Segmentación del habla con independencia de texto utilizando características en el dominio del tiempo

Luis D. Huerta Hernández

Universidad del Istmo Campus Ixtepec
Ciudad Universitaria s/n,Cd. Ixtepec, Oaxaca. C.P. 70110
luisdh2@hotmail.com
Paper received on 12/08/08, accepted on 06/09/08.

Resumen. Recientemente, se han reportado algoritmos para la segmentación fonética, probados con habla continua, con independencia de hablante y de texto. Los enfoques propuestos en el estado del arte basan su procesamiento principalmente en características espectrales o cepstrales del habla, utilizando cuantificaciones vectoriales por unidad de tiempo, que implican transformaciones del habla del dominio del tiempo al dominio de frecuencias. En el algoritmo propuesto se han utilizado cuantificaciones físicas del habla como la intensidad, amplitud y energía como una nueva alternativa para llevar a cabo la segmentación fonética automática, probada bajo condiciones similares que los reportados en el estado del arte. Se hace énfasis en un desempeño competitivo, reducida cantidad de información, y reglas simples utilizadas para llevar a cabo la segmentación fonética del habla. Se describe, la implementación del algoritmo, condiciones experimentales y resultados.

1 Introducción

Actualmente, en las Tecnologías del Habla se están considerando con mayor importancia las unidades de sub-palabras como los fonemas, puesto que para el proceso de reconocimiento estas unidades reducen la complejidad de modelado, de clasificación, y de almacenamiento de información de los lenguajes. Trabajos recientes, se han enfocado en la segmentación fonética del habla con independencia de texto, que consiste en obtener las posiciones de las fronteras entre fonemas, a partir de la onda de habla sin el apoyo de ningún tipo de información conocida previamente, como lo es comúnmente el texto. La segmentación independiente de texto es fundamental en los sistemas de reconocimiento del habla basados en fonemas, puesto que puede afectar la calidad del proceso de reconocimiento. El algoritmo propuesto puede ser de utilidad en la segmentación del habla con independencia de texto para su reconocimiento fonético. También puede ser de utilidad en: sistemas de síntesis de voz que basan su funcionamiento en la concatenación de segmentos; segmentación y etiquetado manual del habla para la creación de bases de datos de investigación fonética, donde la intervención manual es altamente tediosa y consumidora de tiempo. Aunque se han reportado trabajos encausados a la segmentación en subpalabras, éstos han sido probados bajo una serie de restricciones como dependencia de hablante [1] [2], texto [3][4], vocabulario [5] [6], sin hacer uso de habla continua expresada naturalmente y sin considerar la sobre-segmentación [7]. Recientemente

© M. G. Medina Barrera, J. F. Ramírez Cruz, J. H. Sossa Azuela (Eds.).
Advances in Intelligent and Information Technologies.
Research in Computing Science 38, 2008, pp. 129-139.

se propuso un método [8] que suprime todas estas restricciones alcanzando una tasa de correctas detecciones de límites del 76.53% y una tasa de sobre-segmentación cercana al 0 %. El método anteriormente citado, emplea características obtenidas de los espectros Mel, lo cual implica una transformación del dominio del tiempo al dominio de frecuencias de la señal de habla para obtener dichas características en forma de vectores por unidad de tiempo; una vez obtenidas las características en vectores, se procede a utilizarlas para aplicarles el proceso de detección de límites fonéticos. Surge una motivación de segmentar la señal de habla a nivel fonético, utilizando características encontradas en el dominio del tiempo como la intensidad, amplitud y energía, con el propósito de evitar una transformación a codificación vectorial que implica procesamiento extra y reducir el número de características extraídas de la señal. Los resultados obtenidos de la segmentación son comparados con métodos reportados recientemente, probados en condiciones similares.

2 Características acústicas del habla en el dominio del tiempo

El objetivo de analizar características en el dominio del tiempo es conocer la utilidad de algunas de ellas en el proceso de segmentación. Las características comúnmente utilizadas en el dominio del tiempo son la energía, intensidad, frecuencia fundamental y cruces en cero, por mencionar algunas. Ciertamente, las características acústicas en el dominio del tiempo cuantifican los fenómenos físicos producidos al hablar, como puede ser por ejemplo la presión ejercida en el aire (amplitud) por las ondas sonoras. Estos fenómenos físicos se cuantifican teniendo valores escalares en un instante de tiempo t.

