TECNICAS ROBUSTAS DE RECONOCIMIENTO DEL HABLA EN AMBIENTES ADVERSOS

Javier Hernando, Climent Nadeu y José B. Mariño

Departamento de Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña, 08034 Barcelona
E-mail: javier@gps.tsc.upc.es

Abstract

The performance of the existing speech recognition systems degrades rapidly in the presence of background noise. A novel representation of the speech signal, which is based on Linear Prediction of the One-Sided Autocorrelation sequence (OSALPC), has shown to be attractive for speech recognition because of both its high recognition performance with respect to the standard LPC in severe conditions of additive white noise and its computational simplicity. The aim of this work is twofold: 1) to show that OSALPC also achieves good performance in a case of real noisy speech (in a car environment), and 2) to explore its combination with several robust similarity measuring techniques, showing that its performance even improves by filtering and multilabeling conveniently the spectral parameters.

Resumen

El comportamiento de los sistemas actuales de reconocimiento del habla se degrada rápidamente en presencia de ruido de fondo. Recientemente, se ha propuesto una técnica de representación de la señal de voz basada en la predicción lineal de la parte causal de la autocorrelación (OSALPC) que ha mostrado ser atractiva para el reconocimiento de habla ruidosa debido a sus altas prestaciones con respecto a la predicción lineal (LPC) convencional en

condiciones severas de ruido blanco aditivo y a su simplicidad computacional
El propósito de este artículo es doble: 1) mostrar que la técnica OSALPC
obtiene también buenas prestaciones en un un entorno ruidoso real (ruido de
coche), y 2) explorar su combinación con varias técnicas robustas de medida
de similitud, mostrando que sus prestaciones mejoran aún más filtrando
convenientemente los parámetros espectrales y realizando un etiquetado
múltiple de los mismos.

1. Introducción

El problema del reconocimiento del habla en entornos ruidosos permanece sin
resolver incluso en el caso de palabras aisladas y vocabularios pequeños. Por
esta razón, se han propuesto en la literatura [1] diversas técnicas para reducir
ruido en cada una de las etapas del proceso de reconocimiento, particularmente
en extracción de parámetros de la voz y medidas de similitud.
Respecto a la etapa de parametrización, la técnica de predicción lineal (LPC
Linear Predictive Coding) [2] es usada ampliamente en reconocimiento del
habla. Sin embargo, es bien conocido que la técnica convencional LPC es muy
sensible a la presencia de ruido blanco, lo cual provoca tasas de
reconocimiento pobres en condiciones ruidosas.
Recientemente, los autores han propuesto la técnica OSALPC (One-Sided
Autocorrelation Linear Predictive Coding) [3] como una representación
alternativa de la señal de voz en presencia de ruido. Como su nombre indica
en esta técnica se realiza predicción lineal en el dominio de la autocorrelación
en lugar de sobre la señal misma, y su uso en reconocimiento de habla ruidosa
es atractivo debido a sus altas prestaciones con respecto a la técnica LPC
convencional en condiciones severas de ruido blanco aditivo y a su simplicidad
computacional.
El propósito de este trabajo es doble: 1) mostrar que la técnica OSALPC
obtiene también buenas prestaciones en un un entorno ruidoso real (ruido de

coche), y 2) explorar su combinación con varias técnicas robustas de medida de similitud, mostrando que sus prestaciones mejoran aún más filtrando convenientemente los parámetros espectrales [4] [5] [6] [7] y realizando un etiquetado múltiple de los mismos [8].

El artículo está organizado del siguiente modo. En las secciones 2, 3 y 4 se revisarán brevemente las tres técnicas consideradas en este trabajo: parametrización OSALPC, filtrado de parámetros y etiquetado múltiple, respectivamente (para más información ver [9]). La sección 5 está dedicada a mostrar los resultados experimentales obtenidos aplicando estas técnicas, separadamente y en combinación, al reconocimiento multilocutor de palabras aisladas en entorno ruidoso de coche utilizando un sistema de reconocimiento basado en los modelos ocultos de Markov (MOM) y la cuantificación vectorial (CV). Finalmente, en la sección 6 se resumen algunas conclusiones obtenidas a partir de estos resultados

