# RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS MEDIANTE MODELOS OCULTOS DE MARKOV DISCRETOS

J. Hernando, J. Cobo y C. Nadeu

Dpto. Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña

**ABSTRACT:**

Speech recognition in noisy environments remains an unsolved problem, even in the case of isolated word recognition with small vocabularies. Recently, several techniques have been proposed to alleviate this problem. Concretely, the Short-Time Modified Coherence (SMC) parameterization and the Cepstral Projection Distortion (CPD) measure have shown excellent results when tested in a speech recognition system based on Dynamic Time Warping (DTW) and using speech contaminated by additive white noise. In this paper, a new technique based on the AR modeling of the one-sided autocorrelation sequence (OSALPC) is presented and, from a comparative study of these LPC-based techniques in the discrete Hidden Markov Model (DHMM) approach, two main conclusions are attained: 1) the slope cepstral window and a relatively high model order are preferable, and 2) the cepstral representation based on the autocorrelation modeling achieves excellent results.

## 1. INTRODUCCION

El comportamiento de los sistemas actuales de reconocimiento del habla se degrada rápidamente en presencia de ruido de fondo cuando las etapas de entrenamiento y de test no pueden llevarse a cabo en las mismas condiciones ambientales. Por este motivo, en los últimos pocos años se han propuesto algunos métodos y algoritmos en varias etapas del proceso de reconocimiento [1], principalmente en las de extracción de características y medida de similitud, en la dirección de desarrollar un sistema que opere siempre robusta y fiablemente como si hubiera sido entrenado en las mismas condiciones de reconocimiento.

La técnica de predicción lineal (LPC), equivalente a un modelado autorregresivo de la señal, ha mostrado ser de gran utilidad en reconocimiento del habla [2] y, últimamente, se ha puesto de manifiesto que los coeficientes cepstrales del modelo, convenientemente ponderados y usando la distancia euclídea tradicional, ofrecen en general mejores prestaciones que cualquier otro tipo de parámetros asociados al modelo LPC en condiciones

libres de ruido [3]. Sin embargo, estas técnicas no han resultado ser robustas a cambios en las condiciones ambientales.

Para el reconocimiento del habla ruidosa, Hansen y Wakita [4] usaron la distancia euclídea sobre la derivada del logaritmo del espectro, suavizado cepstralmente, que es equivalente a una ponderación de tipo rampa truncada en la distancia euclídea cepstral. Dicha ponderación está relacionada con el hecho de que en presencia de ruido blanco o de banda ancha los coeficientes cepstrales de orden bajo son menos robustos que los de orden alto en el vector cepstral truncado.

Recientemente, Mansour y Juang [5] han propuesto una nueva técnica para el análisis espectral robusto de la señal de voz llamada Coherencia Modificada a Corto Plazo (SMC, "Short-Time Modified Coherence"), basada en el hecho bien conocido de que la secuencia de autocorrelación es menos afectada por el ruido que la señal original. Los mismos autores [6] consideraron, además, que no hay una razón obvia para mantener la propiedad de simetría de la distancia euclídea si se sabe que las señales de referencia y de test tienen diferentes grados de contaminación. Por ello, propusieron una familia de medidas de distancia basada en la operación proyección entre vectores, que tiene en cuenta los efectos del ruido blanco aditivo sobre el vector cepstral.

El propósito de esta comunicación es doble: por un lado, hacer un estudio comparativo de estas técnicas dentro del entorno de los modelos ocultos de Markov discretos; por otro lado, presentar la técnica de predicción lineal de la parte causal de la autocorrelación (OSALPC), inspirada en la representación SMC, como una parametrización robusta de la señal de voz en presencia de ruido.

Esta comunicación está organizada del siguiente modo. En el apartado 2 se introduce la nueva parametrización OSALPC, estableciéndose su relación con las técnicas LPC y SMC. En el apartado 3 se presentan las distancias y ventanas cepstrales que serán objeto de comparación. El apartado 4 está dedicado al uso de la energía y de los rasgos dinámicos de la señal de voz en reconocimiento del habla. En el apartado 5 se muestran los resultados de la aplicación de estas técnicas al reconocimiento de palabras aisladas en ambientes ruidosos usando modelos de Markov discretos, con el fin de comparar su comportamiento. Por último, en el apartado 6 se recogen las principales conclusiones del trabajo.