Se hizo el análisis sobre métricas de la intensidad, amplitud y energía; aunque estas medidas son relacionadas, no presentan exactamente los mismos comportamientos en los límites fonéticos. En este análisis se compararon visualmente los puntos de segmentación contenidos en la señal de habla (establecidos por expertos fonéticos) con la gráfica de los valores de cada característica, y la relación que tienen en el tiempo.

A continuación, se muestra una señal de habla expresada por un hablante femenino, y los límites fonéticos (●) sobre valores de la amplitud, energía e intensidad.

2.1 Amplitud

En la Figura 1, se observa la transcripción fonética en la parte superior, la señal de habla, y la gráfica relacionada con los valores de la amplitud en la parte inferior.

2.2 Energía

La energía puede ser utilizada para distinguir conjuntos vocálicos de consonánticos, así como para la detección de presencia o ausencia de voz sobre señales de alta calidad por medio de umbrales.

Puesto que la energía es obtenida a partir de la amplitud, se observan resultados similares entre ambas características en relación con el mapeo de límites fonéticos.

como se puede observar en la Figura 2. Sin embargo, los valores de amplitud presentan cambios más notorios que los presentados en los valores de la energía.

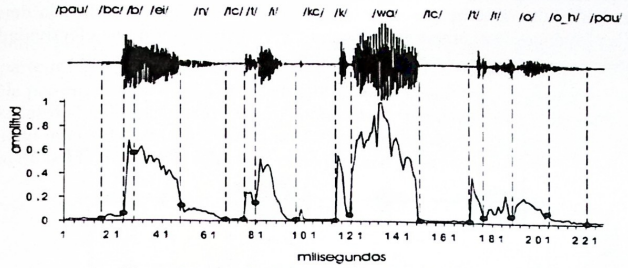

Fig. 1. Límites fonéticos sobre valores de la amplitud

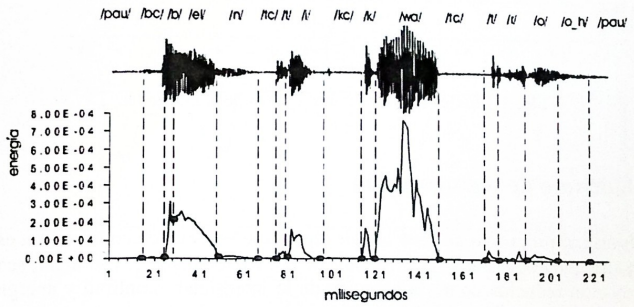

Fig. 2. Límites fonéticos sobre valores de la energía

2.3 Intensidad

La intensidad, al igual que la energía son métricas obtenidas a partir de la amplitud. Cuando el sonido es producido, la energía de la fuente es irradiada sobre el área donde se propagan las ondas sonoras. Tanto la amplitud como la energía de una señal de habla contienen un gran número de pequeños cambios que rodean a los límites fonéticos, de tal forma que se podrían reflejar en el desempeño de la segmentación como falsos límites (inserciones).

Para reducir el inconveniente de las inserciones, una posible solución es aplicar un suavizado de medias sobre los valores de estas características del dominio del tiempo. Por otro lado, aplicando el suavizado podría ocasionar la pérdida de algunos límites fonéticos.

La intensidad tiene la característica de no presentar tantas variaciones en el tiempo por tener baja sensibilidad a los cambios de amplitud, en comparación con la sensibilidad de la energía. En la Figura 3 se muestran los límites fonéticos presentes en la intensidad, de acuerdo a ésta, la intensidad es una métrica que presenta cambios graduales ligeramente perceptibles, que aparentemente, complica la detección de límites fonéticos. Sin embargo, presenta en la segmentación una efectividad similar, e incluso mejor que la amplitud y energía en el dominio del tiempo.