2. Representación OSALPC

A partir de la secuencia de autocorrelación de la señal de voz x(n),

$$R(m) = E\{x(n+m)\, x^*(n)\} \qquad (1)$$

donde E{.} es el operador esperanza, definimos la parte causal de la autocorrelación (PCA) como

$$R^+(m) = \begin{cases} R(m) & m > 0 \\ R(0)/2 & m = 0 \\ 0 & m < 0 \end{cases} \qquad (2)$$

Su transformada de Fourier es el "espectro" complejo

$$S^+(\omega) = \frac{1}{2}\ [\, S(\omega) + jS_H(\omega)\,] \qquad (3)$$

donde $S(\omega)$ es el espectro de la señal, es decir, la transformada de Fourier de $R(m)$, y $S_H(\omega)$ es la transformada de Hilbert de $S(\omega)$.

Debido a la analogía entre $S^+(\omega)$ en (3) y la señal analítica usada en modulación de amplitud, puede definirse una "envolvente" espectral [10] como

$$E(\omega) = |S^+(\omega)| \qquad (4)$$

cuyo cuadrado es el espectro de la secuencia PCA.

Esta característica de envolvente, junto con el alto rango dinámico de los espectros de voz, origina que $E(\omega)$ realce de forma acusada las bandas frecuenciales de mayor potencia. Por tanto, los componentes de ruido correspondientes a otras bandas quedan atenuados fuertemente en $E(\omega)$ con respecto a $S(\omega)$, y $E(\omega)$ resulta ser más robusta al ruido de banda ancha que $S(\omega)$. Por otro lado, es conocido que la secuencia PCA tiene los mismos polos que la propia señal [11].

Ambas propiedades, robustez al ruido de banda ancha y conservación de los polos, sugieren que los polos de la señal de voz pueden estimarse de forma más fiable a partir de la secuencia PCA que directamente a partir de la propia señal cuando ésta está contaminada por ruido de banda ancha. Para ello, del mismo modo que la técnica LPC convencional estima dichos polos realizando predicción lineal de la señal de voz, que es equivalente a suponer un modelo todo-polos para el espectro de la señal $S(\omega)$, podemos considerar una predicción lineal de la secuencia PCA, equivalente a suponer un modelo todo-polos para su espectro $E^2(\omega)$. Esta es la base de la técnica de parametrización OSALPC, propuesta en [3] como una representación robusta de la señal de voz en presencia de ruido.

Se ha propuesto un algoritmo simple para calcular los coeficientes cepstrales correspondientes a la técnica OSALPC, consistente en la aplicación de la técnica LPC convencional sobre la secuencia PCA (ver figura 1). El valor del

máximo índice utilizado de la secuencia PCA se ha optimizado empíricamente para tener en cuenta el compromiso existente entre varianza y resolución de la estimación espectral y se ha tomado igual a la mitad de la duración N de la trama de señal.

En la figura 2 se ilustra la robustez de la técnica OSALPC al ruido blanco aditivo. Como puede verse en la figura, el cuadrado de la envolvente espectral OSALPC realza de forma acusada la banda frecuencial de mayor potencia y es más robusto al ruido que el espectro LPC. La figura muestra también que pueden aparecer picos espurios en la envolvente espectral OSALPC. Probablemente, son debidos a que la técnica OSALPC sólo realiza una deconvolución parcial entre el filtro y la excitación del modelo de producción de voz [3].

La técnica SMC, propuesta por D. Mansour y B.H. Juang en [12], también está basada en la predicción lineal en el dominio de la autocorrelación. En términos de la formulación anterior, la diferencia fundamental entre ambas técnicas consiste de hecho en que la técnica SMC supone un modelo espectral todo-polos para la envolvente $E(\omega)$ en lugar de $E^2(\omega)$.

La técnica OSALPC fue comparada en [3] con las técnicas LPC convencional y SMC usando señales de voz que incluían ruido blanco aditivo. En aquellas pruebas, la técnica OSALPC superó en prestaciones a las otras dos para relaciones señal a ruido bajas.

3. Filtrado de parámetros espectrales

La representación paramétrica de la señal de voz consiste en una secuencia de vectores, uno por trama de señal, cuyos componentes suelen ser algún tipo de coeficientes cepstrales. Esta secuencia bidimensional entra a la etapa de comparación del sistema de reconocimiento donde es clasificada en base a unos modelos y unas medidas de similitud.

