OBTENCIÓN DE SONOGRAMAS CON LOS FORMANTES REALZADOS USANDO EL MÉTODO DE PREDICCIÓN LINEAL (LPC)

[1] Jesús Bobadilla, [2] Pedro Gómez y [1] Jesús Bernal

[1]
Departamento de Informática Aplicada
Escuela Universitaria de Informática
Ctra. de Valencia Km. 7, 28031 Madrid
Tel: +34.91.3367862, Fax: +34.91.3367527
e-mail: jbobi@eui.upm.es, jbernal@eui.upm.es

[2]
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n,
Boadilla del Monte, 28660 Madrid
Tel: +34.91.3367384, Fax: +34.91.3367412
e-mail: pedro@pino.datsi.fi.upm.es

RESUMEN

Una correcta determinación de la posición y evolución de los formantes del habla ayudaría enormemente a avanzar en diversos campos de la ciencia, como la fonética acústica o el reconocimiento y síntesis de la voz. Las características del método de Predicción Lineal (LPC) se adaptan de forma natural a la obtención de formantes, por lo que en este artículo se presentan resultados conseguidos partiendo de algoritmos propios diseñados tomando como base LPC. Los sonogramas habituales (basados en la Transformada de Fourier) reflejan peores resultados en la determinación de la posición y evolución de los formantes.

ABSTRACT

The correct estimation of the speech formant positions and evolution would help in several fields of the speech sciences, such as acoustic phonetics or speech recognition and synthesis. Linear Prediction (LPC) features naturally fit to the formant extraction. This article shows the results obtained starting from original algorithms designed using LPC. The usual spectra (based on the Fourier Transform) show worse results than LPC in establishing the formant positions and evolution.

1. INTRODUCCIÓN

En este apartado se hará un especial énfasis en mostrar las razones que justifican la importancia que tiene la correcta determinación de la posición de los formantes del habla en el reconocimiento de la voz.

El sonido del habla puede ser modelado como la respuesta del tracto vocal a una serie de pulsos. Las frecuencias de resonancia se manifiestan en el espectro con energía máxima. Se les denomina formantes y constituyen una información de vital importancia en el reconocimiento del lenguaje hablado [QUI93].

Según se establezcan el punto y modo de articulación de los órganos bucales, obtendremos diferentes valores de frecuencias en los formantes. Cuanto mayor sea el abocinamiento de la boca, menor será la frecuencia del segundo formante (F2) [MAR94]. Cuanto mayor sea la apertura bucal, mayor será la frecuencia del primer formante (F1). La determinación de los formantes es fundamental en la modelización del habla, y por lo tanto, la obtención de un método automático de cómputo que los calcule, resulta de gran interés en los campos de síntesis y reconocimiento de voz [PIC95], [MAR90b].

En muchos casos, el reconocimiento automático del habla ha sido abordado mediante el uso de técnicas de aprendizaje paramétrico, normalmente cadenas de Markov o redes neuronales [FRE93], [TOH92], [RAB89]. Los parámetros utilizados son usualmente los coeficientes LPC o los valores de una FFT [RAB93], [ROW92]. La calidad de los resultados varía según las técnicas empleadas y los objetivos deseados (uno o varios hablantes, lenguaje conexo o palabras aisladas, tamaño del conjunto de aprendizaje, etc.) [CAS87], [CAS90], pero en todos estos casos, el problema de fondo planteado radica en la distancia conceptual existente entre los parámetros proporcionados y los sonidos que se desea reconocer, distancia que se pretende cubrir mediante los métodos de aprendizaje automático existentes.