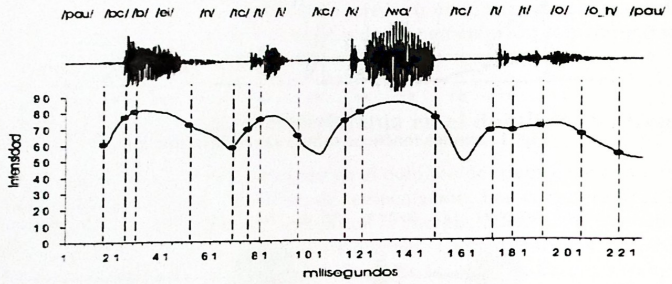

Fig. 3. Límites fonéticos sobre valores de la intensidad

3 Algoritmo de segmentación

El algoritmo propuesto no hace uso de algún tipo especial de codificación espectral o cepstral para determinar la existencia de límites fonéticos. En su lugar, se utiliza información reducida como la obtenida de la intensidad, amplitud y energía. El algoritmo proporciona instancias de tiempo en la que se ha detectado una transición entre fonemas, considerando una tolerancia en la exactitud de ± 20 milisegundos, también se caracteriza por su simplicidad y sencillez, demostrando efectividad, tal como se verá en la sección de experimentos.

3.1 Preprocesamiento

En primera instancia, se aplica el filtro pre-énfasis en la fase de preproceso del algoritmo propuesto. La función del filtro pre-énfasis es obtener un sonido de alta cuesta espectral, haciendo más notorios los cambios espectrales que se presentan en las transiciones de fonemas. Una frecuencia F debe proporcionarse, a partir de ella la cuesta espectral será incrementada en 6 dB/octava. El factor pre-énfasis α es procesado como:

$$\alpha = \exp(-2\prod F\Delta t) \tag{1}$$

Donde Δt es el periodo de muestreo del sonido. El sonido resultante y es obtenido con:

$$y_i = x_i - \alpha\ x_{i-1} \tag{2}$$

Donde cada muestra x del sonido es cambiada, a partir de la última muestra [11]. En la parte inferior de la Figura 4 (c), se puede observar que la intensidad de la señal de habla pre-enfatizada define de una mejor manera los cambios existentes a lo largo de la señal de habla, resultando en un contorno de intensidad mejor definido, mientras que la intensidad sin pre-énfasis no define muchos de los cambios contenidos en la señal de habla, lo cual es relevante en el proceso de segmentación.

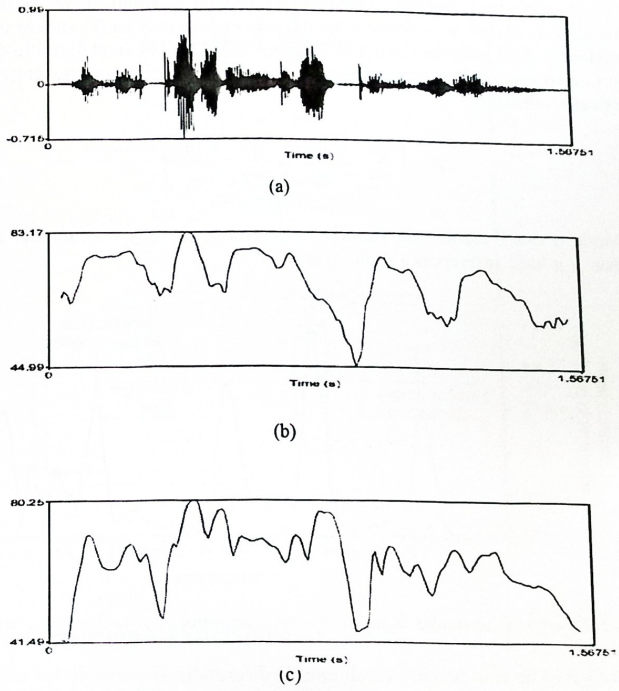

Fig. 4. (a) Señal de habla. (b) Intensidad de la señal sin pre-énfasis. (c) Intensidad de la señal con pre-énfasis

3.2 Detección de límites fonéticos

Este algoritmo hace uso de la amplitud, intensidad y la energía para llevar a cabo la segmentación. El algoritmo toma como entrada la señal de habla para pasarla por el filtro pre-énfasis a partir de una frecuencia *F*. En primera instancia, el algoritmo debe dividir la señal en silencio/sonido, en el cual la energía es comparada contra un umbral establecido, donde el silencio es detectado si el segmento de la señal bajo análisis se encuentra bajo el umbral de sonido, en caso contrario el segmento de la señal será algún tipo de sonido [12].

