



# Medida de Distorsión ARMA. Aplicación a la segmentación de señales del habla natural

F. Martínez<sup>5</sup>, A. Guillamón<sup>6</sup>, A. García<sup>7</sup>, F. González<sup>8</sup>

## RESUMEN

En el presente trabajo se presentan los resultados obtenidos mediante un sistema de segmentación propio totalmente automático de señales del habla natural implementado en el entorno de programación LabView. El sistema propuesto realiza la segmentación de manera secuencial utilizando dos criterios:

- (i) *el pitch* (intentando identificar señales procedentes de sonidos sonoros),
- (ii) *una medida de distorsión* basada en la modelización ARMA de la envolvente espectral (para identificar o segmentar aquellos sonidos reconocidos o descartados en la primera etapa).

Las señales del habla natural sobre las que se ha aplicado el algoritmo de segmentación desarrollado, corresponden a diferentes registros de la *Base de Datos Ahumada*, [3].

**Palabras Claves:** Segmentación de señales, Parametrización ARMA, Análisis Cepstral, Reconocimiento de voz.

## ABSTRACT

This paper introduces the results obtained by means of an automatic segmentation procedure of natural speech signals developed in LabView. The system that it proposes makes the segmentation procedure in two stages using the following criterions:

- (i) Pitch (this criterion is used to look for frames of signals form voiced sounds).
- (ii) A "distortion measure" based on ARMA parametrization of the power spectrum (this criterion is applied to identify or segment the frames obtained in first stage).

The experiments have been computed over signals from a specific speech Spanish database (AHUMADA Database) designed for speaker recognition tasks in Castilian Spanish

**Keywords:** Voice recognition, signals segmentation, ARMA parametrization, cepstral anaysis, speech regognition.

---

<sup>5</sup> Profesor Titular de Universidad. Area de Matemática Aplicada. Departamento de Matemática Aplicada y Estadística. Grupo de investigación Modelos y Sistemas para el Procesado de señales y Series Temporales. Universidad Politécnica de Cartagena (España)

<sup>6</sup> Profesor Titular de Universidad. Area de Estadística e Investigación Operativa. Departamento de Matemática Aplicada y Estadística. Grupo de investigación Modelos y Sistemas para el Procesado de señales y Series Temporales. Universidad Politécnica de Cartagena (España)

<sup>7</sup> Catedrático de Escuela. Area de Matemática Aplicada. Departamento de Matemática Aplicada y Estadística. Grupo de investigación Modelos y Sistemas para el Procesado de señales y Series Temporales. Universidad Politécnica de Cartagena (España)

<sup>8</sup> Ingeniero en Automática y Electrónica Industrial. Grupo de investigación Modelos y Sistemas para el Procesado de señales y Series Temporales. Universidad Politécnica de Cartagena

## 1. INTRODUCCIÓN

Son muy diversos los campos en el tratamiento de señales, en los que la *segmentación* representa una fase previa prácticamente imprescindible; entendiendo por segmentación *la división de las señales en partes discretas, que pueden ser asociadas con un evento específico de los que componen tales señales* [1]. En el caso concreto de las señales del habla, el problema de la segmentación presenta grandes dificultades de partida, ya que en la producción del habla natural cambia continuamente la configuración del tracto vocal y sus modos de excitación. Un sistema de segmentación de señales del habla, interpretado como procedimiento para dividir la señal en partes sostenidas donde las características del sonido presentan una cierta regularidad y en partes transicionales donde las características varían con el tiempo; ha de incluir requisitos tales como [2]:

- capacidad de distinguir suficientemente cambios pequeños en las características de los sonidos
- la utilización de características del sonido que puedan obtenerse fácilmente, a partir de la señal original.

Tradicionalmente, en el campo del análisis de las señales del habla, se han utilizado sistemas semiautomáticos de segmentación que trabajan con ventana fija de corta duración a fin de asumir la hipótesis de estacionariedad de la señal en discurso continuo. El uso de este tipo de ventana presenta grandes inconvenientes sobre todo en el análisis y posterior modelización de transiciones entre sonidos de diversa naturaleza.