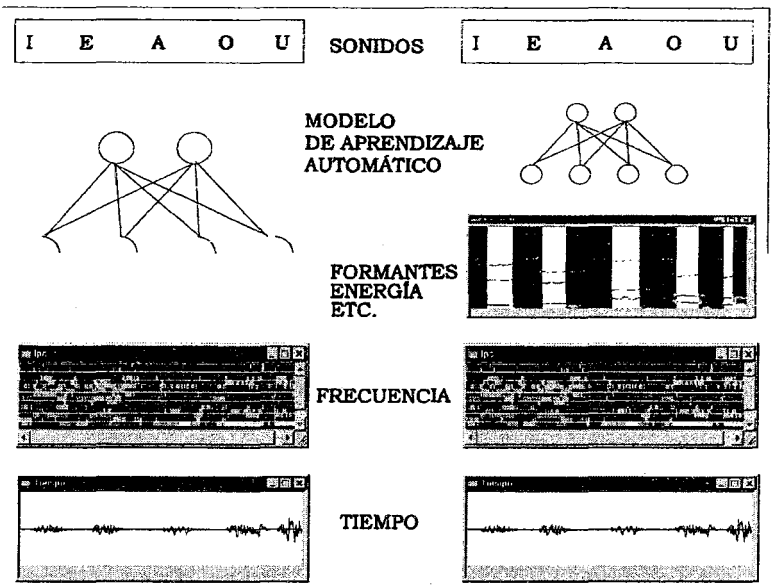


Figura 1.a *Figura 1.b*

Niveles usuales empleados en el reconocimiento del habla, cargando la importancia del proceso en el método de reconocimiento a), o en el cálculo de características de nivel medio b).

El cálculo de formantes proporciona una información de gran valor situada a medio camino entre los parámetros y los sonidos, con lo que la complejidad de las cadenas de Markov o redes neuronales empleadas puede ser reducida de forma sustancial. La figura 1 ilustra esta idea.

Desde un punto de vista computacional, la opción b) de la figura 1, requiere una etapa complementaria, aunque por otra parte, los modelos de aprendizaje automático se reducen. En cualquier caso, el objetivo es conseguir resultados (convergencia) en la etapa de aprendizaje, demasiado complicada en el caso a), así como poder emplear los formantes como modelo del habla, más cercano a nosotros que los parámetros matemáticos de bajo nivel.

2. ANÁLISIS LPC

Entre las ventajas que presenta el método de predicción lineal se encuentran [RAB93]:

- LPC proporciona un modelo adecuado de la señal de voz y sus parámetros se ajustan a las características del tracto vocal, especialmente en los sonidos sonoros del habla cuyas propiedades se aproximan más a la señal estacionaria que en los sonidos sordos [RAN95].

- Los parámetros obtenidos mediante predicción lineal muestran un espectro suavizado que proporciona la información más representativa de la voz. Esto evita perderse en los detalles ofrecidos por la transformación de Fourier.

- LPC es un método preciso, muy adecuado para computación, tanto por su sencillez como por la rapidez de ejecución que presentan algunos de los algoritmos hallados.

- Las pruebas realizadas con este método muestran muy buenos resultados en diversos campos del tratamiento automático de la voz.

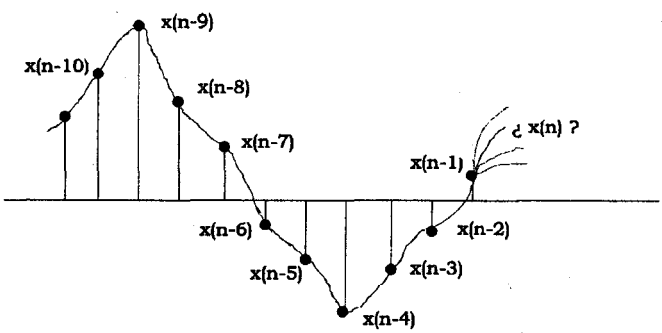


Figura 2
Predicción de los datos adicionales de la señal aplicando LPC

El problema que plantea LPC radica en obtener un cierto número de coeficientes (a_k), los cuales serán usados para conseguir alguna característica de la señal mediante una función de transferencia. Algunos ejemplos de características que se pueden obtener son el espectro o la predicción de datos en una serie numérica(Figura 2).

$$e(n) = x(n) - s(n) = x(n) - \sum_{k=1}^{p} a_k(n)\, x(n-k) \qquad (1)$$

El objetivo es dar un valor, *s(n)*, lo más aproximado al valor que se desconoce, *x(n)*, es decir, una predicción del valor *x(n)* cuyo error sea el menor posible. Esto implica que LPC no va a ser un método exacto, sino que tan sólo dará valores aproximados, aunque esto hará que se decremente el tiempo de cómputo frente a la transformada rápida de Fourier.