En una primera fase el algoritmo determina los límites entre pausas y sonidos, procesando en la siguiente fase aquellos segmentos en los cuales hay presencia de habla.

La segunda fase consiste en detectar las diferencias significativas en la amplitud, intensidad o energía de la señal de habla, para esto se emplea (3), que es una métrica de distancia. Una métrica similar es utilizada en [8, 9, 10] en el dominio de frecuencias, puesto que en la segmentación, se buscan diferencias notorias entre las características analizadas.

$$\delta = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right| \tag{3}$$

Donde *a* es el número de frames considerados antes y después del frame bajo análisis, y *n* hace referencia a dicho frame.

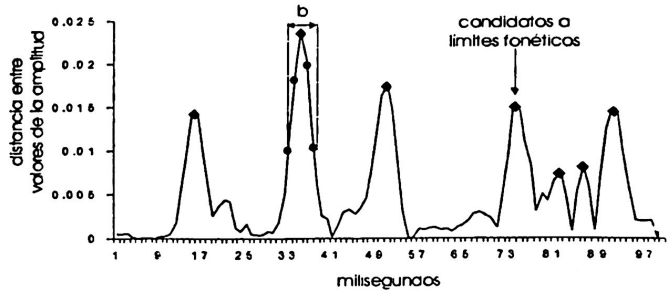

Fig. 5. Diferencias absolutas de las medias de la amplitud, y sus respectivos parámetros

El objetivo de esta función es obtener la diferencia absoluta de las medias de los *a* valores de la amplitud previos y posteriores respecto al frame bajo análisis. De esta manera se obtienen máximos locales que representan una máxima diferencia entre los valores de la amplitud.

Si se denota como *vi* a los valores representados en la Figura 5, los candidatos a límites fonéticos (presentes como máximos locales) deberán cumplir con las siguientes condiciones:

- $v_t > v_{t-1} \ldots v_{t-b}$ y $v_t > v_{t+1} \ldots v_{t+b}$
- $|v_t - v_{t-b}| > 0.0015$ y $|v_t - v_{t+b}| > 0.0015$

El subíndice b en las condiciones es un parámetro que indica cuantos valores previos y posteriores obtenidos con (3) deberán considerarse para cumplir dichas condiciones. Puesto que los cambios de las características físicas entre fonemas se dan manera gradual, por el efecto de la coarticulación, el parámetro b permite detectar solo aquellos máximos locales que se forman gradualmente.

Este algoritmo utiliza frames de 4 ms en la fase de detección de bordes referente a la discriminación entre presencia y ausencia de habla.. En la segunda fase correspondiente al procesamiento del segmento donde existe habla, el tamaño de los frames dependerá de la característica usada para segmentar.

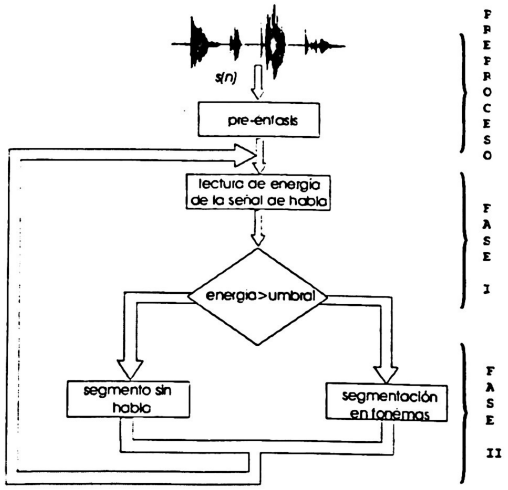

Fig. 6. Diagrama del algoritmo

4 Implementación y experimentos

4.1 PRAAT v.4.4.0.04

El proceso de extracción de las características, la aplicación del preproceso, y el proceso de detección de fonemas, fueron implementados usando el freeware PRAAT v.4.4.0.04 [11].

