Segmentación de Habla a Nivel de Fonemas usando Wavelets Gaussianas y Análisis de Energía en Subbandas.

Cristian-Remington Juárez-Murillo[1], Sergio Suárez-Guerra[1], José-Luis Oropeza-Rodríguez[1]

[1] Laboratorio de Procesamiento Digital de Señales, Centro de Investigación en Computación, Instituto Politécnico Nacional, Ciudad de México, México.
cristianremingtonjm@hotmail.com, {ssuarez, joropeza}@cic.ipn.mx

Resumen. Se describe el uso combinado de técnicas para el análisis de energía en bandas de frecuencia y la transformada wavelet continua (en adelante llamada CWT) usando wavelets gaussianas moduladas para la segmentación acústica de habla a nivel de fonema.

Palabras Clave: Análisis de habla, segmentación de habla, transformada wavelet continua, wavelet.

1 Introducción

Los archivos de un corpus de voz deben ser segmentados y etiquetados. Segmentación es dividir las locuciones en segmentos de tiempo, lo cual se puede hacer en distintos niveles (ya sea a nivel de fonema, difonema, trifonema, PLUs-del inglés *phoneme like units*, sílabas, *dyads*, etc. [11], [12]).

Se han sugerido diversos enfoques para atacar el problema de la segmentación automática del habla, por ejemplo, el uso de la técnica DTW (del inglés *dynamic time warping*) [14], algoritmos evolutivos [10], el uso de fractales [3], [6], técnicas difusas [2], redes neuronales artificiales [4], alineación forzada [4] y técnicas basadas en wavelets [1], [5], [9], [15], [16].

En este trabajo se propone una combinación que no se había intentado, de técnicas para segmentación de habla tomando en cuenta aspectos perceptuales. Se proponen modificaciones a los algoritmos para que no sacrifiquen resolución en tiempo como en las versiones originales. Para cada vecindario, se encontró el valor del parámetro que minimiza el error, obteniéndose errores más pequeños que en la versión original, con la ventaja de que en esta propuesta no se toleran borrados de segmentos.

2 Metodología

Este trabajo está basado en la técnica de detección de eventos propuesta por J. Galka y M. Ziolko en [5]. Se realizaron modificaciones al algoritmo original para tomar en

M.A. Cruz-Chávez (Ed): CICos 2011, ISBN. 978-607-00-5091-6. pp. 252 – 264, 2011

cuenta propiedades del sistema auditivo humano. En [8], Janer García propuso las wavelets gaussianas moduladas en la escala auditiva de Bark para analizar el habla. La escala de Bark divide el rango audible de frecuencias en bandas críticas tomando en cuenta el fenómeno de enmascaramiento [17].

Se implementaron las siguientes modificaciones al algoritmo propuesto en [5]:

- Se cambió el tipo de transformada wavelet. Originalmente se usaba la transformada discreta wavelet (DWT, del inglés *discrete wavelet transform*). Esta vez se usa una transformada wavelet continua (CWT).
- Se cambió el tipo de wavelet utilizada, de symlet-12 a wavelets gaussianas moduladas en la escala de Bark.
- Se elimina el paso llamado de "discretización de tiempo", ya que las subbandas producidas por la CWT tienen todas, la misma cantidad de muestras, al ser una transformada no decimada.
- La función de umbral se modifica para ser:

$$f_{tr}(k) = \alpha \sum_{n=-N}^{N} f(k-n) \tag{1}$$

donde $f(k)$ es la función de detección de eventos [5]. Las razones para implementar dichas modificaciones son: las wavelets gaussianas moduladas en la escala de Bark imitan las propiedades del sistema auditivo humano; la eliminación del paso de "discretización de tiempo" permite que no se sacrifique resolución temporal; y finalmente, la modificación de (1) facilita la búsqueda del mejor valor de α.

2.1 Corpus

Se desarrolló un corpus en español mexicano, el cual contiene todos los fonemas hablados en México. El corpus contiene 150 palabras repetidas 4 veces por dos hablantes. El formato utilizado es WAV, con 12 bits por muestra y una frecuencia de muestreo de 16 KHz. Se utilizó este corpus como los archivos fuente de habla para los experimentos.

