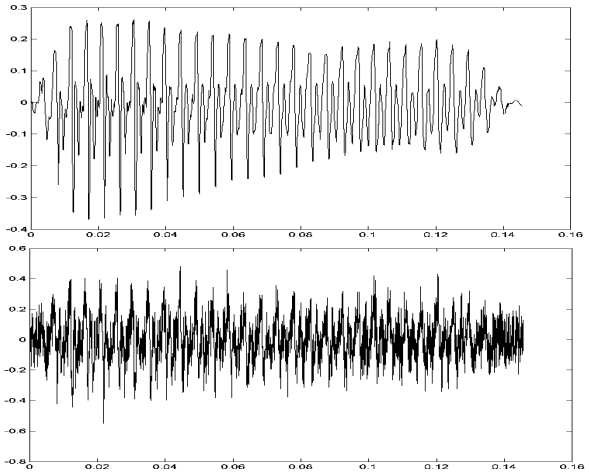

Fig. 3. Porción de señal de voz del tipo sonoro perteneciente al hablante f1. Sin ruido y con ruido respectivamente.

	0%		50%	
	STFT	Wavelet	STFT	Wavelet
f1	769.260,00	839.760,00	2.301.800,00	3.439.700,00
f2	978,61	243,88	1.670,80	472,94
f3	718,44	104,40	1.361,00	353,55
f4	1.136,70	132,48	2.870,60	1.070,50
f5	57,81	72,99	101,37	353,30
m1	10.100,00	1.180,00	15.925,00	2.213,70
m2	1.338,00	120,62	8.749,70	202,44
m3	3.350,40	63,05	5.499,20	228,48
m4	12.548,00	27,35	10.577,00	96,40
m5	6.365,60	33,35	10.558,00	180,88

Tabla 1. Medida de error cuadrático medio para los contornos medidos en Hz.

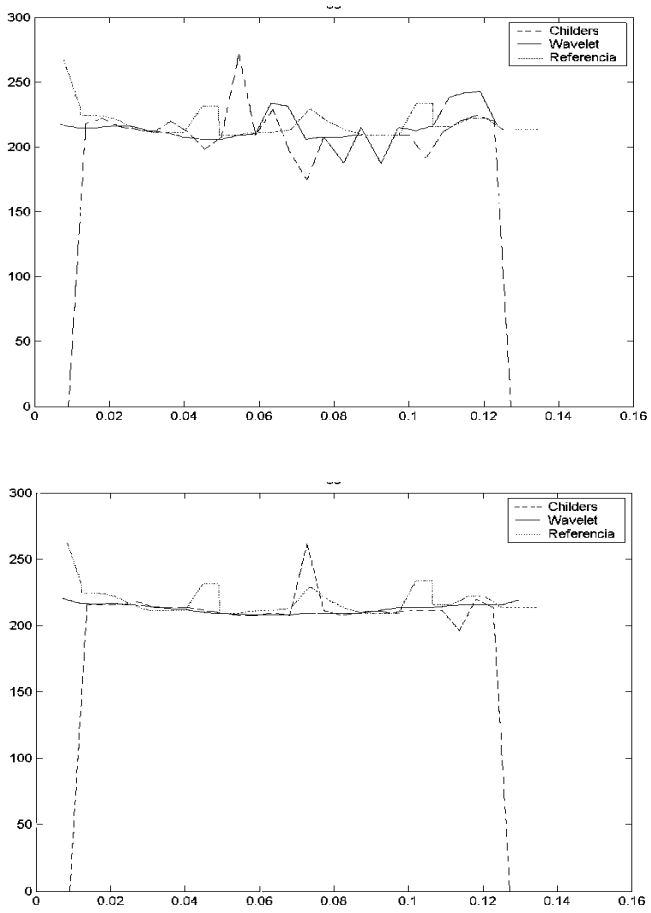

Fig. 4. Contornos de voz de los métodos SIFT y Wavelet, respecto a la referencia.

5. BIBLIOGRAFÍA

[1] A. Bultheel, *Wavelets with applications in signal- and image processing*.http://www.cs.kuleuven.ac.be/ade/WWW/WAVE/contents.html, 2001.

[2] D. Childers, *Speech Processing and Synthesis Toolboxes*, R. Factor, Ed. John Wiley and Sons, 2000.

[3] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Info. Theory*, vol. 38, no. 2, 1992.

[4] N. Gonzáles and D. Docampo, "Application of singularity detection with wavelets for pitch estimation of speech signals," in *EUROSPEECH94*, 1994.

[5] C. Herley. *Digital Signal Processing Handbook*, chapter Wavelets and Filter Banks. Chapman and Hall/ CRCnetBASE, 1999.

[6] L. Janer , "Transformada wavelet aplicada a la extracción de información en señaales de voz," Ph.D. dissertation, Univesitat Politecnica de Catalunya, 1998.

[7] L. Janer, "Modulated gaussian wavelet transform based speech analyser(mgwtsa) pitch detection algorithm (pda)," in *EUROSPEECH*, 1995.

[8] S. Kadambe and G. F. Boundreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Trans. on Info.Theory*, vol. 38, no. 2, 1992.

[9] S. Kadambe and G. Bourdeaux-Bartels, "A comparison of a wavelet functions for pitch detection of speech signals," *International Conference on Acoustics, Speech, and Signal Processing*, 1991.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[11] A. Mojsilovic, M. V. Popovic. On the selection of an optimal wavelet basis for texture characterization. *IEEE Transactions on Image Processing*, 9(12), December 2000.

[12] V. Ngoc and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Eurospeech99*, 1999, pp. 2805–2808.

[13] F. Plante., G. Meyer and W. Ainsworth, "A pitch extraction reference database," in *Eurospeech95*, http://www.liv.ac.uk/Psychology/HMP/projects/pitch .html, 1995.

[14] M. Sakamoto and T. Saitoh, "An automatic pitch-marking method using wavelet transform," in *Proc. of ICSLP2000*, vol. 3, Oct. 2000.

[15] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-97*, Vol. 2, pp. 1323-1326, Apr. 1997.

[16] J. Stegmann and K. A. Fischer. Robust classification of speech based on the dyadic wavelet transform with application to celp coding. In *ICASSP 96*, pages 546–549, 1996.

[17] C.Wendt and A. Petropulu. Pitch determination and speech segmentation using the discrete wavelet transform. *IEEE International Symposium on Circuits and Systems*, 2:45–48, 1996.

[18] M. Wickerhauser. *Adapted Wavelet Analysis: From Theory to Software*. IEEE Press, 1994.