4.2 Bases de datos experimentales

En los experimentos, se hace uso de la base de datos de habla DARPA TIMIT, que fue diseñada para proporcionar datos fonético-acústicos del habla para el desarrollo y evaluación de sistemas de reconocimiento automático del habla. Consiste de expresiones de 630 hablantes que representan la mayoría de los dialectos del inglés americano. Las regiones dialécticas incluidas son: New England, Northern, North Midland, South Midland, Southern, New York City, Western y Army Brat.

El algoritmo es probado con oraciones en inglés del corpus TIMIT, del cual se han extraído 544 señales expresadas por 68 hablantes (34 masculinos y 34 femeninos). El grupo de oraciones incluyen un total de 20647 límites fonéticos, cantidad que será referida como S_t en las medidas de desempeño (4) y (5).

En los experimentos se suprimen muchas de las restricciones bajo las que han sido probados otros algoritmos previamente citados. Los experimentos se llevaron a cabo utilizando habla continua, una gran variedad de hablantes, sin ningún tipo de información adicional que la presente en la señal de habla.

4.3 Medidas de desempeño

El desempeño del algoritmo, fue evaluado con medias de desempeño comúnmente utilizadas, como en [8, 9, 10].

$$D = 100 \cdot \left(\frac{S_d}{S_t} - 1 \right) \tag{4}$$

Donde D es la medida de sobre-segmentación, S_d es el número de puntos de segmentación detectados por el algoritmo, y S_t es el número de puntos reales de segmentación.

$$P_c = 100 \cdot \left(\frac{S_c}{S_t} \right) \tag{5}$$

Donde P_c es el porcentaje de detecciones correctas, y Sc es el número de puntos de segmentación correctos. Los puntos de segmentación son considerados como correctos, si la distancia hacia el punto verdadero de segmentación se encuentra en el rango de ± 20 ms.

4.4 Resultados experimentales

En primera instancia, se hicieron pruebas para detectar bordes, esto es, detectar presencia y ausencia de habla sobre oraciones de las bases de datos, aplicando un umbral de energía de 135 E^{-19} Pa^2 para este efecto. Este nivel de energía sería utilizado como umbral en los siguientes experimentos para detectar bordes (silencio/sonidos).

Puesto que en la mayoría de los límites fonéticos representan cambios notorios entre las características físicas de sus fonemas adyacentes, el parámetro *c* mostrado en la Tabla 1, funge como umbral para obtener solo aquellos máximos locales de altura considerable respecto a otros. Se experimentó con diversos valores tanto para los parámetros *a, b* y *c* cómo para el tamaño de frames en milisegundos, mostrando en la Tabla 1 los valores con los que se obtuvo el mejor desempeño para este algoritmo, con cada característica.

Tabla 1. Parámetros utilizados

Característica	a	b	c	frame en ms.
Intensidad	2	4	2.9	5
Amplitud	2	4	0.035	10
Energía	2	4	0	10

El desempeño en términos de tasa de detecciones correctas y tasa de inserciones se muestra en la Tabla 2, considerando un total de 20647 límites fonéticos por detectar. El uso de la intensidad en este algoritmo genera una tasa de correcta segmentación aproximada al 77% y una tasa de sobre-segmentación debajo del 0%, siendo la intensidad considerablemente más eficiente que la amplitud. Por otro lado, el uso de la energía proporciona un desempeño inferior, puesto que la tasa de inserciones es muy alta.

Tabla 2. Desempeño del algoritmo, utilizando

Característica	Sd	Sc	% Sd	% D
Intensidad	20547	15775	76.40	-0.48
Amplitud	21448	14852	71.93	3.87
Energía	23667	14519	70.32	14.62

A pesar de ser un enfoque simple, proporciona resultados aceptables, tomando en consideración que hace uso reducido de información con 2 valores cada 20 ms cuando no hay presencia de voz, y 4 valores cada 20 ms cuando hay que segmentar los fonemas.

4.5 Comparación con trabajos similares

Los resultados con comparados con algoritmos como [8, 9, 10], que fueron probados en condiciones similares usando la base de datos TIMIT, sin dependencia de texto, de hablante, de vocabulario y haciendo uso de habla continua. En la tabla 3 se muestra información relevante como valores utilizados por unidad de tiempo, fonemas tratados, hablantes involucrados y el porcentaje de detecciones correctas, considerando el aspecto de mantener una sobre-segmentación cercana al 0%.