2.2 Algoritmo propuesto

Los pasos del algoritmo propuesto son:

1. Aplicar un filtro de preénfasis a la señal de habla. El coeficiente del filtro es de 0.95.
2. Normalizar las muestras de la señal de voz.
3. Aplicar la CWT utilizando las wavelets gaussianas moduladas en la escala de Bark.
4. Crear el mapa de importancia, el cual es una matriz, donde cada columna representa una muestra de la señal de voz, y los renglones representan a las subbandas de frecuencia, ordenadas de mayor contenido de energía a

menor, para cada instante de tiempo [5].

5. Aplicar la función de detección de eventos, esto es, detectar cambios significativos en el ordenamiento de las subbandas contenidas en el mapa de importancia [5].
6. Calcular la función de umbral con la ecuación (1), utilizando ciertos valores para α y N.
7. Proponer como fronteras entre fonemas aquellos puntos donde la magnitud de un evento sea mayor que la magnitud de su umbral correspondiente.

2.3 Evaluación de la segmentación

La evaluación de la calidad de segmentación puede realizarse a través de ciertas mediciones, algunas de las cuales son las siguientes:

- Número de segmentos propuestos en la segmentación automática, como se muestra en [16], y se denota por n_a.
- Número de segmentos en la segmentación manual [16], la cual se denota por n_h.
- Tasa de inserción relativa [5], definida por:

$$\lambda = \frac{número\ de\ inserciones}{número\ de\ segmentos\ automáticos} \tag{2}$$

- Tasa de borrado relativa [5], definida por:

$$\mu = \frac{número\ de\ borrados}{número\ de\ segmentos\ de\ referencia} \tag{3}$$

- Error en el número de segmentos [16], definido por:

$$\epsilon_n(w) = \frac{|n_a - n_h|}{n_h} \tag{4}$$

donde w es para una palabra dada.

- Error en la posición de fronteras entre segmentos [16], definida por:

$$\epsilon_p(w) = \sum_j min_i |p_j - q_i| \tag{5}$$

donde p_j es la posición de la j-ésima frontera en la segmentación automática y q_i es la posición de la i-ésima frontera en la segmentación manual.

Sería deseable contar con una medición que refleje tanto el error en el número de segmentos como el error en las posiciones de las fronteras de dichos segmentos. En [16], se propone la siguiente medición:

$$\epsilon(w) = \frac{1}{n_w} \sum_w 5\epsilon_n(w) + \epsilon_p(w) \tag{6}$$

El problema es que numéricamente, ϵ_n tiene valores menores que ϵ_p, debido a que ϵ_p

no toma en cuenta el número de fronteras en la segmentación manual, como sí lo hace ϵ_n. Para evitar esto, se proponen las siguientes mediciones:

$$\epsilon_p{}'(w) = \frac{\epsilon_p(w)}{n_h} \tag{7}$$

y

$$Total\_\epsilon(w) = \sqrt{\epsilon_n(w)^2 + \epsilon_p(w)'^2} \tag{8}$$

3 Diseño de experimentos

Se ejecutaron dos tipos de experimentos. Los experimentos de tipo 1 son para encontrar los valores de α y N en la ecuación (1) tales que los errores de segmentación alcancen un mínimo. Los experimentos de tipo 2 prueban los valores encontrados de α y N en el resto del corpus.

3.1 Experimentos de tipo 1

En la ecuación (1), N representa una cantidad de muestras de la función de eventos. Dicha cantidad de muestras determina un vecindario, denotado por *t_min* (ver [5]). Esto significa que la ecuación (1), o función de umbral va calculando su valor para cada vecindario en la función de eventos, esto es, se calcula de manera adaptativa. La relación entre las cantidades *t_min* y N se calcula con:

$$N = ceil\left(\frac{t\_min}{2}\, fs\right) \tag{9}$$

donde *fs* es la frecuencia de muestreo, en nuestro caso 16 KHz.