Los valores a_k desconocidos, se calculan minimizando el error *e(n)*, para lo cual se aplican mínimos cuadrados. Para ello se forma el error cuadrático medio en el intervalo *'n'* que se desea considerar.

$$L = \sum_{n} e^2(n) = x(n) - \sum_{i=1}^{k} a_i * x(n-i) \quad 0 \le n \le N\text{-}1 \qquad (2)$$

Para obtener el valor mínimo de *L*, se deriva respecto a cada una de las variables $a_j \mid 1 \le j \le k$

$$\frac{\partial L}{\partial a_j} = 0 \qquad 1 \le j \le k \qquad (3)$$

$$\frac{\partial L}{\partial a_j} = \frac{\partial}{\partial a_j} \sum_{n} \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right]^2 = 0 \qquad (4)$$

$$\frac{\partial L}{\partial a_j} = \sum_n \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right]^2 = 0 \qquad (5)$$

$$\frac{\partial L}{\partial a_j} = \sum_n 2 * \frac{\partial}{\partial a_j} \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] \qquad (6)$$

$$\frac{\partial L}{\partial a_j} = -2 * \sum_n x(n-j) * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] = -2 * \sum_n x(n-j) * e(n)$$

$$\frac{\partial L}{\partial a_j} = \sum_n x(n-j) * e(n) = 0 \qquad 1 \le j \le k \qquad (8)$$

La expresión anterior se puede desarrollar como:

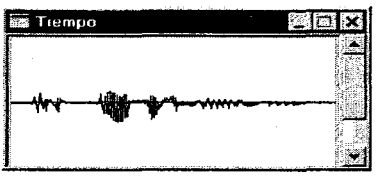
$$\sum_n x(n-j) * \left[x(n) - \sum_{i=1}^{k} a_i * x(n-i) \right] = \qquad (9)$$

$$\sum_n x(n-j) * x(n) - \sum_{i=1}^{k} a_i * \sum_n x(n-j) * x(n-i) = \qquad (10)$$

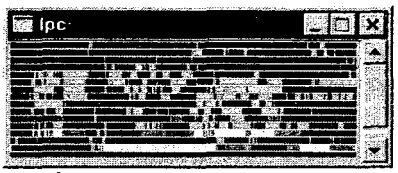
$\boxed{C_{j0} - \sum_{i=1}^{k} a_i * C_{ji}}$ donde se ha definido:

$$C_{ji} = \sum_n x(n-j) * x(n-i) \qquad (11)$$

Despejando las incógnitas 'a_i' obtenemos los parámetros de Predicción Lineal. La siguiente figura muestra la señal de voz correspondiente a la palabra 'economía', los parámetros LPC obtenidos y el espectro de voz asociado.



SEÑAL EN EL TIEMPO



PARÁMETROS LPC

ESPECTRO DE VOZ OBTENIDO

3. ESTRATEGIAS PARA LA OBTENCIÓN DE FORMANTES

Obtener formantes a partir de los picos proporcionados por una FFT, tiene el inconveniente de que la función espectral presenta demasiada información, por lo cual resulta difícil elegir los picos que representan los formantes principales, descartando el resto de los máximos. Al usarse LPC, trabajamos con una función simplificada,

que a modo de envolvente de la FFT conserva la información fundamental, disminuyéndose aquella que no es necesaria para nuestro propósito [RAB93].

Una primera aproximación en la obtención de formantes consiste en calcular las raíces de las ecuaciones planteadas y seleccionar aquellas soluciones que se adapten a los formantes esperados utilizando criterios algorítmicos. Este método presenta la ventaja de ser conceptualmente muy simple, pero existe el grave inconveniente de que requiere una gran carga de computación para ser llevado a cabo.

Otra posibilidad con tiempos de computación mucho más bajos se basa en la utilización de los algoritmos de Levinson-Durbin y Celosía Adaptativa [RAB93]. Después, seleccionar los máximos de las funciones resultantes y escoger (con criterios algorítmicos) los 3 picos que se consideren más adecuados para representar a F1, F2 y F3 [SCH70].

La determinación de picos sobre una curva espectral presenta varios problemas de exactitud:

1. No siempre se encuentran todas las soluciones en el radio evaluado sobre el círculo complejo.
2. A veces aparecen de forma temporal pequeños picos que no deben ser confundidos con formantes.
3. Cuando dos polos están muy cercanos, tienden a unirse en un solo pico, pudiéndose perder de esta manera un posible formante.

La figura 3 ilustra el caso de un pico que representa un formante, pudiéndose apreciar su evolución a lo largo de 6 instantes de tiempo en mitad de una grabación de la vocal 'o'. En $t=1$ la pendiente de subida al pico no es lo suficientemente grande como para poder etiquetarlo como candidato a formante, a medida que pasa el tiempo, el máximo se va consolidando, hasta que se convierte en el

candidato a F2 de la vocal. Esta es la situación opuesta al caso 2 descrito anteriormente.