Recientemente, se ha descubierto que puede mejorarse en gran medida la capacidad discriminante del clasificador procesando de forma apropiada esta representación espectral bidimensional. Esta idea se aplica en el dominio cepstral (o cuefrencial) mediante el enventanado cepstral (o liftado) -producto de la secuencia cepstral por una secuencia de ponderación o ventana- [4]; y en el dominio temporal mediante los llamados rasgos dinámicos o parámetros diferenciales [7].

Por un lado, el enventanado cepstral implica de hecho un filtrado del logaritmo del espectro ya que realiza una convolución periódica del mismo con la transformada de Fourier de la ventana cepstral. Por otro lado, cada parámetro diferencial puede ser visto como la salida de un filtro lineal excitado por la secuencia temporal de coeficientes cepstrales, o cualquier otro parámetro espectral, donde cada muestra corresponde a una trama. Por tanto, pueden interpretarse ambos tipos de procesado desde un punto de vista de filtrado en frecuencia (enventanado cepstral) o en tiempo (parámetros diferenciales), y el análisis frecuencial del proceso de filtrado permite profundizar en su comportamiento [13]. En este sentido, puede hablarse de parámetros filtrados.

Todos los filtros usados hasta ahora en la dimensión frecuencial (análogamente en la dimensión temporal) muestran características paso-banda. De hecho, presentan dos componentes básicos [13]: un componente de diferenciación, que corresponde a un liftado paso-alto (filtrado), y 2) un componente de suavizado que realiza un liftado paso-bajo (filtrado). El componente de diferenciación produce un aumento de resolución en frecuencia (tiempo) del espectro, en el sentido de amplificar su dinámica. El componente de suavizado del filtro atenúa los componentes de alta cuefrencia (frecuencia) poco fiables. Por tanto, podemos interpretar este filtrado como un cambio del compromiso entre resolución en frecuencia (tiempo) y potencia de error del proceso de

estimación espectral que comporta la parametrización para mejorar la capacidad de discriminación del reconocedor.

El orden del modelo de predicción lineal es otro parámetro que permite controlar el mismo tipo de compromiso, ya que un orden alto puede redundar en una mayor resolución frecuencial de la representación de la voz pero puede también producir un aumento del error de la estimación espectral. Cuando la señal de voz es ruidosa, el compromiso puede depender de la relación señal a ruido y de las características del ruido.

En nuestros experimentos se han considerado las tres ventanas cepstrales más usadas:

$$\begin{aligned} &\text{Seno realzado:} && w(n) = 1 + \frac{L}{2} \sin\left(\frac{\pi n}{L}\right) \\ &\text{Rampa:} && w(n) = n \\ &\text{Inversa de la desviación típica:} && w(n) = \frac{1}{\sigma_{c(n)}} \end{aligned} \qquad (5)$$

donde $n = 1,...,L$, y $\sigma_{c(n)}$ es la desviación típica del n-ésimo coeficiente cepstral c(n). Si p denota el orden del filtro de predicción, el valor de L suele ser 3p/2 para la ventana seno realzado [4], y p para la ventana rampa [5] y la inversa de la desviación típica [6].

La versión más común de parámetros filtrados en el tiempo son los llamados coeficientes de regresión o delta-cepstrum [7]. La respuesta impulsional asociada a estos es el polinomio discreto de Legendre de primer grado:

$$h(n) = \begin{cases} -n, & -N \le n \le N \\ 0, & \text{en otro caso} \end{cases} \qquad (6)$$

En nuestro trabajo aplicamos este filtro a las secuencias temporales de coeficientes cepstrales para obtener para cada trama un vector de parámetros

filtrados que suplemente el vector cepstral. La longitud 2N+1 de su respuesta impulsional se optimizó empíricamente.

4. Etiquetado múltiple

Los modelos ocultos de Markov (MOM) [14] se han convertido en los últimos años en la aproximación predominante en decodificación acústico-fonética, debido a la simplicidad de su estructura algorítmica y a sus buenas prestaciones. Por ello, el sistema de reconocimiento utilizado en las pruebas experimentales de este trabajo está basado en MOM.

Básicamente, un MOM es la representación de un proceso estocástico que consta de dos mecanismos interrelacionados: una cadena de Markov subyacente, con un número finito de estados, caracterizada por las probabilidades iniciales y de transición entre estados; y un conjunto de funciones aleatorias observables, cada una de ellas asociadas a un estado. Los parámetros de los modelos de cada unidad de reconocimiento son aprendidos de forma automática a partir de una base de datos de entrenamiento. En el reconocimiento, la decisión se toma en base al modelo o secuencia de modelos más probable.