De acuerdo a Reddy [13], los fonetistas coinciden aproximadamente un 90% de las veces en la posición de las fronteras entre segmentos colocados manualmente cuando se permite un rango de tolerancia de 20 ms. Debido a este fenómeno se escogieron valores de N para que sean equivalentes a 7 ms, 10 ms y 15 ms, los cuales implican vecindarios de 14 ms, 20 ms y 30 ms.

Los valores escogidos para probar al parámetro α son: 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13 y 0.14. Valores más pequeños que los presentados simplemente producían demasiada sobre segmentación, valores más grandes producían sub segmentación, esto es, algunos segmentos no se detectaban.

En este trabajo se considera importante evitar la sub segmentación.

Se escogió un subconjunto de diez palabras, escogidas del corpus desarrollado, de tal forma que dicho subconjunto aún contuviera los 22 fonemas hablados en México. Esas diez palabras se utilizaron para ejecutar los siguientes experimentos:

- Para N que implica 7 ms, los valores para α son: 0.04, 0.05, 0.06, 0.07, 0.08, 0.09. 0.1, 0.11, 0.12, 0.13 y 0.14.
- Para N que implica 10 ms, los valores para α son: 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12 y 0.13.
- Para N que implica 15 ms, los valores para α son: 0.04, 0.05, 0.06, 0.07, 0.08, y 0.09.

3.2 Experimentos tipo 2

Una vez que se encontraron los mejores valores para α y N se aplican en el algoritmo sobre el resto de las palabras del corpus para una repetición por palabra.

4 Resultados experimentales

4.1 Resultados de los experimentos tipo 1

Las mediciones utilizadas para evaluar la calidad de la segmentación fueron ϵ_n, $\epsilon_p{}'$, y $Total\_\epsilon$. Las tablas 1, 2 y 3 describen los errores ϵ_n, $\epsilon_p{}'$, y $Total\_\epsilon$ respectivamente para todos los valores de α cuando el parámetro N corresponde a 7 ms, lo cual implica un vecindario de 14 ms. Las tablas 4, 5 y 6 describen los errores para cada valor de α cuando el parámetro N es de 10 ms, es decir, un vecindario de 20 ms. Las tablas 7, 8 y 9 describen los errores obtenidos para cada valor de α cuando el parámetro N es de 15 ms, es decir, un vecindario de 30 ms.

Primero, con las tablas 3, 6 y 9 buscamos minimizar el error total, observando qué valores de α minimizaron el error total. Al minimizar el error total también se minimizan tanto el error en posición de fronteras como el error en número de segmentos. Luego, para esos valores de α que minimizaron el error total, observamos las casillas correspondientes en las tablas 2, 5 y 8, que representan el error en las posiciones de los segmentos.

Tabla 1. Valores de ϵ_n para cuando N corresponde a 7 ms. Incrementar α disminuye este error.

ϵ_n											
	N → 0.007										
Palabra	alfa= 0.04	alfa= 0.05	alfa= 0.06	alfa= 0.07	alfa = 0.08	alfa= 0.09	alfa= 0.1	alfa= 0.11	alfa= 0.12	alfa= 0.13	alfa= 0.14
"banco"	36.11	20.56	13.67	9.44	8.22	7.33	5.56	3.67	2.44	2.00	1.22
"cepillo"	46.33	32.44	22.89	16.56	12.44	9.00	6.00	4.22	3.00	2.33	1.22
"dentro"	41.11	23.67	16.33	11.44	8.67	8.00	7.33	5.89	4.78	4.11	2.56
"efectividad"	25.73	15.87	11.73	9.53	8.53	7.27	5.07	3.07	1.80	1.20	0.60
"enroscar"	38.55	23.09	16.55	13.27	11.00	9.36	7.73	5.55	3.91	2.55	1.36
"gama"	49.67	30.50	21.00	17.33	14.67	12.17	10.17	7.00	5.17	3.50	2.67
"jugar"	41.00	24.00	16.57	12.86	11.43	8.29	6.00	4.14	2.71	0.86	0.43
"leña"	47.83	28.00	21.00	18.67	16.67	13.83	11.17	7.00	5.50	4.00	2.33
"pecho"	27.86	18.29	15.14	13.86	12.29	11.14	9.43	6.29	4.57	3.43	1.57
"un_yugo"	36.38	25.00	20.25	17.50	15.00	12.75	10.63	6.88	5.38	3.00	2.00
Promedio	39.06	24.14	17.51	14.05	11.89	9.91	7.91	5.37	3.93	2.70	1.60

