ALGORITMO PARA LA SEGMENTACION AUTOMATICA DE LA PALABRA HABLADA

ING. GUALBERTO HIDALGO
ESCUELA POLITECNICA NACIONAL

ING. TANIA PEREZ
ESCUELA POLITECNICA NACIONAL

RESUMEN

En este artículo se presenta un nuevo método para la segmentación automática de la palabra hablada. La segmentación se realiza solamente a un nivel primario de segmentos periódicos, aperiódicos y oclusivos. El método utilizado trabaja solamente en el dominio del tiempo dejando inalterada la onda original. El criterio para la discriminación entre segmentos periódicos y aperiódicos es la similaridad entre los intervalos entre cruces con cero para los segmentos periódicos y la ausencia de la misma para los segmentos aperiódicos. Para la discriminación entre oclusiones de una parte y segmentos periódicos y aperiódicos de otra parte se usa el nivel energético.

ABSTRACT

In this paper a new method for the segmentation of speech is presentad. This segmentation is performed only at the primary level of periodic, aperiodic and occlusive segments. The method used works only in the time domain leaving the speechwave undistorted. The criterium for the discrimination between periodic and aperiodic segments is the similarity between the intervals between zero-crossings for periodic segments and the absence of the same for aperiodic segments. For the discrimination between occlusions on one side and periodic and aperiodic segments on the other side the energetic level is used.

I. INTRODUCCION

La segmentación automática del lenguaje hablado no es un prerrequisito indispensable para el reconocimiento de la locución, puesto que hay algoritmos que incluyen en el mismo proceso tanto la segmentación como el reconocimiento.
Sin embargo dicha segmentación automática de la palabra hablada puede facilitar significativamente el ulterior proceso de su reconocimiento. Para que se justifique el uso de este algoritmo de preprocesamiento, el mismo debe ser simple y, en lo posible, debe dejar la onda bajo proceso inalterada. El presente algoritmo llena ampliamente estas dos condiciones. De hecho, en un cierto sentido, es la continuación lógica del algoritmo de búsqueda de periodicidad de los mismos autores [1].
La segmentacion del lenguaje hablado puede hacerse a diversos niveles. La segmentación más elemental puede considerarse aquella que toma en cuenta solamente las condiciones más notorias y claramente diferenciables de la palabra hablada, estos es: voceada, no voceada y silencio [2]-[4]. Este tipo de segmentación coincide casi totalmente con la que se presenta en el presente artículo, estos es: cuasi-periódica, aperiódica y oclusión. Otros algoritmos realizan la segmentación a un nivel más detallado [5]-[10]. Existen también algoritmos cuyo objeto es la segmentación de la palabra hablada en unidades de información tipo frase [11].

El presente artículo se estructura como sigue:
En la sección II se explican los principios básicos de la segmentación de la locución. En la sección III se presenta el algoritmo y su diagrama de flujo. Finalmente en la sección IV se da una evaluación del mismo.

II. CONCEPTOS FUNDAMENTALES

Debido a la naturaleza continua del lenguaje hablado, en el cual los fonemas se influencian unos a otros en un proceso conocido como coarticulación, la segmentación, casi en forma inevitable, presenta la siguiente característica. El instante en el cual la onda de voz realiza una transición de un estado a otro no puede ubicarse con matemática precisión. Hay, casi siempre, una región de incertidumbre dentro de la cual puede localizarse la marca de separación entre estados.
Esto no obstante, para una segmentación confiable la condición estacionaria de los diferentes estados en los cuales se segmenta la voz deben estar claramente definidos. Teniendo esto en mente el presente algoritmo ha seleccionado tres clases de estados para la segmentación de la palabra hablada, esto es:

periódico
aperiódico
oclusión

El nivel energético no puede ser un discriminante confiable entre los segmentos periódicos y aperiódicos. De hecho, aunque generalmente hablando, los segmentos periódicos son predominantemente más energéticos que los aperiódicos, ocasionalmente un segmento aperiódico puede ser bastante más energético que un segmento periódico débil. Así pues en el presente artículo, el único criterio utilizado para la discriminación entre segmentos periódicos y aperiódicos es el siguiente:
Si se define T como la duración de un período en una onda periódica, el segmento es periódico si