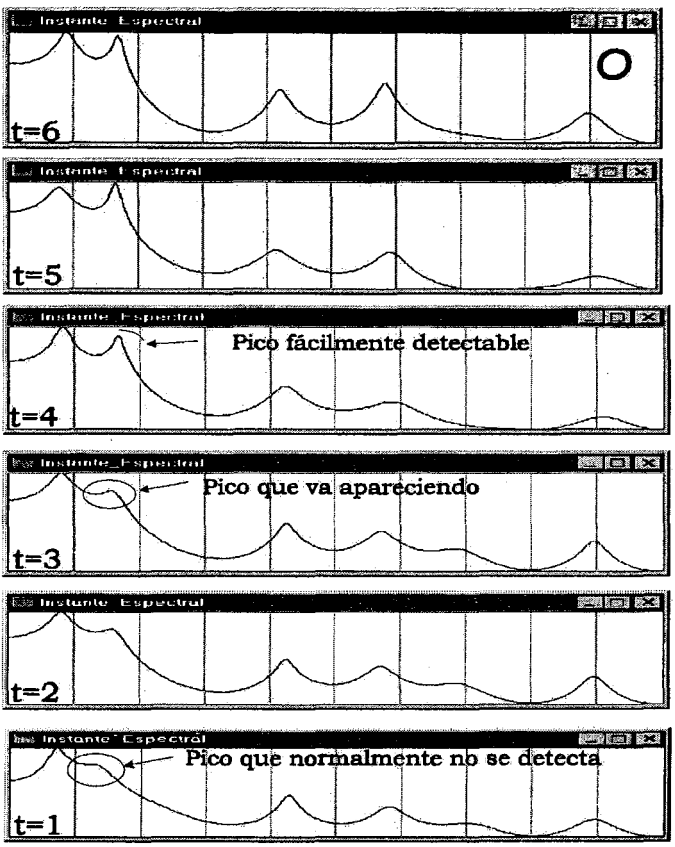


Figura 3

Ejemplo de evolución a lo largo del tiempo de los picos en las funciones espectrales

En la determinación de formantes se debe asumir un grado de inexactitud en los resultados que, tal y como ocurre en la naturaleza, se puede intentar corregir en niveles superiores del proceso de

reconocimiento, por ejemplo realizando comprobaciones de consistencia a nivel de palabra, sintáctico, etc. [CAS87]

4. DESARROLLO

Los gráficos presentados en el artículo como ventanas Windows, pertenecen a una aplicación informática desarrollada para probar las distintas teorías que se van creando con el fin de obtener resultados satisfactorios en diversos campos del procesamiento de la voz. De esta manera se va trabajando en un ciclo de teoría $\Rightarrow$ programación $\Rightarrow$ prueba $\Rightarrow$ error , que va generando la suficiente documentación y unidades orientadas a objetos como para poder imaginar las siguientes teorías y probarlas hasta obtener los aciertos esperados.

Como resultado de los métodos y algoritmos diseñados, se han obtenido diversas funciones espectrales con las que se consiguen sonogramas del habla que presentan los formantes realzados. A modo de ejemplo de las fases matemáticas y algorítmicas empleadas, se presentan los siguientes espectrogramas en los que se puede apreciar gráficamente los resultados conseguidos.

En el último caso, la señal se eleva en las zonas de formantes, y se rebaja en el fondo de los valles situados entre máximos relativos. Este resultado asegura una buena determinación espectral de los formantes del habla.

En el caso de espectros tridimensionales estas operaciones son menos importantes, puesto que la separación entre formantes se visualiza con claridad, sin embargo, cuando se hace uso de espectros en los que la tercera dimensión se codifica con colores (o escala de grises), las transformaciones de 'realzado' son muy útiles para diferenciar e identificar formantes.