Tabla 2. Valores de $\epsilon_p{}'$ para cuando N corresponde a 7 ms. Disminuir α disminuye este error.

ϵ_p'											
	N → 0.007										
Palabra	alfa= 0.04	alfa= 0.05	alfa= 0.06	alfa= 0.07	alfa= 0.08	alfa= 0.09	alfa= 0.1	alfa= 0.11	alfa= 0.12	alfa= 0.13	alfa= 0.14
"banco"	1.15	1.87	2.64	5.45	6.20	7.38	11.59	16.74	21.61	21.61	22.53
"cepillo"	1.10	1.10	1.10	3.24	3.56	5.11	12.91	23.29	35.20	44.54	44.54
"dentro"	1.94	2.47	2.88	3.33	6.72	6.72	6.72	9.05	9.05	9.17	26.83
"efectividad"	1.73	2.42	4.19	4.36	4.86	6.85	12.35	20.58	24.92	32.29	39.53
"enroscar"	1.74	2.87	3.33	4.50	4.64	6.66	7.84	10.05	15.01	17.06	22.84
"gama"	1.76	2.21	6.78	6.79	8.42	8.65	9.25	17.12	17.12	32.53	32.53
"jugar"	0.84	1.94	2.89	4.64	5.06	7.07	8.06	10.04	12.56	21.71	31.60
"leña"	1.52	2.56	2.56	2.56	2.56	6.94	6.94	15.18	21.40	26.84	48.54
"pecho"	1.00	2.12	2.16	3.71	3.72	4.90	5.35	10.08	10.23	20.11	22.18
"un_yugo"	1.88	1.88	2.30	2.99	3.57	3.66	5.27	7.82	14.10	26.23	26.23
Promedio	1.47	2.14	3.08	4.16	4.93	6.39	8.63	13.99	18.12	25.21	31.74

Tabla 3. Valores de $Total\_\epsilon$ para cuando N corresponde a 7 ms. Algunos valores de α y N producen valores mínimos en el error.

Total_ϵ											
	N → 0.007										
Palabra	alfa= 0.04	alfa= 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09	alfa = 0.10	alfa = 0.11	alfa = 0.12	alfa = 0.13	alfa = 0.14
"banco"	36.13	20.64	13.92	10.90	10.30	10.40	12.86	17.14	21.75	21.71	22.57
"cepillo"	46.35	32.46	22.92	16.87	12.94	10.35	14.24	23.67	35.33	44.60	44.56
"dentro"	41.16	23.80	16.59	11.92	10.97	10.45	9.95	10.79	10.23	10.05	26.95
"efectividad"	25.79	16.05	12.46	10.48	9.82	9.99	13.35	20.80	24.99	32.31	39.53
"enroscar"	38.58	23.27	16.88	14.01	11.94	11.49	11.01	11.48	15.51	17.25	22.89
"gama"	49.70	30.58	22.07	18.61	16.91	14.93	13.75	18.50	17.88	32.72	32.64
"jugar"	41.01	24.08	16.82	13.67	12.50	10.89	10.04	10.86	12.85	21.72	31.61
"leña"	47.86	28.12	21.16	18.84	16.86	15.48	13.15	16.72	22.09	27.13	48.60
"pecho"	27.88	18.41	15.30	14.35	12.84	12.17	10.84	11.88	11.21	20.40	22.24
"un_yugo"	36.42	25.07	20.38	17.75	15.42	13.27	11.86	10.41	15.09	26.40	26.31
Promedio	39.09	24.25	17.85	14.74	13.05	11.94	12.10	15.23	18.69	25.43	31.79

Tabla 4. Valores de ϵ_n para cuando N corresponde a 10 ms. Incrementar α disminuye este error.