$$f(t) = f(t+nT) \qquad (2)$$

esto es, T es el intervalo de tiempo después del cual la onda se repite.
La palabra hablada no posee este tipo de periodicidad, en realidad en la palabra

hablada no existen dos períodos iguales. La relación entre dos períodos consecutivos de la locución es de similaridad. Por lo tanto al referirse al lenguaje hablado es mejor hablar de "cuasi-periodicidad" antes que de periodicidad. De acuerdo a esto, la palabra hablada podrá considerarse "cuasi-periódica" si, para dos "cuasi-períodos" consecutivos, tiene vigencia la siguiente relación:

$$f(t) \approx f(t+nT) \qquad (2)$$

Puede decirse que esta regla establece un criterio evanescente para la discriminación entre segmentos periódicos y cuasi-periódicos; ciertamente, sin embargo en la práctica se manifiesta como un criterio aceptable como puede confirmarse por los resultados obtenidos.
Aun cuando es incorrecto hablar de periodicidad en el lenguaje hablado, en adelante, por razones de brevedad, los términos periódico y periodicidad se usarán para significar cuasi-periódico y cuasi-periodicidad, respectivamente.
Para la discriminación de oclusiones de una parte, y segmentos periódicos o aperiódicos de otra parte, un criterio confiable es el nivel energético. En efecto la oclusión que corresponde a los instantes durante los cuales la boca está cerrada (entre palabras o como preparación para una plosiva), es el menos energético de los niveles de locución. Se puede decir que la misma coincide con el ruido ambiental que puede ser muy pequeño en una habitación normal de trabajo si el micrófono utilizado es altamente directivo. Desafortunadamente este criterio falla para voces femeninas para las cuales una fricativa débil es clasificada erróneamente algunas veces como oclusión.

EL ALGORITMO

Como se mencionó anteriormente el presente algoritmo usa los resultados de otro algoritmo que sirve para detectar periodicidad [1].
Basado en dicho algoritmo el programa para la segmentación de la palabra hablada es un típico algoritmo para detectar transiciones entre estados. Este programa primero establece el nivel energético de la señal para decidir si el segmento bajo análisis es una oclusión o no. Como se dijo anteriormente, la oclusión corresponde a los instantes durante los cuales el locutor cierra su boca en preparación para la brusca liberación de energía de una plosiva, o para hacer una pausa entre palabras. Si el nivel energético de la onda esta sobre el umbral del silencio u oclusión, entonces debe decidirse si el segmento es periódico o aperiódico. Esta tarea se realiza utilizando el criterio dado en la relación (2). Esto es, para dos períodos consecutivos de la locución cualquier magnitud o parámetro correspondiente debe ser similar. En el presente análisis el criterio utilizado es la similitud de los intervalos entre cruces con cero de la onda de locución con respecto a un eje corrido sobre el eje normal horizontal trazado a nivel cero, y con una pendiente que sigue aproximadamente las variaciones energéticas supra-periódicas de la locución. Esta primera parte del proceso que coincide con el algoritmo para detectar periodicidad [1], establece si el segmento bajo análisis es periódico, aperiódico o una oclusión. Una vez que se ha realizado esta detección, para completar el proceso de segmentación solamente debe determinarse los instantes en los cuales tiene lugar un cambio de estado en la onda analizada. Esto es, el algoritmo debe trabajar como un detector de transiciones de estados finitos.
En el caso presente existen tres estados: periódico, aperiódico, oclusión. Estos estados se identifican por medio de la variable INDEX que tiene el valor 1 para las oclusiones, 2 para los segmentos aperiódicos y 3 para los segmentos periódicos. Cuando comienza el proceso el algoritmo se encuentra en un pre-estado neutro que tiene el valor de la variable INDEX = 0. Puesto que el algoritmo necesita un cierto tiempo para decidir que clase de onda está siendo procesada, hay siempre un pequeño intervalo de la onda al comienzo que permanece sin clasificar. Tan pronto como se hace la primera decisión el algoritmo entra en alguno de los tres estados.
Teóricamente es posible la transición de un estado a cualquier otro estado, en la realidad la transición segmento periódico-oclusión o su alterna, oclusión-segmento periódico, se observaron muy raramente. En otras palabras, la transición segmento periódico-oclusión o su inversa están casi siempre suavizadas por un intervalo aperiódico. Por lo tanto el diagrama de cambio de estados será el indicado en la fig. 1.