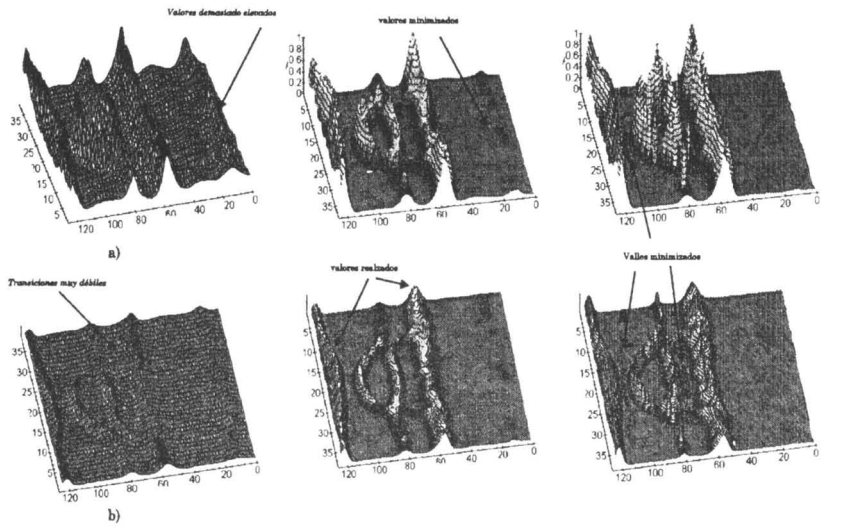


Figura 4
Espectro del triptongo 'ioi' obtenido usando diferentes métodos y algoritmos propios basados en LPC. Los espectrogramas superiores e inferiores representan distintas vistas de las mismas figuras

Estos espectros se presentan para probar las funciones empleadas son bidimensionales, codificándose la tercera dimensión mediante una escala de colores. En este artículo se imprime como escala de grises, perdiéndose buena parte de su calidad, sin embargo, el lector podrá comprobar en líneas generales la correcta determinación de los formantes.

En el gráfico se visualiza la secuencia 'ieaou'. En el caso a), el espectro ha sido obtenido mediante una aplicación desarrollada previamente a este trabajo. El método para pasar a frecuencias ha sido el de predicción lineal. En este primer caso se puede apreciar que existe saturación y los formantes no se determinan visualmente con

facilidad. El resto de los espectrogramas se obtienen aplicando diversos algoritmos y filtros digitales de señal.

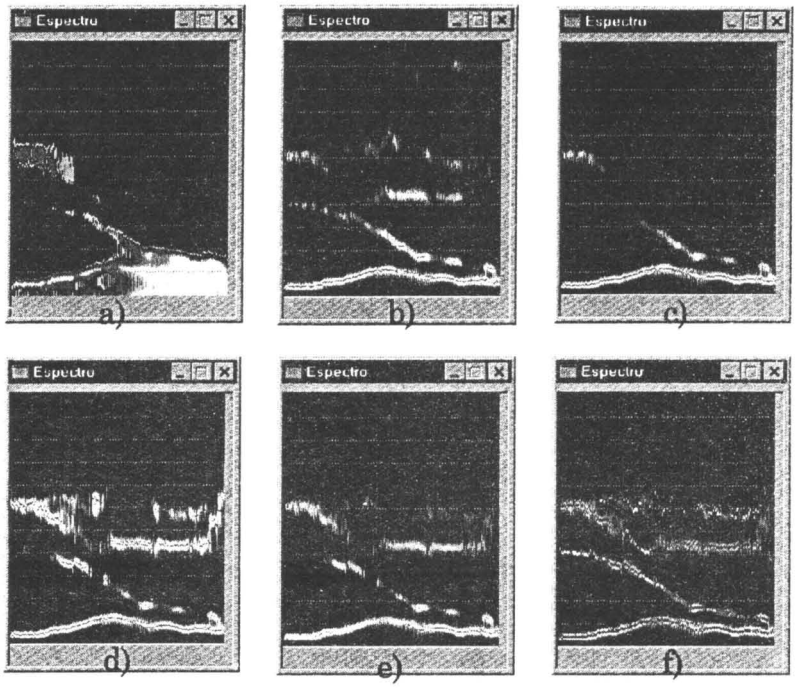


Figura 5
Espectros del sonido 'ieaou' usando diferentes funciones y filtros de señal para su obtención.

Los siguientes espectros han sido obtenidos utilizando la grabación de los sonidos 'eme ene eɲe'.