El sistema desarrollado que se presenta en este trabajo, utiliza secuencialmente dos criterios para segmentar de manera automática una señal temporal grabada. En una primera fase se realiza un barrido previo con ventana y paso prefijado, obteniéndose *la variación del pitch* a lo largo de la señal y los distintos tramos se agrupan atendiendo a la variabilidad del pitch. Los tramos temporales obtenidos se corresponderán con sonidos puros, tanto sonoros como sordos.

El problema fundamental que presenta un sistema de segmentación con las anteriores características, vuelve a ser el estudio de los tramos que agrupan transiciones. Con el fin de solventar este problema, se plantea el estudio de estos tramos temporales desde una perspectiva diferente. En una segunda etapa, a partir de la envolvente espectral ARMA se construye una medida de similitud entre fragmentos de longitud fija (*medida de distorsión ARMA*) con el fin de agrupar o discriminar la hipótesis de que ambos segmentos provengan de un mismo sonido, obteniéndose, a través de este último criterio, la segmentación definitiva de la señal temporal. Se han utilizado como señales test, registros obtenidos de la *Base de Datos Ahumada* correspondientes a diferentes señales y locutores.

Seguidamente, mediante una adaptación del *AIC (Akaike's information-theoretic criteria)* a la modelización ARMA, abordamos el problema de la determinación del número óptimo de parámetros (*polos y ceros*) que se deben utilizar.

Se presentarán algunos resultados obtenidos mediante la aplicación del algoritmo de segmentación elaborado sobre diversos registros de la citada base de datos, [3]. Se han querido incluir entre los registros analizados diversas situaciones del habla natural, que desde el punto de vista fonético entrañan un notable grado de complejidad.

Por último, se destacan las principales conclusiones que se pueden entresacar de los resultados presentados.

## 2. MATERIAL Y MÉTODOS

### Frecuencia fundamental o pitch

El análisis cepstral, introducido por Bogert y otros [4], fue aplicado en sus inicios al estudio de señales sísmicas, debido a su capacidad para representar por separado los efectos de la fuente de una señal y de los ecos producidos por el sistema de transmisión de la misma.

Dada una señal temporal  $s(n)$ , se puede definir su *cepstrum*  $C(t)$  como *el espectro de potencia del espectro de potencia logarítmico de la señal*, es decir:

$$C(t) = \left| \mathbf{F} \left( \log | \mathbf{F}(s(n)) |^2 \right) \right|^2 \quad (1)$$

En el presente trabajo se considera la señal modelizada por un proceso  $ARMA(p,q)$ , cuya ecuación en diferencias viene dada por

$$s(n) = \sum_{k=1}^p a_k \cdot s(n-k) + G \cdot \sum_{l=0}^q b_l \cdot u(n-l) \quad (2)$$

Denotando por  $H(z)$ , a la función de transferencia y  $S(z)$  y  $U(z)$  a la *transformada-z* de las señales de salida y entrada se tendrá:

$$H(z) = \frac{S(z)}{U(z)} \Rightarrow S(z) = U(z) \cdot H(z)$$

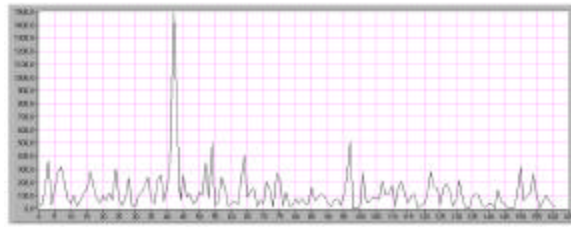
y por tanto:

$$\mathbf{F}(\log |S(z)|^2) = \mathbf{F}(\log |U(z)|^2) + \mathbf{F}(\log |H(z)|^2) \quad (4)$$