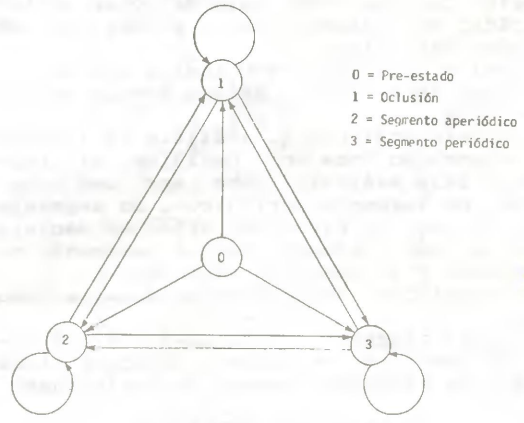

Fig. 1. Diagrama de cambio de estados.

Las transiciones periódico-aperiódico o aperiódico-periódico están precedidas a veces por un intervalo incierto en el cual hay una secuencia de pequeños fragmentos clasificados alternativamente como periódicos y aperiódicos. Igualmente hay ocasiones en las cuales el algoritmo momentáneamente se extravía, y estando muy adentro de un segmento periódico, clasifica algunos intervalos como aperiódicos o viceversa. Estos errores claramente identificables son eliminados al final del proceso con un tratamiento adecuado. Al final del proceso se tienen dos arreglos, uno numérico y otro literal que sirven

para la presentación gráfica de los resultados. El arreglo numérico consta de una secuencia de números que representan el orden (o la ubicación temporal) de la muestra en la cual la transición ocurre. El arreglo literal está formado por una secuencia de las letras: "A", "O", "P". Donde:

"A" indica aperiódico
"O" indica oclusión
"P" indica periódico

Esta secuencia de letras se corresponden con los instantes en que una transición ha tenido lugar, y que están dados por el arreglo numérico anterior.

El diagrama de flujo para la segmentación automática de la locución se da en la fig. 2

EXPLICACION DEL DIAGRAMA DE FLUJO

Como se desprende de la fig. 2 el proceso comienza con la fijación del valor de algunas constantes:

NUMAX = 330
NUMIN = 104 para hombres
DIMIN = 3

NUMAX = 171
NUMIN = 60 para mujeres
DIMIN = 2

El pre-estado se indica con el valor de INDEX = 0
Las constantes NUMAX, NUMIN, DIMIN se utilizan por cuanto el algoritmo para la segmentación automática de la palabra hablada es una continuación y complemento lógico del algoritmo para detectar periodicidad del mismo autor y utiliza las mismas subrutinas.
El valor de INDEX = 0 indica que el algoritmo está en un estado previo a toda decisión.
En cuanto comienza el análisis de la onda el algoritmo toma una decisión, el fragmento bajo análisis debe ser una oclusión, un segmento periódico o un segmento aperiódico. El algoritmo entonces declara que la onda comienza con el segmento apropiado y el análisis continúa.
La transición entre estados decurre como sigue:
a) Si el fragmento bajo análisis se detecta como una oclusión, entonces tiene lugar la siguiente cadena de decisiones:

si el valor previo de INDEX = 0
entonces el algoritmo establece que la segmentación comienza con una oclusión y almacena el orden de la muestra en el cual se supone que comienza la oclusión y la letra "O". La variable INDEX toma el valor 1. El proceso continúa.
si el valor previo de INDEX = 1
entonces el algoritmo decide que el estado de oclusión continúa, nada se cambia ni almacena, El proceso continúa.
si el valor previo de INDEX = 2
entonces el algoritmo decide que ha tenido lugar una transición segmento aperiódico-oclusión. Se almacena el orden de la muestra en la cual se supone que tiene lugar la transición y el carácter "O". El valor de la variable INDEX se fija en 1. El proceso continúa.
Si el valor previo de INDEX = 3
Entonces el algoritmo decide que una transición segmento periódico-oclusión ha tenido lugar. Se almacena el orden de la muestra en la cual se supone que ocurre la transición y el caracter "O". El valor de INDEX se fija en 1. El proceso continúa.