En los modelos ocultos de Markov discretos (MOMD) las probabilidades de observación están representadas mediante distribuciones de probabilidad discretas de un conjunto finito de símbolos. Por ello, los vectores de parámetros espectrales de la voz son cuantificados previamente utilizando un cuantificador vectorial a cuyas palabras-código se asignan los símbolos de los modelos.

En la cuantificación vectorial (CV) convencional, utilizada en los MOMD, para cada vector de parámetros de entrada el cuantificador realiza una decisión drástica acerca de cuál de sus palabras-código es la adecuada y, por tanto, se descarta la información asociada al grado de proximidad a las otras palabras-

código. Esta información puede ser especialmente importante en el caso de habla ruidosa, ya que esta decisión puede ser afectada fácilmente por el ruido.

En contraste con la CV convencional, el etiquetado múltiple realiza una decisión flexible acerca de qué palabra-código es la más cercana al vector de entrada, generando un vector de salida cuyos componentes indican la proximidad relativa del vector de entrada a cada palabra-código.

Sea $\{\mathbf{v}_k\}_{k=1...C}$ el diccionario del cuantificador de etiquetado múltiple, donde C es el tamaño del diccionario, y sea $\mathbf{x}_t$ el vector cepstral enventanado en el instante t. El cuantificador de etiquetado múltiple utilizado en este trabajo [8] asigna al vector de entrada $\mathbf{x}_t$ un vector de salida $\mathbf{O}_t=\{w(\mathbf{x}_t,\mathbf{v}_k)\}_{k=1...C}$ estimado como

$$w(\mathbf{x}_t,\mathbf{v}_k) = \frac{1/d(\mathbf{x}_t,\mathbf{v}_k)}{\sum_{m=1}^{C} 1/d(\mathbf{x}_t,\mathbf{v}_m)} \qquad (7)$$

donde $d(\mathbf{x}_t,\mathbf{v}_k)$ es la distancia entre $\mathbf{v}_k$ y $\mathbf{x}_t$.

Estos componentes son positivos, suman 1 y son decrecientes con $d(\mathbf{x}_t,\mathbf{v}_k)$. Por tanto, puede interpretarse $\mathbf{O}_t$ como un vector que describe la probabilidad de que un vector de entrada $\mathbf{x}_t$ corresponda a la clase representada por cada palabra-código $\mathbf{v}_k$.

Los algoritmos de los MOMD deben ser generalizados para incorporar la salida del etiquetado múltiple. Para un estado j, la probabilidad de que el vector $\mathbf{x}_t$ sea observado se escribe como

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{C} w(\mathbf{x}_t,\mathbf{v}_k)\, b_j(k) \quad , \qquad (8)$$

donde $b_j(k)$ denota la probabilidad de observación discreta asociada a la palabra-código $\mathbf{v}_k$ y el estado j.

El algoritmo clásico de Viterbi para la evaluación de los modelos se generaliza simplemente utilizando (8) en lugar de $b_j(k)$. Con respecto al entrenamiento, las fórmulas de reestimación de Baum-Welch para las probabilidades de transición y las probabilidades iniciales se generalizan del mismo modo. En cuanto a la reestimación de $b_j(k)$, la estimación de máxima verosimilitud conduce a la siguiente fórmula para una secuencia de entrenamiento de longitud T

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)\,\frac{w(\mathbf{x}_t,\mathbf{v}_k)\,b_j(k)}{b_j(\mathbf{x}_t)}}{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)} \qquad (9)$$

donde $\alpha_t(j)$ y $\beta_t(j)$ son las probabilidades hacia adelante y atrás [14], respectivamente.