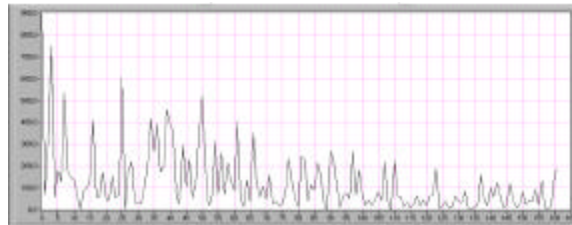
Dado que el espectro de potencia logarítmico es una función par, se puede obtener el cepstrum de una señal como *el cuadrado de la transformada coseno del espectro logarítmico de potencia de la señal*, es decir:

$$C(t) = \left\{ \int_0^\infty \log |S(z)|^2 \cos(zt) dz \right\}^2 \quad (5)$$

Precisamente, su poder para separar los efectos de la fuente, primer sumando de la expresión (4) y la transmisión, segundo sumando, es lo que hace que análisis cepstral resulte fundamental a la hora de distinguir entre sonidos sonoros y sordos [5]. Así, en sonidos vocálicos, su cepstrum característico presentará en la parte baja de las *quefrecias* los formantes y posteriormente aparecerá un pico de amplitud muy superior a los demás, que representará el periodo de vibración de las cuerdas vocales o pitch. Es una característica representativa de sonidos no sonoros, como consonantes fricativas, la ausencia de este pico. Las siguientes gráficas muestran esta característica:



**Figura 1.- Cepstrum en sonidos sonoros**



**Figura 2. Cepstrum en sonidos sordos**

### Envolvente ARMA

Actualmente, en muchas aplicaciones del procesamiento de señales, como puede ser el procesamiento del habla, una señal  $s(n)$  puede ser modelada como un proceso ARMA, cuya ecuación en diferencias hemos presentado en (2).

Teniendo en cuenta este modelo, si se asume que la señal  $s(n)$  es generada por la excitación del sistema ARMA mediante un impulso o un ruido blanco, la función de transferencia  $H(z)$  vendrá dada por:

$$H(z) = G \frac{1 + \sum_{l=1}^q b_l \cdot z^{-l}}{1 - \sum_{k=1}^p a_k \cdot z^{-k}} \quad (6)$$

Denotando por  $E$  el error cometido al aproximar el espectro de potencia de la señal ( $P(\mathbf{w})$ ) y la envolvente espectral ARMA ( $\bar{P}(\mathbf{w})$ ), [6], se prueba que

$$E = \frac{1}{2p} \int_{-p}^p P(\mathbf{w}) \cdot \frac{D(\mathbf{w})}{N(\mathbf{w})} \cdot d\mathbf{w} \quad (7)$$

donde

$$N(\mathbf{w}) = \left| 1 + \sum_{l=1}^q b_l \cdot e^{-j l \mathbf{w}} \right|^2$$

$$D(\mathbf{w}) = \left| 1 - \sum_{k=1}^p a_k \cdot e^{-j k \mathbf{w}} \right|^2$$

Dado que el problema de la minimización del error  $E$  conduce a un sistema de  $p+q$  ecuaciones no lineales, éste ha de ser resuelto iterativamente. Muchos son los métodos de la teoría de la optimización [7], que pueden ser aplicados a este problema; de entre los cuales en este trabajo hemos optado por el

propuesto en [8], que consiste en un método general para la obtención de una estimación inicial de un proceso  $ARMA(p,q)$ , suponiendo el cómputo de los parámetros del modelo en dos etapas:

- (i) *estimación de los parámetros AR*,  $\{a_k, 1 \leq k \leq p\}$
- (ii) *estimación de los parámetros MA*,  $\{b_l, 1 \leq l \leq q\}$

Veamos más detalladamente el proceso realizado.