b) Si el segmento bajo análisis se detecta como aperiódico:
si el valor previo de INDEX = 0
el algoritmo decide que la segmentación comienza con un segmento aperiódico y almacena el orden de la muestra en el cual comienza el segmento aperiódico y el literal "A". El valor de INDEX se fija en 2. El proceso continúa
si el valor previo de INDEX = 1
entonces el algoritmo decide que ha tenido lugar una transición oclusión-segmento aperiódico. Se almacena el orden de la muestra en la cual ocurre la transición y el carácter "A", el valor de INDEX se fija en 2. El proceso continúa,
si el valor previo de INDEX = 2
entonces el algoritmo decide que el segmento aperiódico continúa. Nada se cambia, ningún valor se almacena. El proceso continúa.
si el valor previo de INDEX = 3
entonces el algoritmo decide que ha tenido lugar una transición segmento periódico-segmento aperiódico. Se almacena el orden de la muestra en el cual se supone que ocurre la transición y el carácter "A". El valor de INDEX se fija en 2. El proceso continúa.

c) Si el segmento bajo análisis se detecta como periódico:
si el valor previo de INDEX = 0
entonces el algoritmo decide que el proceso comienza con un segmento periódico. Se almacena el orden de la muestra en la cual la transición tiene lugar y la letra "P". El valor de INDEX se fija en 3. El proceso continúa.
si el valor previo de INDEX = 1
entonces el algoritmo decide que ha tenido lugar una transición oclusión-segmento periódico. Se almacena el orden de la muestra en la cual se supone que tiene lugar la transición y la letra "P". El valor de INDEX se fija en 3. El proceso continúa.
si el valor previo de INDEX = 2
entonces el algoritmo decide que ha tenido lugar una transición segmento aperiódico-segmento periódico. Se almacena el orden de la muestra en la cual se supone que ocurre la transición y el carácter "P". El valor de INDEX se fija en 3. El proceso continúa.
si el valor previo de INDEX = 3
entonces el algoritmo establece que el segmento periódico continúa. No se almacena ningún valor, la variable INDEX conserva su valor 3.
Cuando no hay más datos disponibles el arreglo que contiene el orden de las muestras en las cuales ocurren las transiciones y el literal que indica la clase de estado o segmento que comienza en dicho instante, esto es los arreglos NUMERAL y LITERAL$ se procesan para eliminar las pérdidas de estado de corta duración que ocurren durante el proceso de segmentación. Como resultado de este proceso se obtienen los arreglos NUMERAL2 y

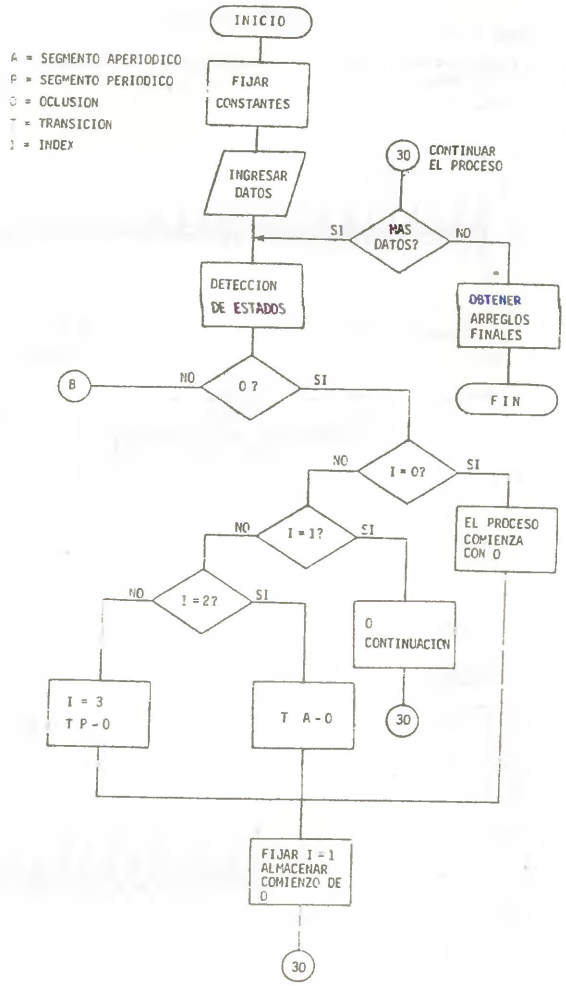

Fig. 2a. Diagrama de flujo del programa de segmentación.