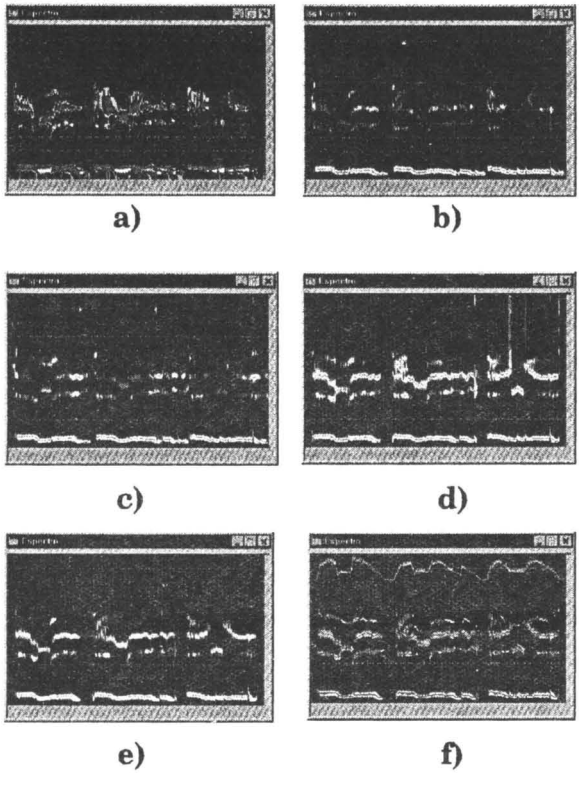


Figura 6
Espectros de los sonidos 'eme ene eɲe' usando diferentes funciones y filtros digitales de señal para su obtención.

5. RESULTADOS

Una vez comentada la existencia de distintos algoritmos y funciones matemáticas ideados para la obtención de los espectros, en este apartado se presentarán casos de ejemplo de espectros finales de sonidos sonoros.

La claridad de un espectro no depende únicamente de la idoneidad de los procesos matemáticos e informáticos con los que se obtiene, sino que además influyen significativamente factores tales como la calidad de la grabación de partida, la precisión del soporte físico de visualización/impresión, etc. Por ello se ha decidido establecer 'pares de prueba', formados por espectros obtenidos con el método propuesto y sus correspondientes réplicas mediante espectros generados a partir de predicción lineal básica.

La figura 7 presenta en primer lugar un espectro tradicional LPC en a) y el obtenido con nuestros algoritmos en b), pertenecientes a la secuencia de consonantes sonoras vibrantes y laterales: 'ere ɛ̄re ele eʎe'. Aunque en el espectro LPC se aprecian las características básicas de estos sonidos, en el caso b) la claridad es bastante mayor, y los formantes aparecen más nítidos y delimitados.

Los casos c) y d) se corresponden con la secuencia de sonidos 'imi eme ama omo umu'. En el caso d) se evita la saturación de la escala y se delimitan más claramente los formantes de las vocales. Las posiciones frecuenciales de la nasal quedan bien definidos en ambos casos.

La figura 8 se centra en el grupo oclusivo. Las oclusivas sordas se muestran en los casos a) y b) mediante los sonidos 'epe ete eke'. En b), sólo se representan las secciones sonoras. Como se puede apreciar, los formantes están bien delimitados, y su evolución corresponde con la esperada.

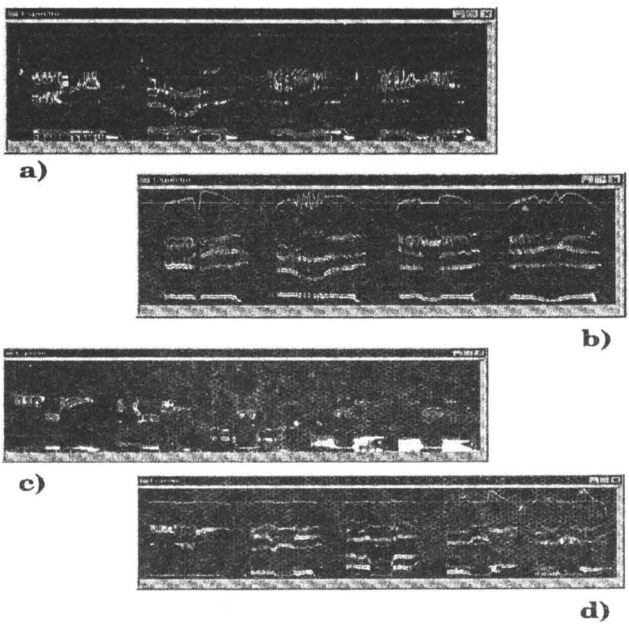


Figura 7
Sonidos 'ere ɛ̄re ele eʎe' y 'imi eme ama omo umu' empleando los espectros propuestos y sus correspondientes LPC

Los casos c) y d) contienen consonantes fricativas sonoras. La grabación corresponde a los sonidos 'eβe eðe eγe'. Aunque en el caso c) la trayectoria de los formantes es clara, su intensidad es pequeña. En d) los formantes aparecen suficientemente realzados.