Los algoritmos con los que se compara el algoritmo propuesto, se basan en características vectoriales en el dominio de frecuencias, además de hacer uso de un post-procesamiento denominado "ajuste", en el cual se obtienen finalmente los límites fonéticos, radicando aquí una principal diferencia, puesto que el algoritmo aquí descrito utiliza características en el dominio del tiempo, basado en simples reglas, sin post-procesamiento alguno.

Cada uno de los valores expuestos en la Tabla 3, han sido reportados por sus respectivos autores, tomando en cuenta que se busca una tasa de sobre-segmentación cercana al 0%.

Tabla 3. Comparativa de algoritmos considerando diversos aspectos

Característica usada	Valores extraídos cada 20 ms	Hablantes	Límites fonéticos tratados	% Correcta detección
Intensidad	4	68	20647	76.40
Espectros Mel[8]	8	48	17930	76.53
PCBF[9]	15	48	17930	73.56
PCBF[10]	15	20	6200	75.80

Los límites fonéticos difíciles de detectar, son principalmente aquellos incluidos en diptongos, aunado al hecho que existen límites fonéticos detectados por los algoritmos, ligeramente fuera del rango de los ± 20 ms.

5 Conclusiones

Se hace uso de diversas características acústicas del habla presentes en el dominio del tiempo como la energía, amplitud e intensidad de manera independiente, sin ser antes probadas de manera individual en la segmentación fonética con independencia de texto. A diferencia de métodos como [1], que hacen uso de varias características en el dominio del tiempo y frecuencias simultáneamente para efectuar la segmentación, en el algoritmo aquí propuesto se utilizan a lo mucho dos características de este tipo. La ventaja de este algoritmo es que promueve un procesamiento simple, tanto en la extracción de dichas características como en su procesamiento, sin tener la necesidad de hacer uso de algún tipo de codificación del habla más elaborada como vectores de características.

Referencias

1. Suh Y. and Lee Y. Segmentation of continuous speech using multi-layer perceptron. IEEE Trans. Speech and Audio Proc., 1999.

2. Fernández L. Aportaciones a la Mejora de los Sistemas de Reconocimiento. PhD thesis. Universidad de Vigo. 2001.3. R. Egas et al. (1999). Adapting k-d Trees to Visual Retrieval, Proc. Visual 99, LNCS 1614: 533-540.
3. Hansen J. Pellom B. Automatic segmentation of speech recorded on unknown noisy channel characteristics. Speech Communication, 1998.
4. Mayora O. Segmentazione automatica di fonemi per aplicazioni di riconoscimento vocale. Technical report, Universitµa di Genova, 2000.
5. Bernard E. Cole R. Hu Z., Schalwyk J. Speech recognition using syllabe-like units. ICSLP '96, 1996.
6. Korhonen P. Unsupervised segmentation of continuous speech using vector autoregressive modeling. Master's thesis. Helsinki University of Technology. 2004.
7. Dalsgaard P. Petek B., Andersen O. On the robust automatic segmentation of spontaneous speech. Proceedings of ICSLP '96, 1996.8. M. S. Lew (2000). Next-Generation Web Searches for Visual Content, IEEE Computer, 33(11):46-52.
8. Aversano G. and Esposito A., "Automátic Parameter Estimation for a Context-Independent Speech Segmentation Algorithm", TSD 2002, LNAI 2448, pp. 293-300, 2002 Springer Verlag Berlin Heidelberg 2002.
9. Aversano G. and Esposito A., "A new text-independent method for phoneme segmentation". in Proc. the 44th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516--519. 2001.
10. Saraswhati S., Geetha T.V, and Saravanan K., "Integrating Language Independent Segmentation and Language Dependent Phoneme Based Modeling for Tamil Speech Recognition System", Asian Journal of Information Technology 5 (1) : 38-43, 2006.
11. Boersma. P. "Praat, a system for doing phonetics by computer". Glot International 5:9/10. 341-345.
12. Juang B. Rabiner L. Fundamentals of Speech Recognition. Prentice Hall, 1993.