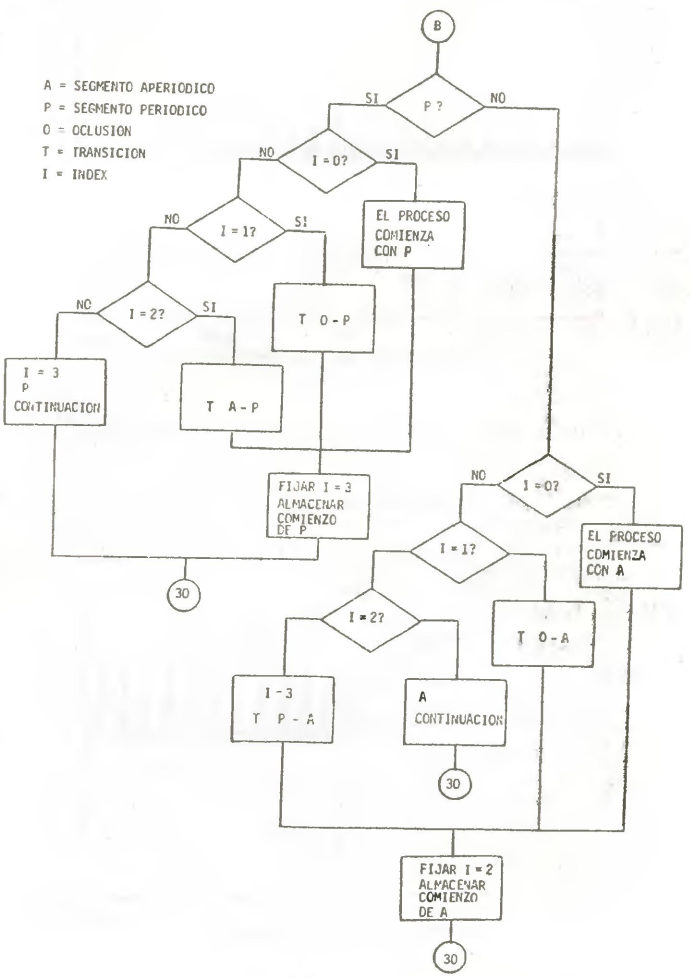

Fig. 2b. Diagrama de flujo del programa de segmentación.

LITER2$ que indican el resultado final de la segmentación de la manera que sigue: por ejemplo los pares
4 P
2020 A
2800 O
3400 A
3800 P, etc.
indican que la onda analizada es:
periódica de la muestra 4 a la 2020
aperiódica de la muestra 2020 a la 2800
oclusión de la muestra 2800 a la 3400
aperiódica de la muestra 3400 a la 3800
periódica de la muestra 3800 en adelante.
Estos pares de datos alfanuméricos se almacenan para la posterior representación gráfica de la onda segmentada.

IV. EVALUACION DEL ALGORITMO

Se consideró que la mejor forma de presentar los resultados era dibujar la onda analizada de la siguiente manera:

Bajo la onda de locución existe una línea con flechas que indican de que muestra a que muestra la onda tiene un determinado estado. Ver la fig. 3 para un ejemplo con voz masculina.

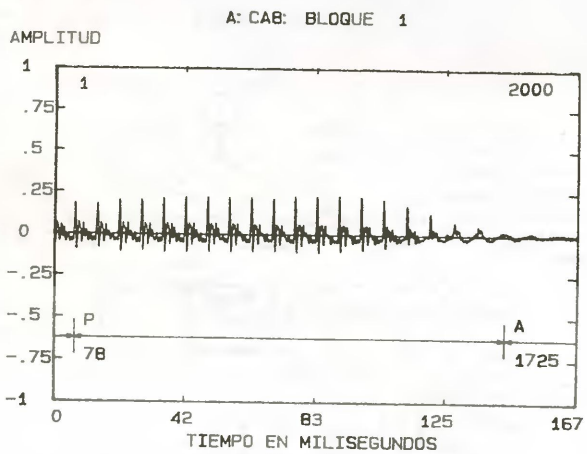

Fig. 3a.