- (i) Estimación de los parámetros AR.: El cálculo de los *parámetros AR*,  $\{a_k, 1 \leq k \leq p\}$ , se ha llevado a cabo a partir de la matriz de autocorrelación asociada a la señal mediante el método de Durbin [6, 8].
- (ii) Estimación de los parámetros MA.: A partir de las autocovarianzas  $c_j$  de la señal  $s(n)$ ,

$$c_j = \frac{1}{N} \sum_{n=0}^{N-j} \left( s(n) - \bar{s} \right) \left( s(n+j) - \bar{s} \right) \quad (8)$$

donde:

$$\bar{s} = \frac{1}{N} \sum_{n=0}^{N-1} s(n) \quad (9)$$

siguiendo el modelo propuesto en [7], se calcula la secuencia de *covarianzas modificadas* que denotaremos por  $c'_j$ ,

$$c'_j = \begin{cases} \sum_{i=0}^p \sum_{k=0}^p a_i \cdot a_k \cdot c_{|j+i-k|} & p > 0 \\ c_j & p = 0 \end{cases} \quad (10)$$

donde  $j=0,1,\dots,q$ ,  $a_0=1$ . Es inmediato demostrar que la *autocovarianza* de la señal  $s(n)$  en un intervalo finito  $0 \leq n \leq N-1$ , satisface:

$$\sum_{k=1}^p a_k \cdot c_{|q+i-k|} = c_{q+i}$$

A partir de estos valores, los parámetros *MA*,  $b_l, 1 \leq l \leq q$ , se obtienen mediante un algoritmo iterativo tipo Newton-Raphson, [8]. Así, para cada iteración  $i$  se sigue el siguiente esquema:

$$\mathbf{t}^{i+1} = \mathbf{t}^i - (\mathbf{T}^i)^{-1} \mathbf{f}_i \quad (11)$$

donde:

$$\mathbf{t} = (t_0, t_1, \dots, t_q),$$

$$\mathbf{f}^* = (f_0, f_1, \dots, f_q)$$

$$f_j = \sum_{k=0}^{q-j} t_k \cdot t_{k+j} - c'_j,$$

para  $j=0,1,\dots$ , siendo:

$$T = \begin{pmatrix} \mathbf{t}_0 & \mathbf{t}_1 & \dots & \mathbf{t}_q \\ \mathbf{t}_1 & \dots & \mathbf{t}_q & 0 \\ \dots & \dots & 0 & \dots \\ \mathbf{t}_q & 0 & \dots & 0 \end{pmatrix} + \begin{pmatrix} \mathbf{t}_0 & \mathbf{t}_1 & \dots & \mathbf{t}_q \\ 0 & \mathbf{t}_0 & \dots & \mathbf{t}_{q-1} \\ \dots & 0 & \dots & \dots \\ 0 & \dots & 0 & \mathbf{t}_0 \end{pmatrix}$$

Suelen utilizarse como valores iniciales para el proceso:

$$\mathbf{t}_0 = \sqrt{c'_0}, \mathbf{t}_1 = \mathbf{t}_2 = \dots = \mathbf{t}_q = 0$$

Cuando se verifica  $|f_j| < \varepsilon$ ,  $j=0,1,\dots,q$ , para algún umbral  $\varepsilon$  prefijado, el proceso iterativo se detiene y los parámetros *MA* se obtienen a partir de:

$$b_l = \frac{\mathbf{t}_l}{\mathbf{t}_0} \quad (12)$$

para  $l=1,2,\dots,q$ , siendo la ganancia del modelo *ARMA*:

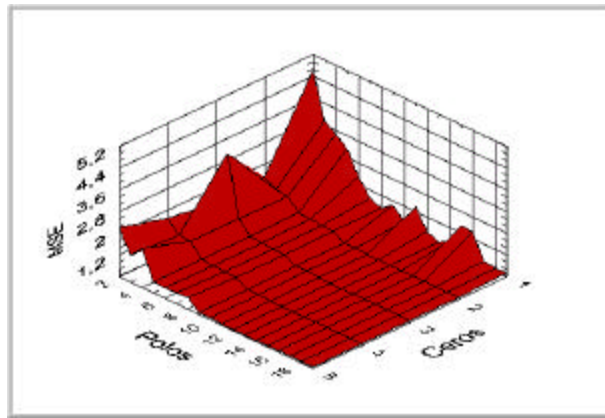
$$G^2 = N \cdot \mathbf{t}_0^2 \quad (13)$$

Para finalizar, indicar que se utiliza el método propuesto por Schür-Cohn, [10], con el fin de garantizar la estabilidad del sistema obtenido. Este método aporta una condición necesaria y suficiente para asegurar la estabilidad, a partir de los llamados *coeficientes PARCOR*, sin necesidad de calcular las raíces del polinomio característico asociado a los parámetros AR.