ϵ_n										
	N → 0.010									
Palabra	alfa = 0.04	alfa = 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09	alfa = 0.1	alfa = 0.11	alfa = 0.12	alfa = 0.13
"banco"	13.33	9.00	8.11	6.78	5.00	3.56	1.78	0.67	0.33	0.22
"cepillo"	24.56	18.00	11.67	7.33	5.00	2.78	1.44	0.44	0.33	0.56
"dentro"	16.44	10.00	7.67	7.00	6.00	4.44	2.78	0.67	0.11	0.56
"efectividad"	12.47	9.93	8.20	6.80	5.07	2.60	1.07	0.53	0.20	0.20
"enroscar"	16.73	12.55	10.82	7.91	6.36	3.73	1.64	0.91	0.00	0.55
"gama"	20.33	16.50	14.67	12.67	9.00	5.33	3.33	2.33	0.50	0.17
"jugar"	16.43	12.00	10.29	7.71	5.57	2.43	1.00	0.29	0.43	0.86
"leña"	22.50	19.33	16.67	12.33	7.00	3.17	2.50	1.50	0.33	0.17
"pecho"	15.14	12.86	11.57	10.57	8.86	4.43	2.71	1.29	0.29	0.57
"un_yugo"	19.50	16.00	14.38	11.38	6.75	2.88	0.75	0.00	0.50	0.75
Promedio	17.74	13.62	11.40	9.05	6.46	3.53	1.90	0.86	0.30	0.46

Tabla 5. Valores de $\epsilon_p{}'$ para cuando N corresponde a 10 ms. Disminuir α disminuye este error.

ϵ_p'										
	N → 0.010									
Palabra	alfa = 0.04	alfa = 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09	alfa = 0.1	alfa = 0.11	alfa = 0.12	alfa = 0.13
"banco"	3.44	5.81	8.79	7.17	12.24	16.75	19.66	44.82	46.36	87.87
"cepillo"	1.10	3.06	4.92	11.11	16.29	36.71	61.51	65.35	78.38	97.20
"dentro"	3.47	3.62	9.05	9.05	9.05	9.69	18.81	44.63	64.40	129.76
"efectividad"	3.69	4.69	4.86	8.26	11.17	25.41	35.73	47.43	78.08	90.18
"enroscar"	3.75	4.47	4.64	8.15	13.20	17.09	30.95	55.38	104.68	487.45
"gama"	6.79	8.64	8.64	8.65	14.24	17.71	36.12	36.12	60.17	125.43
"jugar"	2.77	4.64	6.49	6.50	10.04	12.11	19.89	61.09	77.07	168.48
"leña"	2.56	2.56	2.56	8.81	10.21	24.95	24.98	88.28	131.13	131.13
"pecho"	5.94	7.56	7.56	7.57	12.46	15.50	31.61	31.61	52.65	109.75
"un_yugo"	2.46	4.12	5.77	5.77	8.93	10.76	17.68	54.30	68.50	149.76
Promedio	3.60	4.92	6.33	8.10	11.78	18.67	29.70	52.90	76.14	157.70

Tabla 6. Valores de $Total\_\epsilon$ para cuando N corresponde a 10 ms. Algunos valores de α y N producen valores mínimos en el error.