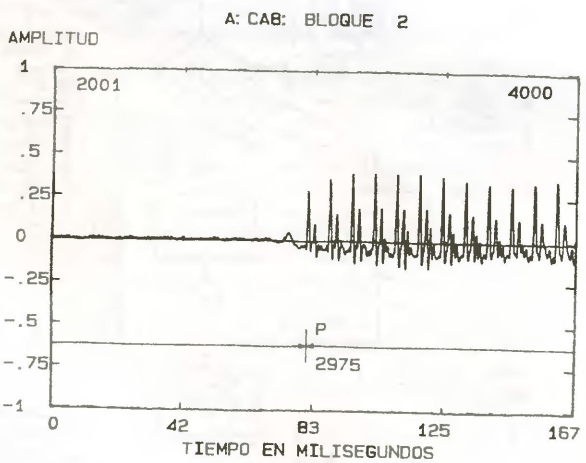

Fig. 3b.

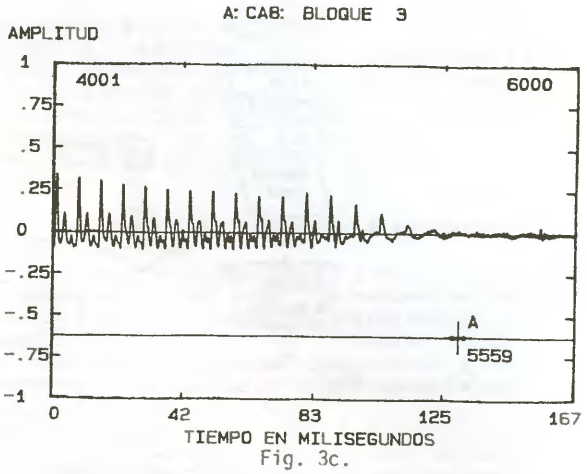

Fig. 3c.

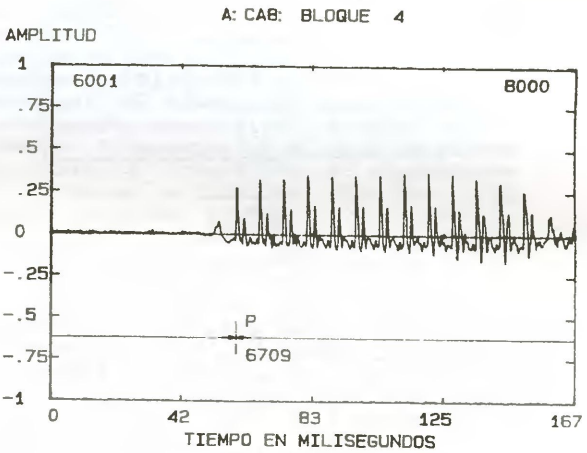

Fig. 3d.

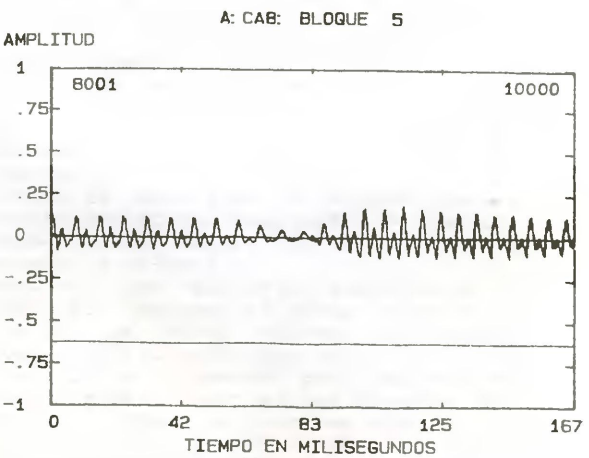

Fig. 3e.