## 2. PREDICCION LINEAL DE LA PARTE CAUSAL DE LA AUTOCORRELACION (OSALPC)

A partir de la secuencia de autocorrelación R(n) definimos su parte causal como

$$R^{+}(n) = \begin{cases} R(n) & n>0 \\ R(0)/2 & n=0 \\ 0 & n<0 \end{cases} \qquad (1)$$

Su transformada de Fourier es el espectro complejo

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

donde $S(\omega)$ es el espectro, es decir, la transformada de Fourier de $R(n)$, y $S_H(\omega)$ es la transformada de Hilbert de $S(\omega)$. Debido a la analogía entre $S^{+}(\omega)$ y la señal analítica usada en modulación de amplitud, se puede definir una "envolvente" espectral [7] como

$$E(\omega) = |S^{+}(\omega)| \qquad (3)$$

Esta característica de envolvente, junto al alto rango dinámico del espectro de la señal de voz, origina que el cuadrado de la envolvente espectral $E^2(\omega)$, que es además el espectro de $R^{+}(n)$, sea más robusto al ruido que el propio espectro $S(\omega)$. Además, es un hecho bien conocido que $R^{+}(n)$ tiene los mismos polos y con la misma multiplicidad que la señal.

Ambas propiedades conducen a considerar la predicción lineal de $R^{+}(n)$ como una técnica robusta de representación de la señal de voz. Al igual que la técnica LPC standard asume un modelo todo polo para $S(\omega)$, esta nueva técnica -a la que hemos llamado OSALPC- equivale a suponer un modelo todo polo para $E^2(\omega)$. Ello da lugar a que nuestra técnica sólo realice una desconvolución parcial de la señal de voz [8].

La relación de nuestra técnica con la LPC standard es obvia: nuestra técnica consiste en aplicar el algoritmo LPC standard sobre $R^{+}(n)$. En cuanto a su relación con la técnica SMC [5], las principales diferencias son que esta última utiliza: 1) el estimador coherencia para calcular la secuencia de autocorrelación, mientras que nosotros usamos el clásico estimador sesgado; y 2) un conformador espectral para calcular las entradas al algoritmo de Levinson-Durbin. El uso de este conformador no se ve justificado en el desarrollo teórico que nos ha llevado a la propuesta de la técnica OSALPC ni en los resultados de reconocimiento[8].

Un aspecto importante de cualquiera de estas tres técnicas es el orden del modelo. Es un hecho bien conocido que los coeficientes de orden alto de la secuencia de autocorrelación son menos sensibles al ruido blanco o de banda ancha que los de orden bajo. Por ello, es conveniente órdenes relativamente altos en reconocimiento de habla ruidosa. Un orden excesivamente alto, sin embargo, provocará errores en la estimación ya que la fiabilidad estadística de los coeficientes de la secuencia de autocorrelación disminuye con el orden.

## 3. MEDIDAS DE DISTANCIA Y VENTANAS CEPSTRALES

La distancia euclídea entre dos vectores de coeficientes cepstrales LPC ponderados se define como

$$d_E = \sum_{n=1}^{L} [\, w(n) \, ( c_t(n) - c_r(n) ) \,]^2 \qquad (4)$$

donde $c_t(n)$ y $c_r(n)$ son los n-ésimos coeficientes cepstrales de las tramas de test y de referencia respectivamente, L es el número de coeficientes cepstrales y w(n) es la ponderación aplicada al n-ésimo coeficiente.

El conjunto de pesos w(n) define una ventana cepstral. Las dos más usadas en reconocimiento del habla son

Ventana seno : $w(n) = 1 + \frac{L}{2} \sin \left(\frac{\pi n}{L}\right)$ (5.a)

Ventana rampa: $w(n) = n$ (5.b)

propuestas en [3] y [4] respectivamente, donde n=1,....L. Si M denota el orden del modelo LPC, el valor de L es 3M/2 en la ventana seno y M en la ventana rampa.

Como resultado de la ponderación cepstral, se obtiene una versión suavizada del espectro que depende tanto del tipo de ventana como del orden del modelo. Uno de los propósitos del apartado 5 es obtener el grado óptimo de suavizado en condiciones ruidosas.