Total_ϵ										
	N → 0.010									
Palabra	alfa= 0.04	alfa= 0.05	alfa= 0.06	alfa= 0.07	alfa= 0.08	alfa= 0.09	alfa= 0.10	alfa= 0.11	alfa= 0.12	alfa= 0.13
"banco"	13.77	10.71	11.96	9.86	13.22	17.12	19.74	44.83	46.36	87.87
"cepillo"	24.58	18.26	12.66	13.31	17.04	36.81	61.53	65.35	78.38	97.20
"dentro"	16.81	10.64	11.86	11.44	10.86	10.66	19.01	44.63	64.40	129.76
"efectividad"	13.00	10.99	9.53	10.70	12.27	25.54	35.75	47.43	78.08	90.18
"enroscar"	17.14	13.32	11.77	11.36	14.65	17.49	30.99	55.39	104.68	487.45
"gama"	21.44	18.63	17.02	15.34	16.85	18.49	36.28	36.20	60.17	125.43
"jugar"	16.66	12.86	12.16	10.08	11.48	12.35	19.92	61.09	77.07	168.48
"leña"	22.65	19.50	16.86	15.15	12.38	25.15	25.11	88.30	131.13	131.13
"pecho"	16.27	14.92	13.82	13.00	15.29	16.12	31.72	31.63	52.65	109.75
"un_yugo"	19.65	16.52	15.49	12.76	11.19	11.14	17.70	54.30	68.51	149.76
Promedio	18.20	14.63	13.31	12.30	13.52	19.09	29.77	52.92	76.14	157.70

Tabla 7. Valores de ϵ_n para cuando N corresponde a 15 ms. Incrementar α disminuye este error.

ϵ_n						
	N → 0.015					
Palabra	alfa = 0.04	alfa = 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09
"banco"	7.78	6.78	4.44	2.00	0.78	0.11
"cepillo"	13.33	7.67	3.89	1.67	0.22	0.67
"dentro"	8.44	7.00	5.67	2.89	0.56	0.11
"efectividad"	9.07	6.40	3.07	1.20	0.40	0.27
"enroscar"	11.45	7.91	5.27	1.73	0.18	0.82
"gama"	15.00	11.83	7.67	3.83	1.83	0.17
"jugar"	10.57	7.29	4.29	2.14	0.29	0.71
"leña"	17.00	12.00	5.50	2.33	0.50	0.50
"pecho"	12.43	9.71	5.14	2.43	0.14	0.29
"un_yugo"	14.25	11.25	5.00	0.38	0.63	0.88
Promedio	11.93	8.78	4.99	2.06	0.55	0.45

Tabla 8. Valores de $\epsilon_p{}'$ para cuando N corresponde a 15 ms. Disminuir α disminuye este error.

ϵ_p'						
	N → 0.015					
Palabra	alfa = 0.04	alfa = 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09
"banco"	8.79	8.79	13.41	19.66	26.12	62.97
"cepillo"	3.07	10.53	23.41	47.06	75.10	102.05
"dentro"	5.67	9.05	9.09	29.95	55.41	62.90
"efectividad"	4.83	6.26	20.75	52.42	70.05	89.22
"enroscar"	4.64	10.57	17.38	47.47	104.22	433.09
"gama"	8.64	12.43	21.63	53.90	59.85	299.31
"jugar"	6.04	6.50	15.51	20.81	71.14	125.82
"leña"	3.79	4.43	28.10	38.85	124.52	154.36
"pecho"	4.21	9.88	19.90	33.49	69.99	74.52
"un_yugo"	3.19	4.47	9.55	45.99	102.54	211.50
Promedio	5.29	8.29	17.87	38.96	75.89	161.57

Tabla 9. Valores de $Total\_\epsilon$ para cuando N corresponde a 15 ms. Algunos valores de α y N producen valores mínimos en el error.

Total_ϵ						
	N → 0.015					
Palabra	alfa = 0.04	alfa = 0.05	alfa = 0.06	alfa = 0.07	alfa = 0.08	alfa = 0.09
"banco"	11.74	11.10	14.13	19.76	26.13	62.97
"cepillo"	13.68	13.02	23.73	47.09	75.10	102.05
"dentro"	10.17	11.44	10.71	30.09	55.41	62.90
"efectividad"	10.27	8.96	20.98	52.43	70.05	89.22
"enroscar"	12.36	13.21	18.16	47.50	104.22	433.09
"gama"	17.31	17.16	22.95	54.04	59.88	299.31
"jugar"	12.18	9.76	16.09	20.92	71.14	125.82
"leña"	17.42	12.79	28.64	38.92	124.53	154.36
"pecho"	13.12	13.86	20.56	33.58	69.99	74.52
"un_yugo"	14.60	12.10	10.78	45.99	102.54	211.51
Promedio	13.29	12.34	18.67	39.03	75.90	161.58

Para escoger los mejores valores de α y N, se tomaron las siguientes consideraciones:

- El error Total_ϵ debe ser minimizado, lo cual hace que los errores ϵ_n y $\epsilon_p{}'$ también se minimicen ambos al mismo tiempo.
- Escoger los que generen los segmentos más largos.