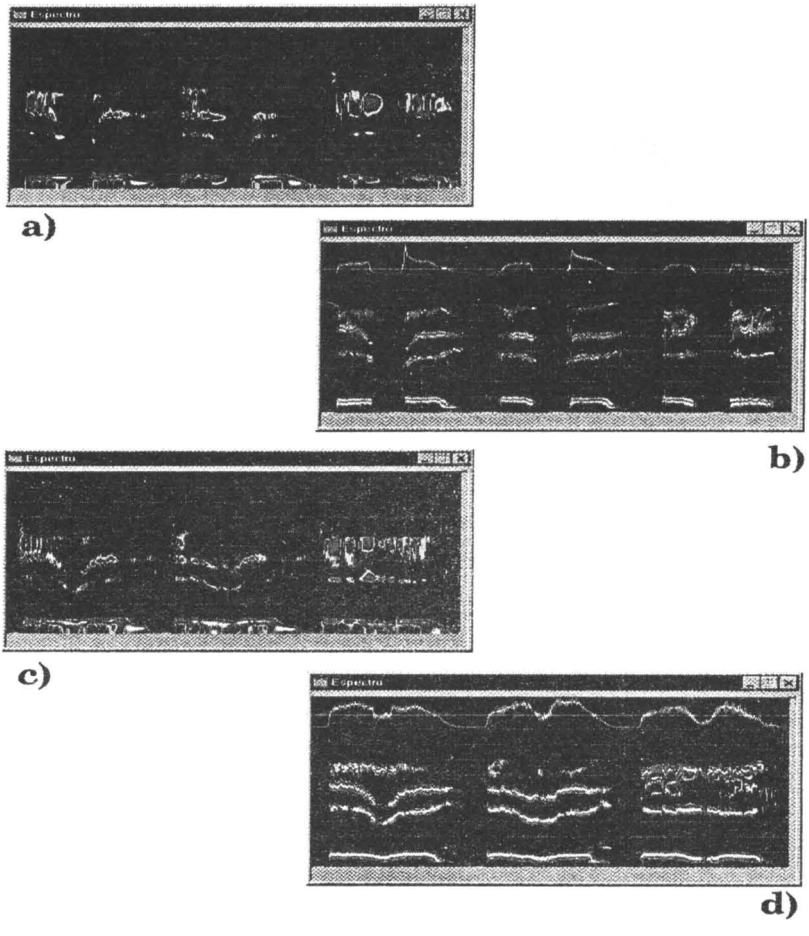


Figura 8

Sonidos 'epe ete eke' y 'eβe eðe eγe' empleando los espectros propuestos y sus correspondientes LPC

Por último, en la figura 9 se presentan tres espectros, el primero de ellos contiene las palabras ayer, hoy, mañana. Podemos apreciar una buena precisión en las transiciones de los formantes al evolucionar de una vocal a otra ('hoy') y las características de los sonidos nasales en 'mañana'.

El segundo espectro contiene los triptongos que empiezan y acaban en la vocal 'i'. Aparecen bien definidas las transiciones desde y hasta las 'ies'. Se puede observar como los formantes segundo y tercero bajan hasta cada una de las vocales centrales de los triptongos.

Los espectros tridimensionales del apartado anterior se corresponden con el sonido 'ioi' de esta figura.

El último espectro de la figura 9 muestra como se visualiza una frase completa. En este caso, podemos observar conjuntamente sonidos vocálicos, nasales, fricativo sonoro y vibrantes. La evolución de los formantes aparece muy nítida en todo el intervalo mostrado. Las nasales se distinguen bien y la vibrante múltiple (forzada para realzar sus características) se aprecia con claridad.