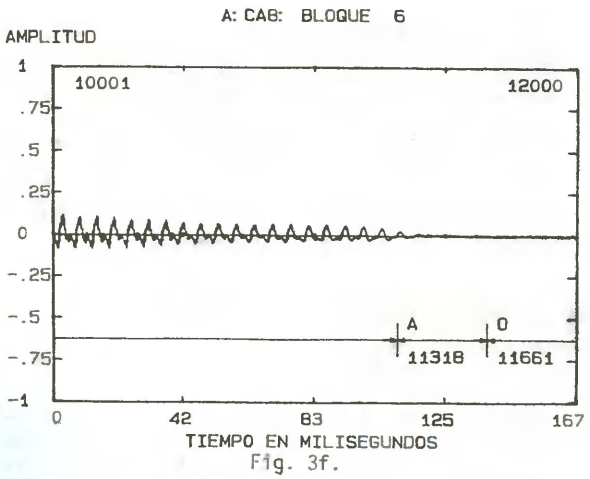

Fig. 3f.

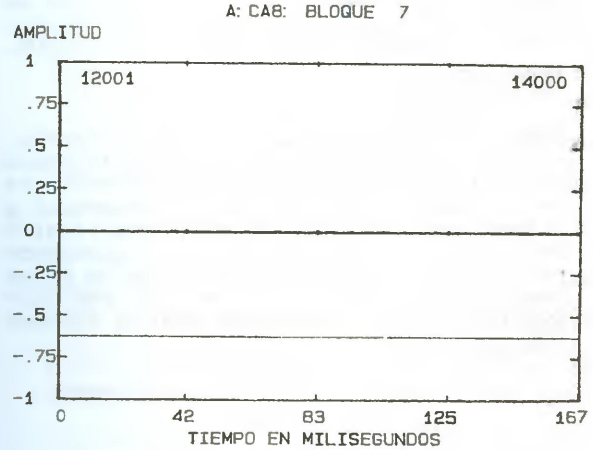

Fig. 3g.

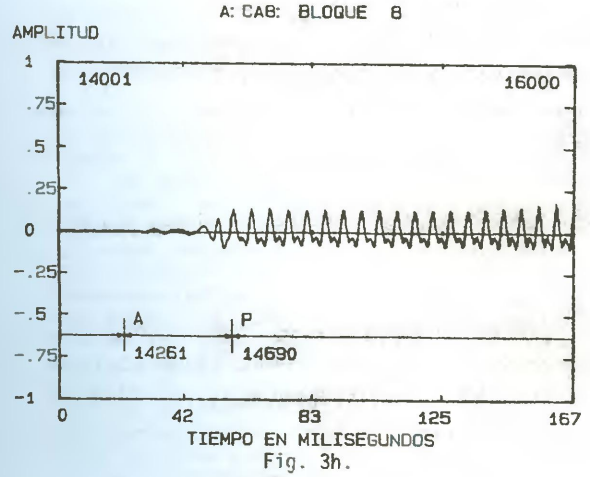

Fig. 3h.

De las voces que se han procesado es posible concluir que el algoritmo trabaja aceptablemente para voces masculinas. Para voces femeninas los errores son más frecuentes, pero se espera que esto podrá mejorarse a través de las conclusiones que se extraigan de la observación de un mayor número de voces femeninas, de las cuales, al momento, hay una cantidad muy restringida.

Para el caso de voces masculinas existen ciertas voces para las cuales el algoritmo trabaja tan bien que las segmentación parece hecha manualmente. Este es el caso de la voz presentada en la fig. 3.

Para captar periodicidad el algoritmo necesita a veces dos o tres períodos. A veces la periodicidad se capta desde el primer período. En general, debido a la incertidumbre que prevalece durante la transición, existe frecuentemente un retardo en la localización de la transición.

Para una evaluación más objetiva de los resultados obtenidos hasta el momento, se elaboró la siguiente tabla de errores con datos correspondientes a 10 voces masculinas:

TABLA DE ERRORES

Segmentación Automática / Columnas — Filas: Segmentación Visual	Segmento periódico	Segmento aperiódico	Oclusión	% Error
Segmento periódico	108	12	0	10%
Segmento aperiódico	0	151	5	3.2%
Oclusión	0	2	50	3.8%

Como puede verse de la tabla anterior el error más frecuente es la confusión periódico-aperiódico, este error debe clasificarse como una substitución, esto significa que el algoritmo clasifica como aperiódico un segmento que es claramente periódico. El otro caso de substitución se da cuando un segmento aperiódico, ordinariamente una fricativa débil, se considera una oclusión. Se espera que este error pueda evitarse ubicando el micrófono más cerca de la boca del locutor. El otro error en el cual una oclusión se clasifica como un segmento aperiódico se categoriza como una omisión, esto significa que el algoritmo ignora de plano una oclusión. Este error se origina probablemente de una ubicación inexacta del comienzo y fin de las oclusiones y puede ser reducido con una determinación más cuidadosa de la duración de las oclusiones.