Sin embargo, en las pruebas realizadas en este trabajo se empleó una fórmula de reestimación alternativa. De hecho, se obtuvieron mejores tasas de reconocimiento utilizando la siguiente expresión [9]

$$\bar{b}'_j(k) = \frac{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)\,w(\mathbf{x}_t,\mathbf{v}_k)}{\sum_{t=1}^{T} \alpha_t(j)\,\beta_t(j)} \qquad (10)$$

La expresión (10) beneficia la probabilidad de las palabras-código más cercanas al vector de entrada, ya que se han eliminado los términos $b_j(k)$ y

$b_j(\mathbf{x}_t)$, que ponderan los componentes del vector de salida del cuantificador $w(\mathbf{x}_t,\mathbf{v}_k)$ en la fórmula (9). Debido a la eliminación de estos términos correspondientes a valores previos de las probabilidades de observación, el uso de esta fórmula de reestimación requiere menor número de iteraciones y, por tanto, reduce la carga computacional del entrenamiento. Sin embargo, la fórmula (10) no garantiza la convergencia del proceso de entrenamiento. Por otro lado, (7) y (8) pueden simplificarse en la práctica usando sólo los K valores más significativos de $w(\mathbf{x}_t,\mathbf{v}_k)$ para cada $\mathbf{x}_t$, donde K es menor que el tamaño C del diccionario. Todo ello reduce extraordinariamente la carga computacional de los modelos ocultos de Markov con etiquetado múltiple (MOMEM).

El método de etiquetado múltiple presentado en este trabajo es similar a los descritos en [15] y [16]. Las principales discrepancias con respecto a ellos son la posibilidad de usar cualquier medida de distancia en (7), la generalización alternativa de los algoritmos de los MOM (10), y la simplificación de (7) y (8) usando sólo los K palabras-código más cercanas.

Los modelos semicontinuos (MOMSC) [17] están también estrechamente relacionados con el etiquetado múltiple. Sin embargo, los componentes del vector de salida $w(\mathbf{x}_t,\mathbf{v}_k)$, que se estiman desde un punto de vista determinista en el etiquetado múltiple, se estiman desde un punto de vista estocástico en los MOMSC. Concretamente, mientras que en el etiquetado múltiple las palabras-código se identifican con el centroide -media- de cada agrupación, en los MOMSC el diccionario se modela como una familia paramétrica de densidades gaussianas, caracterizadas por la media y la varianza de cada agrupación.

Las tasas de reconocimiento obtenidas con los MOMEM y los MOMSC son similares y superan considerablemente las obtenidas con los MOMD convencionales, en condiciones libres de ruido y en presencia de ruido blanco aditivo [8]. Sin embargo, los algoritmos correspondientes a los MOMEM son más eficientes computacionalmente que los correspondientes a los MOMSC.

5. Resultados experimentales

La base de datos usada en nuestros experimentos procede del proyect ESPRIT-ARS y consiste en 25 repeticiones de los dígitos pronunciados por 4 locutores, 2 hombres y 2 mujeres, grabadas en diferentes entornos ruidosos: 5 repeticiones con el motor y el ventilador funcionando y 20 más con el motor encendido y diferentes velocidades del ventilador -10 con el coche parado, 5 con el coche circulando a 70 km/h y 5 con el coche circulando a 130 km/h-. El sistema de reconocimiento fue entrenado con las señales pronunciadas con el motor y el ventilador parados, es decir, en condiciones libres de ruido, y en las pruebas de reconocimiento sólo se usaron las señales ruidosas.

En la etapa de parametrización, se dividió la señal de voz, previamente muestreada a 8 kHz, detectada manualmente y preenfatizada con $1-0.95z^{-1}$, en tramas de 30 ms a un ritmo de 15 ms y se caracterizó cada trama con L parámetros cepstrales enventanados, estimados bien mediante la técnica LPC convencional, bien mediante la representación OSALPC. En algunas pruebas también se calcularon los parámetros diferenciales de la trama. Cada información se cuantificó separadamente usando diccionarios de 64 palabras-código mediante CV convencional o etiquetado múltiple. Se caracterizó cada dígito mediante un MOM de izquierda a derecha, sin saltos, de 10 estados.

Los primeros experimentos llevados a cabo consistieron en optimizar empíricamente el orden de predicción y el tipo de ventana cepstral con CV convencional y sin utilizar parámetros diferenciales. Resultados de reconocimiento preliminares habían mostrado que ni el orden del modelo ni la ventana cepstral son importantes en esta tarea en condiciones libres de ruido Sin embargo, en presencia de ruido las tasas de reconocimiento han resultado ser muy sensibles a ambos factores. Los resultados obtenidos usando la técnica LPC convencional y la representación OSALPC pueden verse en la tabla 1, en función de la velocidad del coche, para órdenes de predicción 8, 12 y 16 :

para las ventanas cepstrales seno realzado, rampa e inversa de la desviación típica (IDT).