### Número óptimo de polos y ceros

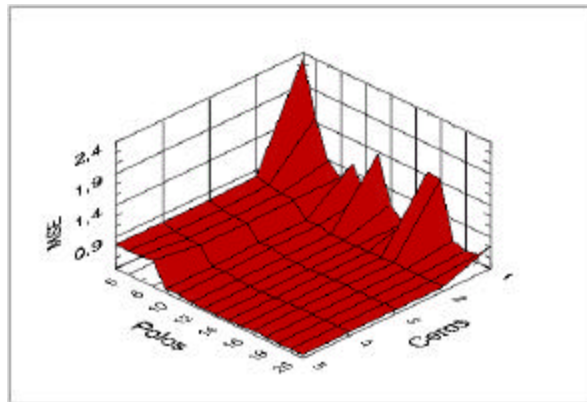
La selección del orden de un modelo ARMA presenta siempre mayor dificultad que para un modelo AR, ya que se ha de elegir un orden para el numerador y otro para el denominador. En numerosas aplicaciones como, por ejemplo, en el campo de las señales del habla, el orden del denominador es más crítico que el del numerador, al determinar el primero el número de modos o frecuencias naturales presentes en la señal. En este trabajo, para la selección del orden de un modelo ARMA, se propone una adaptación del AIC (Akaike's information-theoretic criteria) desarrollado para modelos AR, [7].

En la Figura 3 se representa el MSE (Mean Square Error) variando el número de polos desde 1 a 17 y el número de ceros desde 0 (modelo AR) a 4. Para un número de ceros mayor que 4 resulta un filtro inestable.



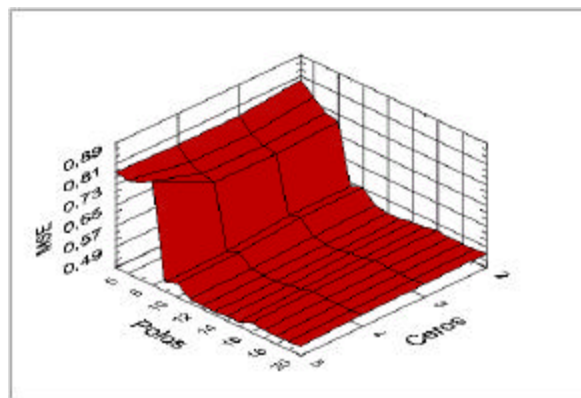
**Figura 3. Análisis del MSE para AR y ARMA y diferente número de polos y ceros en la vocal “a”.**

Con el fin de poder apreciar los cambios relativos al variar el número de polos, en la Figura 4, se parte de un número de polos  $p=6$ . Se observa que un aumento del número de coeficientes a partir de  $p=10$  no proporciona una disminución significativa del MSE.



**Figura 4. Análisis del MSE para ARMA y diferente número de polos y ceros en la vocal “a”.**

Si se representa esta misma gráfica a partir de un número de ceros  $q=2$ , se observan mejor los posibles cambios en el MSE respecto a la variación del número de ceros. En la Figura 5 se puede observar que estas variaciones no resultan significativas.



**Figura 5. Análisis del MSE para ARMA y diferente número de polos y ceros en la vocal “a”.**

Aplicando el anterior criterio para la selección del orden de un modelo ARMA sobre distintas situaciones fonológicas presentes en el habla natural, se ha optado por trabajar con el modelo ARMA(12,2) para la parametrización de registros de 32 ms de duración.