CONCLUSION

No obstante sus limitaciones el presente algoritmo puede considerarse un punto de partida aceptable para el reconocimiento de la palabra hablada. En realidad para el ulterior procesamiento de la palabra hablada lo que interesa no es la ubicación matemáticamente precisa de los instantes en los que tiene lugar la transición, un objetivo inalcanzable debido a la coarticulación, sino más bien la detección cierta de que una transición ha tenido lugar, ya que este conocimiento permitirá un tratamiento adecuado de la locución. Teniendo en cuenta esto, los retardos en la determinación de los instantes de transición no son errores tan graves como las confusiones de estado, que afortunadamente no son tan frecuentes. Además es de esperarse que las conclusiones extraídas de los resultados obtenidos hasta el momento, permitirán una optimización del algoritmo.

Referencias.

[1] G. E. Hidalgo, T. Pérez, "Un algoritmo para la detección de periodicidad sin prefiltrado." XII Jornadas en Ingeniería Eléctrica y Electrónica, 1991.

[2] C. K. Un and H. H. Lee, "Voiced/ unvoiced/ silence disrimination of speech by delta modulation," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, pp. 398-407, Aug. 1980.

[3] B. V. Cox and L. K. Timothy,"Nonparametric Rank-Order Statistics Apllied to Robust Voiced-Unvoiced-Silence Classification," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, pp. 550-561, Oct. 1980.

[4] R. J, DiFrancesco,"Real-Time Speech Segmentation Using Pitch and Convexity Jump Models: Application to Variable Rate Speech Coding," IEEE Trans. Acoust., Speech, Signal Processing, Vol 38, pp. 741-748, May 1990.

[5] R. A. Cole and L. Hou, "Segmentation and Broad Classification of Continuous Speech," ICASSP 88, New York, S10.12, pp. 453-456.

[6] J. R. Glass and V. W. Zue,"Multi-Level Acoustic Segmentation of Continuous Speech," ICASSP 88, New York, S10.6, pp. 429-432.

[7] K. L. Brown, V. R. Algazi,"Characterization of Spectral Transitions with Applications to Acoustic Sub-Word Segmentation and Automatic Speech Recognition," ICASSP 89, Glasgow, S3.7, pp. 104-107.

[8] V. Zue, J. Glass, M. Phillips, and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the Summit System," ICASSP 89, Glasgow, S.8.1, pp. 389-392.

[9] R. Andre-Obrecht, "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 36, pp. 29-40, January 1988.

[10] K. Atazaki, Y. Komori, T Kawabata and K. Shikano,"Phoneme Segmentation Using Spectrogram Reading Knowledge,"ICASSP 89, Glasgow, S8.2, pp. 393-396.

[11] D. Huber, "A Statistical Approach to the segmentation and Broad Classification of Continuous Speech into Phrase-Sized Information Units," ICASSP 89, Glasgow, S12.1, pp. 600-603.

HIDALGO, GUALBERTO

Ingeniero en Electrónica y Telecomunicaciones graduado en la Escuela Politécnica Nacional en el año de 1974. Obtuvo el título de Master en Ingeniería de Comunicaciones en el Imperial College de Londres, Inglaterra, en 1979. Actualmente se desempeña como profesor a tiempo completo en la Escuela Politécnica Nacional y es estudiante externo de la Universidad de Londres. Dirige el PROYECTO CONUEP 88-01, RECONOCIMIENTO DE FONEMAS POR COMPUTADOR.

PEREZ, TANIA

Ingeniero en Electrónica y Telecomunicaciones graduada en el Instituto Bonch Bruyevich, Leningrado, 1977. Actualmente se desempeña como profesora principal a tiempo completo en la Escuela Politécnica Nacional y realiza estudios de postgrado en Computación e Informática en la misma Institución. Participa en el PROYECTO CONUEP 88-01, RECONOCIMIENTO DE FONEMAS POR COMPUTADOR.