Se obtuvieron los mejores resultados usando ventana IDT para la técnica LPC convencional y ventana rampa para OSALPC, y orden de predicción 16 en ambos casos. Es decir, un orden de predicción relativamente alto y una ventana cepstral no simétrica.

De hecho, un valor relativamente alto del orden de predicción puede proporcionar estimaciones más robustas de la autocorrelación en presencia de ruido de banda ancha debido a que la sensibilidad de la autocorrelación a este tipo de ruido tiende a decrecer con el retardo. Sin embargo, ordenes del modelo demasiado altos dan lugar a tasas de reconocimiento pobres debido a la aparición de picos espurios en las estimaciones espectrales. Por otro lado, una ventana cepstral no simétrica es más apropiada en presencia de ruido de banda ancha debido a que los coeficientes cepstrales de menor índice están más afectados por este tipo de ruido que los coeficientes de mayor índice.

En cuanto a la comparación entre ambas técnicas de parametrización, la representación OSALPC supera notablemente la técnica LPC convencional en condiciones severas de ruido. Por contra, en condiciones poco ruidosas, las tasas de la técnica LPC convencional son mejores que las de la representación OSALPC debido a que esta última técnica realiza una deconvolución imperfecta de la señal de voz.

Respecto a los parámetros filtrados en el tiempo, el uso del delta-cepstrum y la delta-energía, en el caso de la técnica LPC convencional, y el uso del delta-cepstrum, en el caso de la técnica OSALPC, proporciona resultados excelentes. Los mejores resultados se obtuvieron usando una ventana de 240 ms de duración para la estimación de los parámetros filtrados.

También se obtuvieron resultados excelentes aplicando etiquetado múltiple en lugar de CV convencional. Además, estos resultados superaron siempre a los obtenidos con MOMSC.

La combinación de todas estas técnicas proporcionó todavía mejores resultados que los obtenidos aplicando cada técnica por separado. En la tabla 2 se comparan las tasas de reconocimiento obtenidas con el orden y las ventanas cepstrales óptimos en función del tipo de parametrización -LPC u OSALPC- y cuantificación vectorial -convencional (MOMD) o etiquetado múltiple (MOMEM)- empleadas, y utilizando o no parámetros filtrados en el tiempo (Δc, ΔE), además del cepstrum estático (c).

Como puede observarse en la tabla 2, la técnica OSALPC obtiene resultados excelentes en condiciones severas de ruido, pero los resultados de la técnica LPC convencional son mejores que los de la representación OSALPC en condiciones casi libres de ruido (0 km/h) cuando no se usan parámetros filtrados en el tiempo. Sin embargo, utilizando delta-cepstrum, la técnica OSALPC supera a la LPC convencional en todas las condiciones consideradas. Los mejores resultados se obtuvieron utilizando parametrización OSALPC. delta-cepstrum y etiquetado múltiple.

6. Conclusiones

A partir de la aplicación de la parametrización OSALPC, el filtrado de parámetros y el etiquetado múltiple al reconocimiento de habla en entorno ruidoso de coche, utilizando un sistema basado en cuantificación vectorial y modelos ocultos de Markov, han podido extraerse las siguientes conclusiones a) Cuando se utilizan técnicas de predicción lineal, son convenientes órdenes d predicción relativamente altos y ventanas cepstrales no simétricas; b) L representación cepstral basada en la predicción lineal de la parte causal de l autocorrelación (OSALPC) proporciona resultados excelentes en condicione severas de ruido; c) La inclusión de parámetros filtrados en el tiempo es mu útil en todas las condiciones consideradas; d) El etiquetado múltiple supe notablemente a la cuantificación vectorial convencional; e) Se obtienen todav mejores resultados combinando estas técnicas que aplicándolas por separado.

Referencias

[1] B.H. Juang, "Speech Recognition in Adverse Environments", Computer Speech and Language, vol. 5, pp. 275-294, 1991.

[2] F. Itakura, "Minimum prediction residual principle apllied to speech recognition", IEEE Trans. ASSP, vol. 23, pp. 67-72, 1975.

[3] J. Hernando, C. Nadeu, E. Lleida, "On the AR Modeling of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition", Proc. ICSLP'92, Banff, octubre de 1992, pp. 1593-1596.