### Medida de Distorsión

El segundo de los criterios utilizados tiene como finalidad verificar la hipótesis de que dos segmentos contiguos, provengan de un mismo sonido.

Así, sean las señales  $s_1(n)$ ,  $s_2(n)$  en el intervalo  $0 \leq n \leq N-1$ , con parametrizaciones:

$$\begin{aligned}\mathbf{a}^1 &= (\mathbf{a}_0^1, \mathbf{a}_1^1, \mathbf{a}_2^1, \dots, \mathbf{a}_p^1) \\ \mathbf{a}^2 &= (\mathbf{a}_0^2, \mathbf{a}_1^2, \mathbf{a}_2^2, \dots, \mathbf{a}_p^2)\end{aligned}$$

Siguiendo el modelo propuesto en [11, 12], se puede definir una medida de similitud entre dos segmentos de la señal a partir de:

$$D = \frac{\hat{E}^{(p)}}{E^{(p)}} \quad (14)$$

donde por  $E^{(p)}$  se representa el error cuadrático medio asociado a la señal  $s_1(n)$ , con su parametrización y por  $\hat{E}^{(p)}$  el error cuadrático medio asociado a la señal  $s_2(n)$  con la parametrización de  $s_1(n)$ . Puesto que este cociente es siempre  $\geq 1$  con el fin de construir una “medida” entre las dos señales, definiremos la *medida de distorsión ARMA* como:

$$d(s_1, s_2) = \frac{1}{2} \cdot (d(s_1, s_2) + d(s_2, s_1)) \quad (15)$$

donde:

$$d(s_1, s_2) = D(s_1, s_2) - 1 \quad (16)$$

En resumen, la aplicación  $d(s_1, s_2)$  entre las señales  $s_1(n)$  y  $s_2(n)$ , satisface:

- (i)  $d(s_1, s_2) \geq 0$ , para cada par de señales  $s_1, s_2$
- (ii)  $d(s_1, s_2) = d(s_2, s_1)$ , para cada par de señales  $s_1, s_2$

Si bien esta aplicación no verifica la desigualdad triangular, si nos da una idea de la similitud o diferencia entre dos señales o tramos temporales de una misma señal y, por tanto, parece adecuada su utilización como criterio con el fin de decidir si dos señales o tramos provienen de un mismo sonido.

### 3. RESULTADOS

El sistema de segmentación de señales propuesto en este trabajo, se ha implementado utilizando el entorno de programación gráfica LabView de National Instruments, buscando poder disponer de una herramienta de trabajo de fácil acceso y manejo por parte del usuario. El software que se ha desarrollado permite en su pantalla de salida, que el usuario pueda controlar los parámetros fundamentales implicados en las dos etapas del proceso de segmentación. Asimismo, en dicha pantalla se presentan bien gráficamente, bien numéricamente, los resultados del procesamiento de señales a través de este algoritmo. En este sentido cabe destacar que el programa ofrece la posibilidad de estudiar el número óptimo de parámetros ARMA (polos y ceros) para la modelización de los tramos temporales en la segunda etapa del sistema de segmentación.



El algoritmo presentado se ha aplicado sobre un gran número de registros de señales del habla de la base de datos Ahumada, buscando trabajar en contextos y determinados eventos del habla natural, que supongan situaciones de severa dificultad para cualquier sistema de segmentación. A modo de ejemplo, se presentan dos pantallas de salida del programa desarrollado, cuya elección obedece a que representan situaciones de seria dificultad en el habla natural. Así, en la Figura 6, se muestra la pantalla de salida del programa desarrollado, que corresponde al registro "...siete seis...". En ella se observan los fragmentos resultantes al aplicar el algoritmo.

Se ha considerado el registro "...siete seis...", por dos motivos distintos. Por un lado, supone un estándar en el procesado de señales acústicas y, por otro, presentan dipongos y diferentes sonidos sordos.