Recientemente, estudios analíticos y observaciones experimentales llevados a cabo por Mansour y Juang [6] revelan que, en presencia de ruido blanco aditivo, el mayor desajuste entre vectores cepstrales LPC limpios y ruidosos es la reducción de las normas. También muestran que los vectores cepstrales de norma más alta son menos sensibles que los vectores cepstrales de norma más baja y que el ángulo entre dos vectores cepstrales es menos sensible que la distancia euclídea tradicional. Estas consideraciones condujeron a la propuesta de una familia de distancias basadas en la operación de proyección entre vectores (CPD, "Cepstral Projection Distances"). Los mejores resultados en [6] se obtuvieron usando:

$$d_P = |\, C_t \,| \, ( 1 - \cos \beta ) = |\, C_t \,| - \frac{C_t^T C_r}{|\, C_r \,|} \qquad (6)$$

que es la distancia que usaremos en nuestro estudio comparativo.

## 4. PARAMETROS REGRESIVOS Y ENERGIA

Es sabido que las representaciones espectrales dinámicas de la señal de voz se ven menos afectadas por las variaciones de las condiciones ambientales y del canal de transmisión que las representaciones estáticas [9]. Por ello, paralelamente al desarrollo de nuevas representaciones espectrales, un objeto de este trabajo es profundizar en la implementación de los parámetros regresivos: algoritmo de estimación, intervalo de estimación y número óptimo de parámetros.

En cuanto al algoritmo de estimación, se han considerado cuatro posibilidades: 1) utilización de los coeficientes de regresión de orden n para estimar el parámetro regresivo $\Delta_n c$ [10]; 2) utilización de los coeficientes de regresión de primer orden para estimar $\Delta c$ y aplicación iterativa de los mismos para obtener los parámetros de órdenes superiores; 3) aproximación polinómica [11]; y 4) truncamiento de la respuesta impulsional del derivador discreto ideal. En cuanto al intervalo de estimación de cada parámetro regresivo y el número óptimo de parámetros, como se verá, se han optimizado en función de los resultados de reconocimiento.

Por otro lado, se ha comprobado que el uso de la energía de la trama y sus correspondientes parámetros regresivos también contribuye al robustecimiento del sistema. Intentaremos reflejar a continuación las prestaciones de cada una de estas técnicas desde el punto de vista de mejora de la tasa de reconocimiento en presencia de ruido.

## 5. RESULTADOS EXPERIMENTALES

Este apartado muestra la aplicación de las técnicas descritas anteriormente al reconocimiento multilocutor de palabras aisladas en el entorno de los modelos ocultos de Markov discretos y en presencia de ruido blanco aditivo.

### 5.1. Sistema de reconocimiento y base de datos

En primer lugar, la señal de voz fue filtrada de 100 a 3400 Hz. con un filtro "antialiasing", muestreada a 8 KHz y codificada con dos bytes por muestra. La señal digital obtenida fue marcada manualmente para determinar el inicio y el fin de cada palabra. Las marcas obtenidas de este modo fueron usadas en todos los experimentos con el fin de eliminar los errores debidos al problema de la detección. Se entrenó siempre el sistema usando señal de voz libre de ruido. La señal de voz ruidosa se simuló añadiendo ruido blanco gaussiano de media cero a la señal limpia de manera que se obtuviese la SNR deseada.

En la etapa de parametrización del sistema de reconocimiento, la señal se dividió en tramas de 30 ms. de duración con un desplazamiento de 15 ms.. y cada trama se caracterizó por L parámetros cepstrales y, en su caso, la energía y los correspondientes parámetros regresivos. Cuando se hizo uso de varias informaciones, cada una de ellas se cuantificó independientemente. Cada palabra se caracterizó por un modelo de Markov discreto y las fases de entrenamiento y test se realizaron mediante los algoritmos de Baum-Welch y Viterbi, respectivamente.

La base de datos empleada consiste en diez repeticiones de los dígitos catalanes correspondientes a 7 hombres y 3 mujeres (1000 palabras, en total), que se han considerado pronunciadas en un ambiente libre de ruido. Las pruebas se han llevado a cabo dividiendo la base de datos en dos bloques, cada uno de ellos con la mitad de repeticiones de cada dígito y locutor, de forma que las señales a reconocer no tomaran parte en el entrenamiento. Por problemas de coste computacional, sólo se hizo una partición de la base y se realizaron dos pruebas alternando los papeles de cada uno de los bloques.