[4] B. H. Juang, L.R. Rabiner, J.G. Wilpon, "On the Use of Banspass Liftering in Speech Recognition", IEEE Trans. ASSP, vol. 35, pp. 947-954, 1987.

[5] B.A. Hanson, "Spectral Slope Based Distortion Measures for All-Pole Models of Speech", H. Wakita, IEEE Trans. ASSP, vol. 35, pp. 968-973, 1987.

[6] Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition", IEEE Trans. ASSP, vol. 35, pp. 1414-1422, 1987.

[7] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. ASSP, vol. 34, pp. 52-59, 1986.

[8] J. Hernando, J.B. Mariño, C. Nadeu, "Multiple Multilabeling to Improve HMM-Based Speech Recognition in Noise", Proc. EUROSPEECH'93, Berlin, septiembre de 1993, pp. 1643-1646.

[9] J. Hernando, "Técnicas de Procesado y Representación de la Señal de Voz para el Reconocimiento del Habla en Ambientes Ruidosos", Tesis Doctoral, Dpto. Teoría de la Señal y Comunicaciones, Universidad Politécnica de Cataluña, Barcelona, mayo de 1993.

[10] M.A. Lagunas, M. Amengual, "Non-Linear Spectral Estimation", Proc. ICASSP'87, Dallas, abril de 1987, pp. 2035-2038.

[11] D.P. McGinn, D.H. Johnson, "Reduction of All-Pole Parameter Estimation Biass by Successive Autocorrelation", Proc. ICASSP'83, Boston, April 1983, pp. 1088-1091.

[12] D. Mansour, B.H. Juang, "The Short-Time Modified Coherence Representation and Noisy Speech Recogniton", IEEE Trans. ASSP, vol. 37, pp. 795-804, 1989.

[13] C. Nadeu, B.H. Juang, "Filtering of Spectral Parameters for Speech Recogniton", Proc. ICSLP'94, Yokohama, Septiembre 1994, pp. 1927-30.

[14] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, vol. 77, 1989.

[15] H.P. Tseng, M.J. Sabin, E.A. Lee, "Fuzzy Vector Quantization Applied to Hidden Markov Modeling", Proc. ICASSP-87, Dallas, abril de 1987, pp. 641-644.

[16] M. Nishimura, "HMM-Based Speech Recognition Using Multi-Dimensional Multi-Labeling", K. Toshioka, Proc. ICASSP'87, Dallas, April 1987, pp. 1163-1166.

[17] X.D. Huang, "Phoneme Classification Using Semicontinuous Hidden Markov Models", IEEE Trans. ASSP, vol. 40, pp. 1062-1067, 1992.

		LPC			OSALPC		
Orden	Vent. ceps./Veloc.	0 km/h	70 km/h	130 km/h	0 km/h	70 km/h	130 km/h
8	Seno realzado	93.7	88.9	58.2	91.2	83.4	71.7
	Rampa	93.2	85.1	59.7	91.7	85.1	68.0
	IDT	93.0	84.5	61.2	93.5	82.6	72.2
12	Seno realzado	96.7	93.9	71.0	96.7	89.3	74.5
	Rampa	93.2	91.4	75.2	91.0	87.1	76.2
	IDT	95.2	84.1	60.0	95.5	87.9	77.2
16	Seno realzado	90.7	86.1	66.2	92.7	85.5	69.7
	Rampa	92.7	85.6	72.0	96.0	94.6	85.0
	IDT	97.5	92.1	79.0	95.5	91.2	80.7

Tabla 1: Tasas de reconocimiento obtenidas utilizando cepstrum LPC y OSALPC con varios órdenes de predicción y ventanas cepstrales.

Param.	Modelos	Parámetros / Veloc.	0 km/h	70 km/h	130 km/h
LPC	MOMD	c	97.5	92.1	79.0
LPC	MOMEM	c	98.2	94.9	81.7
OSALPC	MOMD	c	96.0	94.7	85.0
OSALPC	MOMEM	c	97.7	92.1	91.2
LPC	MOMD	c, Δc, ΔE	98.5	96.6	92.0
LPC	MOMEM	c, Δc, ΔE	99.2	96.6	94.0
OSALPC	MOMD	c, Δc	99.5	96.1	95.5
OSALPC	MOMEM	c, Δc	99.5	98.1	95.0

Tabla 2: Comparación de varias combinaciones de técnicas.