En la Figura 7 se recoge el resultado de la segmentación del registro "... bien el salón ..." procedente de la misma base de datos. Esta señal presenta dos situaciones de interés, por un lado, se pueden observar varias transiciones vocal-consonante nasal (*acoplamiento nasal*) y, por otro, y quizás más interesante, la presencia de una situación más compleja, un diptongo seguido de una consonante nasal, [13].

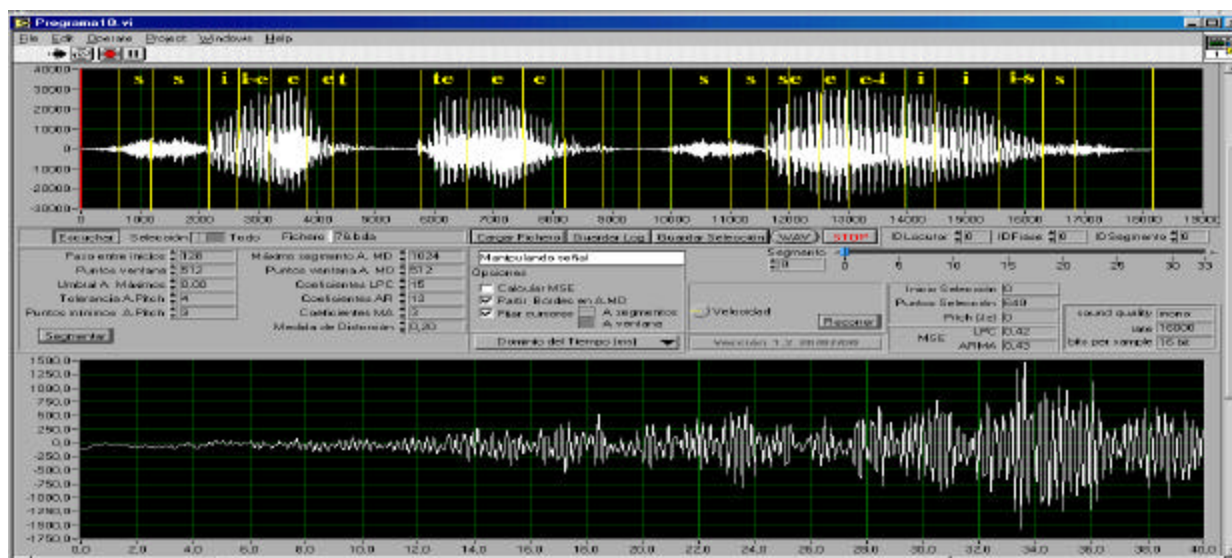


Figura 6.- Segmentación de la señal "siete seis".