Los resultados de los experimentos tipo 1 y los criterios antes mencionados muestran que los valores escogidos son:

$$\alpha = 0.1$$
$$N => 0.007 \text{ ms} = 112 \text{ muestras} \quad (10)$$

Un ejemplo de segmentación automática generada por el algoritmo se muestra en la figura 1.

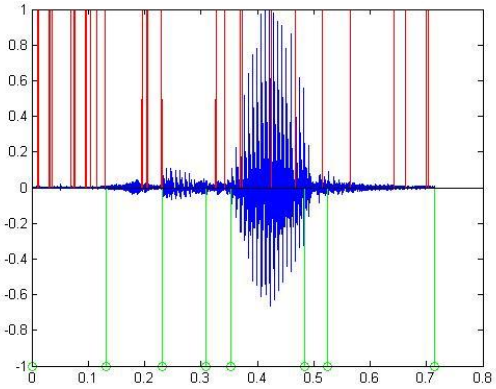

Fig. 1. Ejemplo de segmentación automática (rojo), segmentación manual (verde).

4.2 Resultados de los experimentos tipo 2

En este conjunto de experimentos se probaron los valores de α y N son usados para el resto del corpus (para una repetición por palabra). La tabla 10 muestra los valores promedio, máximo y mínimo par a las mediciones de error aplicadas.

Tabla 10. Resultados para todas las clases de palabras en el corpus.

	Promedio	Max	Min
λ	0.89129	0.94872	0.73913
μ	0	0	0
ϵ_n	8.8932	18.5	2.83333
ϵ_p	72.9354	207.348	6.62012
ϵ	117.401	233.062	60.2768
$\epsilon_p{}'$	9.77216	29.6211	1.10335
Total_ϵ	13.754	30.0643	7.882

4.3 Comparación

Aunque no se puede hacer una comparación precisa o completa con los resultados publicados en [5], algunas mediciones se pueden comparar tomando en cuenta el tamaño del vecindario *t_min*. En la tabla 11 se pueden observar los resultados (publicados originalmente como una tabla) en el artículo [5].

Tabla 11. Resultados de exactitud promedio en milisegundos reportados en [5].

	t_min				
α	5 ms	10 ms	15 ms	20 ms	30 ms
0	7.6	11.2	12.7	13.9	16.1
0.2	7.7	11.2	12.7	13.9	16.1
0.6	9	11.5	12.9	14	16.1
1	10.8	12.2	13.2	14	16.1
1.4	12.6	13.3	13.7	14	15.4
1.8	13.2	13.5	13.7	13.8	14.5

Comparando la exactitud promedio reportada por Galka y Ziolko con el promedio de $\epsilon_p{}'$ (la cual es también una medida promedio de qué tan cercanas están las fronteras de segmentos automáticos a las fronteras de segmentos manuales) en las tablas 3, 6 y 9 se puede concluir que para los vecindarios de 15 ms, 20 ms y 30 ms se logró una distancia menor entre fronteras automáticas y fronteras manuales que en [5].

Conclusiones

En este artículo se modificó un algoritmo de segmentación de habla basado en el ordenamiento de las subbandas wavelet de acuerdo a su energía, para tomar en cuenta aspectos perceptuales del oído humano utilizando para ello la escala de Bark. El error obtenido es menor que el error reportado en la versión original del algoritmo, con la ventaja de que en este trabajo se evitan los borrados de segmentos.