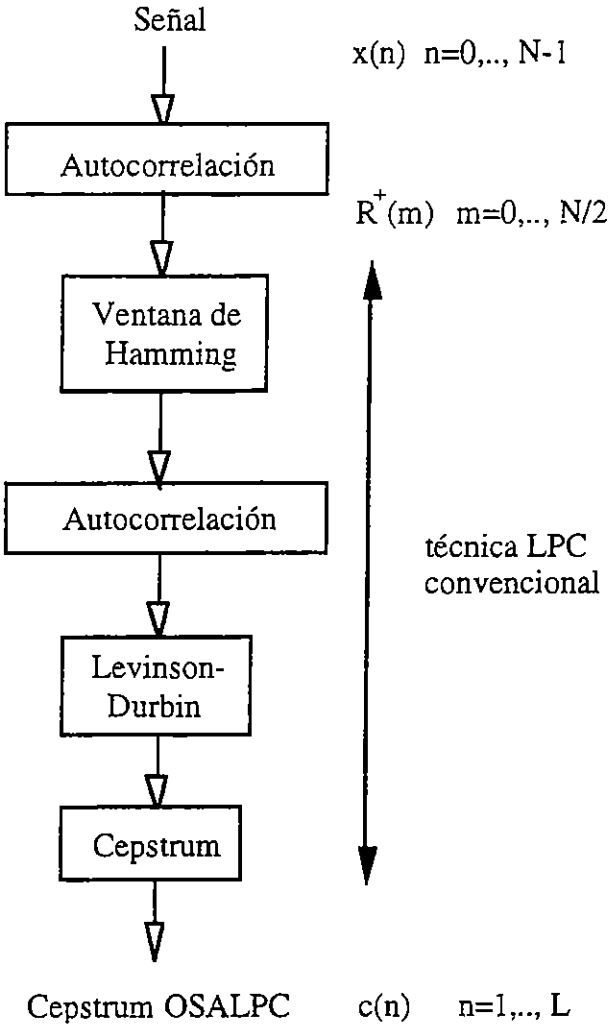

Figura 1: Diagrama de bloques para el cálculo del cepstrum OSALPC.

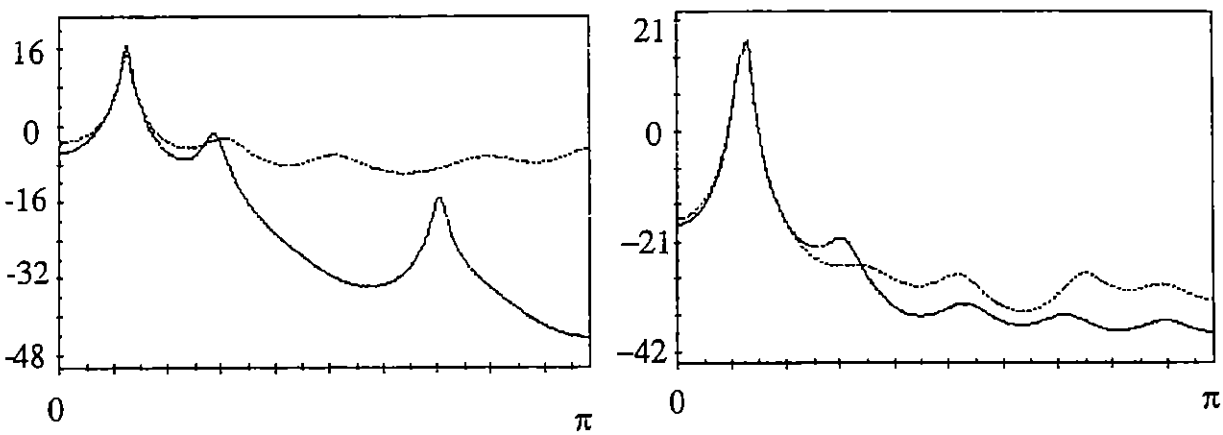

Figura 2: Robustez de la representación OSALPC al ruido blanco aditivo: a) espectro LPC y b) cuadrado de la envolvente OSALPC de una trama de voz sonora en condiciones libres de ruido (línea continua) y relación señal a ruido de 0 dB (línea de puntos).