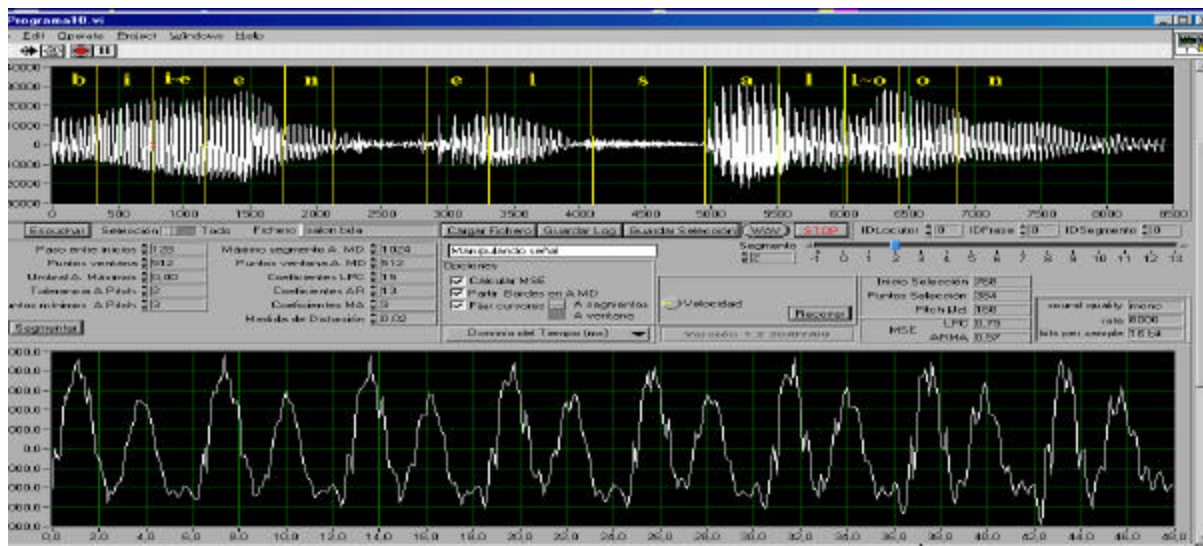


Figura 7.- Segmentación de la señal “bien el salón”.

#### 4. CONCLUSIONES

Los resultados presentados en este trabajo ponen de manifiesto la adecuación del algoritmo de segmentación de señales del habla natural presentado, que utilizando los parámetros: *Nivel de Pitch* y *Medida de la Distorsión ARMA* entre señales, permite una satisfactoria agrupación de los segmentos *de longitud variable*, tanto para sonidos sostenidos como en las transiciones entre sonidos del habla de distinta naturaleza, tales como los fricativos o en eventos del habla en los que esté implicada la estructura *vocal - consonante nasal*.

También se puede destacar que el método propuesto realiza una segmentación adecuada de señales del habla natural, proporcionando tramos temporales correspondientes a sonidos puros, pilar básico para el desarrollo de todo sistema ASRS (Automatic Speaker Recognition System).

Para finalizar, a partir de la adaptación a la modelización ARMA del criterio AIC, se ha determinado de manera empírica la parametrización ARMA óptima para el tratamiento de señales del habla natural de 32 ms de duración, siendo ésta de 12 parámetros AR y 2 parámetros MA, es decir, un modelo ARMA(12,2) resulta óptimo para señales del habla de esta duración.

#### REFERENCIAS

1. Reddy D.R., "Speech Recognition by Machine: A Review", The Journal of The Acoustical Society of America, vol. 64, pp.501-531, 1976.
2. Reddy D.R., "Segmentation of Speech Sounds", The Journal Of The Acoustical Society of America, vol. 40, Number 2, 1966.
3. García Jiménez R.Y. y Díaz Gómez J.L., "Base de Datos para Identificación y Clasificación de Locutores", Politécnica de Madrid, Proyecto Fin de Carrera, 1997.
4. Bogert B..P. y otros, Proceeding of the Symposium on Time Series Analysis. John Wiley & Sons, Inc. Capítulo 5, 1963.
5. Rabiner L.R., Schafer R.W., "Digital Processing of Speech Signals", Prentice- Hall, Capítulo 8, 1988.
6. Makhoul J., "Linear Prediction: A Tutorial Review", Proceeding of the IEEE, vol. 63, pp. 99- 118, 1975.

7. Therrien C.W., "*Discrete Random Signals and Statistical Signals Processing*", Prentice-Hall, 1992.
8. Box G. and Jenkins G. "*Time series analysis: forecasting and control*", Prentice-Hall, 1986.
9. Wilson G.T., "Factorization of the generating function of a pure moving average process", SIAM Jour. Num. Analysis, Vol. 6, 1969.
10. Proakis J.G., Manolakis D.G., "*Tratamiento digital de señales*", Prentice-Hall, 1998.
11. Martínez F. García-Sánchez A. y Peñalver D., "*Segmentación de señales mediante la medida de la Distorsión L.P.C.*", Caseib2000.
12. Rabiner L., Juang B.H., "*Fundamentals of Speech Recognition*", Prentice-Hall, 1993.
13. Su L.-S., Li K.-P., Fu K.S., "*Identification of speakers by use of nasal coarticulation*", The Journal of The Acoustical Society of America, vol.56, pp. 1876-1882, 1974.