Comparando las tablas como se menciona en la sección 4.1 se obtiene, que el método original, para un vecindario 15 ms tiene una separación promedio mínima de 12.7 ms, mientras que en el método aquí propuesto se obtiene 8.63 ms; esto es, que las fronteras automáticas están más próximas a las manuales. Para el vecindario de 20 ms, el método original reporta 13.82 ms; el método de este trabajo 4.16 ms. Para el vecindario de 30 ms, el método original reporta 14.5 ms; el método de este trabajo 4.92 ms. De estos tres valores, el mejor es el primero porque es el que aumenta menos la sobre segmentación.

Como trabajo futuro se experimentará combinando el algoritmo presentado con la técnica de alineamiento forzado para disminuir la sobre segmentación.

Agradecimientos

Este trabajo fue apoyado por CONACyT y por el Instituto Politécnico Nacional a través del proyecto SIP 20110889.

Referencias

[1]. Alani, A., Deriche, M.: A novel approach to speech segmentation using the wavelet transform. Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, vol. 1, pp. 127-130 (1999)

[2]. De Mori, R., Laface, P.: Use of fuzzy algorithms for phonetic and phonemic labeling of continuous speech. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-2, Issue 2, pp. 136-148 (1980)

[3]. Fantinato, P. C. et al.: A fractal-based approach for speech segmentation. Tenth IEEE International Symposium on Multimedia, pp. 551-555 (2008)

[4]. Finster, H.: Automatic speech segmentation using neural network and phonetic transcription. International Joint Conference on Neural Networks, vol. 4, pp. 734-736 (1992)

[5]. Galka, J., Ziolko, M.: Wavelets in speech segmentation. The 14th IEEE Mediterranean Electromagnetical Conference, pp. 876-879 (2008)

[6]. Grieder, W., Kisner, W.: Speech segmentation by variance fractal dimension. 1994 Canadian Conference on Electrical and Computer Engineering, Conference Proceedings, vol. 2, pp. 481-485 (1994)

[7]. Hosom, J. P.: Automatic time alignment of phonemes using acoustic-phonetic information. Thesis (PhD). Oregon Graduate Institute of Science and Technology, pp. 35-36 (2000)

[8]. Janer, L.: Transformada wavelet aplicada a la extracción de información en señales de voz. Tesis doctoral. Universitat Politècnica de Catalunya. Barcelona, pp. 63-73 (1998)

[9]. Janer, L., Marti, J., Nadeu, C., Leida-Solano, E.: Wavelet transforms for non-uniform speech recognition systems. Fourth International Conference on Spoken Language, vol. 4, pp. 2348-2351 (1996)

[10]. Milone, D. H., Merelo, J. J., Rufiner, H. L.: Evolutionary algorithm for speech segmentation. Proceedings of the 2002 Congress on Evolutionary Computation, vol. 2, pp. 1115-1120 (2002)

[11]. Oropeza-Rodgriguez, J-L.: Algoritmos y métodos para el reconocimiento de voz en español mediante sílabas. Tesis doctoral. Instituto Politécnico Nacional, Centro de Investigación en Computación, pp. 28-29 (2006)

[12]. Rabiner, L., Biing-Hwang, J.: Fundamentals of speech recognition. Prentice Hall International, pp. 436-437 (1993)

[13]. Reddy, R.: Speech recognition by machine: a review. Proceedings of the IEEE, vol. 64, pp. 501-531 (1976)

[14]. Spohrer, J.C., Brown, P. F., Roth, R.: Automatic Labeling of Speech. IEEE International Conference on Acoustics, Speech and Signal Processing, vol 7, pp. 1641- 1644 (1982)

[15]. Wendt, C., Petropulu, A. P.: Pitch determination and speech segmentation using the discrete wavelet transform. International Symposium on Circuits and Systems, vol. 2, pp. 45-48 (1996)

[16]. Ziolko, B., Manandhar, S., Wilson, R. C.: Phoneme segmentation of speech. 18th International Conference on Pattern Recognition, vol. 4, pp. 282-285 (2006)
[17]. Zwicker, E.: Subdivision of the audible frequency range into critical bands (frequenzgruppen). Journal of the Acoustical Society of America, vol. 33, issue 2, pp. 248-248 (1961)