Utilizando únicamente la información espectral, estimada mediante la técnica LPC standard, se hicieron unas primeras pruebas para optimizar el número de estados de cada modelo así como el tamaño del codebook. Pruebas preliminares aconsejaban para este tipo de aplicaciones la no utilización de preénfasis y el uso de la ventana cepstral rampa de longitud 12 [8] y de un codebook de 64 codewords [12]. Tras varias pruebas de reconocimiento realizadas con estas restricciones, el compromiso coste computacional-tasa de reconocimiento nos llevó a considerar modelos de izquierda a derecha de 10 estados sin posibilidad de transición entre estados no consecutivos. Posteriormente, fijada la estructura de los modelos intentamos optimizar el tamaño del codebook, pero de nuevo obtuvimos los mejores resultados para 64 codewords.

### 5.2. Parametrización LPC y distancia euclídea

Como ya hemos mencionado, pruebas preliminares aconsejaban el uso de una ventana rampa de longitud 12 y, por tanto, de un orden del modelo LPC igual a 12. Para reafirmar la conveniencia de esta ventana cepstral y de este orden del modelo en nuestra aplicación concreta, realizamos varias pruebas de reconocimiento. En la tabla I se observan las tasas de reconocimiento en tanto por ciento y para diferentes SNR obtenidas usando ventana rampa y varios órdenes, y en la tabla II las correspondientes a orden 12 y distintas ventanas (rect.= rectangular, var.= inversa de la desviación típica). Estos resultados nos confirmaron en el uso de la ventana rampa de longitud 12.

Seguidamente, procedimos al uso de los parámetros regresivos y de la energía para robustecer el sistema de reconocimiento. Las mejores prestaciones se obtuvieron cuando se

estimaron los parámetros regresivos aplicando iterativamente los coeficientes de regresión de primer orden. Por otro lado, se observó que la incorporación de parámetros regresivos de orden superior mejoraba la tasa de reconocimiento si bien cada vez de forma más inapreciable. El compromiso coste computacional-tasa de reconocimiento nos decidió a usar sólo los dos primeros parámetros regresivos del cepstrum y de la energía. En cuanto al intervalo de estimación, los mejores resultados se obtuvieron usando un intervalo de 90 ms. para el primer parámetro regresivo y de 120 ms. para el segundo. Por otro lado, el tamaño de los codebooks se fijo a 64 para todas la informaciones tras las correspondientes pruebas de optimización. Los resultados obtenidos se recogen el la tabla III. En ella se observa cómo mejora ostensiblemente la tasa de reconocimiento al añadir más informaciones al sistema.

Tabla I

| L | ∞dB | 20dB | 10dB | 0dB |
|---|---|---|---|---|
| 8 | 99.7 | 95.7 | 72.3 | 34.1 |
| 10 | 99.9 | 97.6 | 85.3 | 43.8 |
| 12 | 99.8 | 98.9 | 89.5 | 54.2 |
| 14 | 99.8 | 97.8 | 83.0 | 52.6 |
| 16 | 99.8 | 93.2 | 70.7 | 41.2 |

Tabla II

| vent. | ∞dB | 20dB | 10dB | 0dB |
|---|---|---|---|---|
| seno | 99.7 | 96.2 | 73.7 | 29.0 |
| rect. | 99.8 | 66.1 | 34.0 | 22.8 |
| rampa | 99.8 | 98.9 | 89.5 | 54.2 |
| var. | 99.6 | 97.9 | 82.1 | 41.7 |

Tabla III

| codebook | ∞dB | 20dB | 10dB | 0dB |
|---|---|---|---|---|
| c | 99.8 | 98.9 | 89.5 | 54.2 |
| c,Δc | 99.9 | 99.9 | 94.4 | 60.2 |
| c,Δc,Δ2c | 99.9 | 99.9 | 95.2 | 61.4 |
| c,Δc,Δ2c,e | 99.9 | 99.9 | 95.5 | 63.6 |
| c,Δc,Δ2c,e,Δe | 100 | 99.9 | 96.0 | 66.6 |
| c,Δc,Δ2c,e,Δe,Δ2e | 100 | 100 | 96.1 | 69.6 |