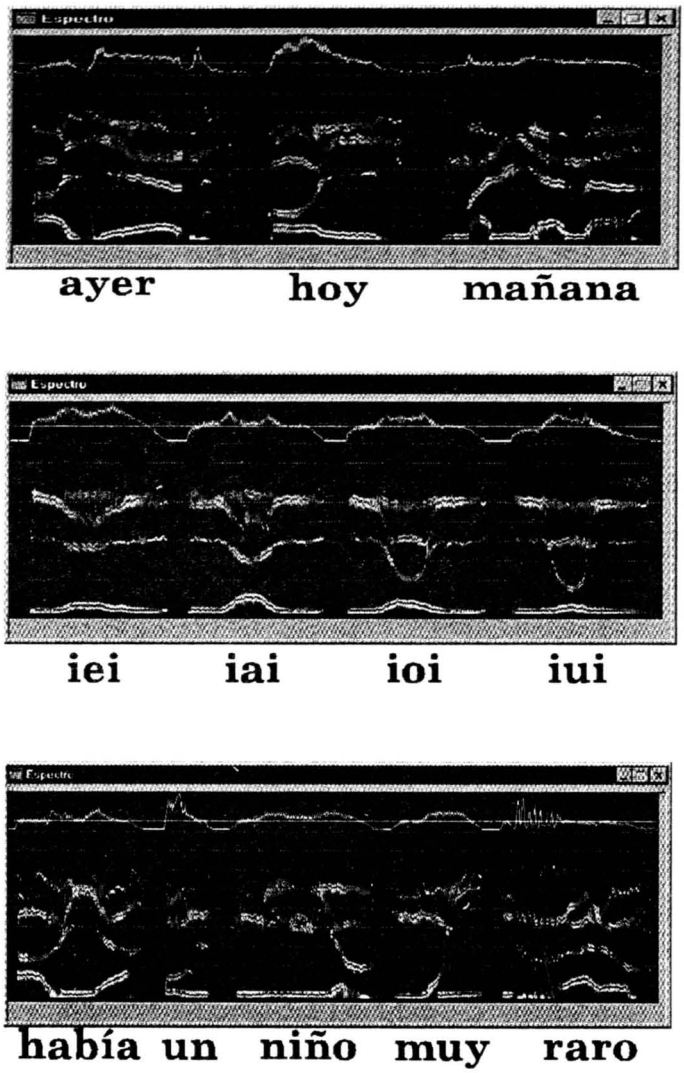


Figura 9

Espectros finales utilizando el método propuesto

6. CONCLUSIONES

Los algoritmos basados en el método LPC presentan una adecuada capacidad para seleccionar de forma automática la información que debe ser tomada en cuenta para la obtención de formantes, pudiéndose desechar aquella que es menos representativa para la consecución de este fin.

La obtención de espectros claros, requiere la utilización de varios filtros que realicen las siguientes funciones:

1. Eliminar formantes erróneos.
2. Disminuir la importancia de las zonas que no proporcionan información espectral básica.
3. Realzar los formantes.
4. Minimizar los valores de las zonas situadas entre formantes.

Los resultados obtenidos se pueden considerar satisfactorios, presentándose espectros bastante más claros que sus correspondientes calculados con otros métodos y herramientas.

Los algoritmos desarrollados proporcionan en su conjunto una buena calidad en la visualización de los espectros de voz, calidad basada en la correcta determinación de las características espectrales buscadas.

7. REFERENCIAS BIBLIOGRÁFICAS

[CAS87] F. Casacubierta, E. Vidal, *Reconocimiento automático del habla*, Marcombo, 1987

[CAS90] F. Casacubierta, E. Vidal, *Reconocimiento automático del habla*, Estudios de Fonética Experimental, 1990, Vol. 4, pp. 167-178

[FRE93] J. Freeman, D. Skapura, *Redes neuronales, algoritmos, aplicaciones y técnicas de programación*, Addison-Wesley/Díaz de Santos, 1993

[MAR90b] R. Marti, J., "Situación actual de la síntesis de voz", *Estudios de Fonética Experimental*, 1990, Vol. 4, pp. 167-178

[MAR94] E. Martínez Celdrán, *Fonética*, Martínez Celdrán E., Teide, 1994

[PIC95] J.M. Pickett, H. Bunell, S. Revoile, "Phonetics of intervocalic consonant perception: retrospect and prospect", *Phonetica*, Vol. 52, 1995, pp. 1-40

[QUI93] A. Quilis, *Tratado de fonología y fonética españolas*, Gredos, 1993

[RAB89] L.R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77, Nº 2, 1989, pp. 257-286

[RAB93] L.R. Rabiner, *Fundamentals of speech recognition*, Biing-Hwang Juang, Prentice Hall, 1993

[RAN95] M. Rangoussi, A. Delopoulos, "Recognition of unvoiced stops from their time-frequency representation", *International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, 1995, pp. 792-795

[ROW92] C. Rowden, *Speech processing*, Mc Graw Hill, 1992

[SCH70] R.W. Schafer, L.R. Rabiner, "System for automatic formant analysis of voiced speech", *The Journal of the*

Acoustic Society of America, Vol. 47, Nº 2, 1970, pp. 634-648

[TOH92] Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka, *Speech perception, production and linguistic structure*, IOS Press, 1992