### 5.3 Parametrización OSALPC y distancia de proyección

En la tabla IV se recogen los resultados de reconocimiento obtenidos utilizando la parametrización OSALPC y la distancia de proyección, usando 1 codebook (c) y 6 codebooks (c,Δc,Δ2c,e,Δe,Δ2e), y comparándolos con los obtenidos con la parametrización LPC y la distancia euclídea. Se han realizado pruebas con todas las combinaciones posibles de parametrizaciones, distancias y número de codebooks; excepto OSALPC con distancia de proyección y 6 codebooks, ya que en este caso no era de esperar una mejora de los resultados. Todas las pruebas se hicieron sin preénfasis, con ventana rampa de longitud 12 y codebooks independientes de 64 codewords cada uno.

Tabla IV

| param. | dist. | nº cdb. | ∞ dB | 20 dB | 10 dB | 0 dB |
|---|---|---|---|---|---|---|
| LPC | euc. | 1 | 99.8 | 98.9 | 89.5 | 54.2 |
| LPC | proy. | 1 | 99.8 | 98.9 | 90.0 | 56.5 |
| OSALPC | euc. | 1 | 98.6 | 97.7 | 93.1 | 73.1 |
| OSALPC | proy | 1 | 98.4 | 97.5 | 93.0 | 70.5 |
| LPC | euc. | 6 | 100 | 100 | 96.1 | 69.6 |
| LPC | proy. | 6 | 99.9 | 99.9 | 95.7 | 63.4 |
| OSALPC | euc. | 6 | 99.9 | 99.8 | 95.6 | 63.0 |

## 6. CONCLUSIONES

En esta comunicación hemos presentado una nueva técnica de parametrización robusta de la señal de voz, para el reconocimiento del habla ruidosa, consistente en la predicción lineal de la parte causal de la secuencia de autocorrelación (OSALPC), inspirada en la "Coherencia Modificada a Corto Plazo" propuesta en [5].

Después de un estudio comparativo de esta nueva técnica con la parametrización LPC standard y otras técnicas de robustez propuestas recientemente, como la distancia de proyección y el uso de la energía y los parámetros regresivos, se ha concluido que en reconocimiento de habla ruidosa:

- cuando se usa la parametrización LPC standard, son preferibles un orden de modelo relativamente alto, una ventana cepstral de tipo rampa y el uso de la energía y de parámetros regresivos;

- la distancia de proyección usada en el codebook de los DHMMs, tal como la hemos implementado, no proporciona una mejora ostensible;

- la representación cepstral basada en el modelado autorregresivo en el dominio de la autocorrelación alcanza excelentes resultados (los mejores de la comunicación) con distancia euclídea en condiciones severas de ruido, pero cuando se usan varias informaciones los resultados son peores que los de la parametrización LPC en las mismas condiciones.

### REFERENCIAS


[1] B.H. Juang, Computer Speech and Language, 1991, vol. 5, pp. 275-294.
[2] F. Itakura, IEEE Trans. ASSP, vol. 23, 1975, pp. 67-72.
[3] B. H. Juang, L.R. Rabiner y J. G. Wilpon, IEEE Trans. ASSP, vol. 35, 1987, pp. 947-954.
[4] B.A. Hansen y H. Wakita, IEEE Trans. ASSP, vol. 35, 1987, pp. 968-973.
[5] D. Mansour y B.H. Juang, IEEE Trans. ASSP, vol. 37, 1989, pp. 795-804.
[6] D. Mansour y B. H. Juang, IEEE Trans. ASSP, vol. 37, 1989, pp. 1659-1671.
[7] M.A. Lagunas y M. Amengual, Proc. ICASSP-87, Dallas, pp. 2035-2038.
[8] J. Hernando y C. Nadeu , Proc. EUROSPEECH-91, Génova, pp. 91-94.
[9] S. Furui, IEEE Trans. ASSP, vol. 34, 1986, pp. 52-59.
[10] B.A. Hanson y T.H. Applebaum, Proc. ICASSP-90, Albuquerque, pp. 857-860.
[11] J.G. Wilpon, C.H. Lee y L.R. Rabiner, Proc. ICASSP-91, Toronto, pp. 349-352.
[12] J.B. Mariño et al, Proc. EUSIPCO-90, Barcelona, pp. 1363-1366.