TECNICAS DE MODELADO AR ROBUSTO DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS

J. Hernando, D. Riu y C. Nadeu

Departamento de Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña
E.T.S.I. Telecomunicación, Apdo. 30002, 08080 Barcelona

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. 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 several LPC-based techniques in the discrete Hidden Markov Model (DHMM) approach, two main conclusions are attained when the speech is contaminated by additive white noise: 1) the slope cepstral window and a relatively high model order are preferable, and 2) the cepstral representation of the speech signal 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, Hanson 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.

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 presentan la distancia cepstral y las ponderaciones que serán objeto de comparación. En el apartado 3 se introduce la nueva parametrización OSALPC, estableciéndose su relación con las técnicas LPC y SMC. En el apartado 4 se muestran los resultados de la aplicación de estas técnicas al reconocimiento de palabras aisladas en ambientes ruidosos usando modelos de Markov discretos cuando la señal está contaminada con ruido blanco aditivo. Por último, en el apartado 5 se recogen las principales conclusiones del trabajo.

2. MEDIDA DE DISTANCIA Y VENTANAS CEPSTRALES

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

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

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

Ventana rampa: $w(n) = n$ (2.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 de esta comunicación es obtener el grado óptimo de suavizado en condiciones ruidosas.

3. PREDICCION LINEAL DE LA PARTE CAUSAL DE LA AUTOCORRELACION

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

$$R^+(n) = \begin{cases} R(n) & n>0 \\ R(0)/2 & n=0 \\ 0 & n<0 \end{cases} \qquad (3)$$

Su transformada de Fourier es el espectro complejo

$$S^+(\omega) = \frac{1}{2}\left[S(\omega) + jS_H(\omega)\right] \qquad (4)$$

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 [6] como

$$E(\omega) = |S^+(\omega)| \qquad (5)$$

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 [7].

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 [7].

4. 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.

4.1. 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 cuantificada 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 obtenidos mediante una de las técnicas de predicción lineal expuestas en el apartado anterior. Posteriormente, se aplicó sobre estos parámetros una cuantificación vectorial. Para ello, se utilizó un codebook de 64 codewords entrenado mediante el algoritmo de Lloyd. El tamaño del codebook había sido optimizado en pruebas preliminares de reconocimiento.

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. 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.

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.

4.2. Resultados de reconocimiento

Los primeros experimentos llevados a cabo consistieron en optimizar empíricamente el orden del modelo y el tipo de ponderación cepstral usando la parametrización standard LPC. Los resultados preliminares mostraron que ni el orden del modelo ni el tipo de ponderación cepstral eran importantes en nuestra aplicación en condiciones libres de ruido. Sin embargo, en presencia de ruido blanco aditivo los resultados de reconocimiento resultaron ser muy sensibles a ambos factores. Los mejores se obtuvieron con orden del modelo igual a 12 y ponderación cepstral de tipo rampa. Para ilustrar dicha sensibilidad, en la tabla I se muestran 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).

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

En la tabla III se muestran los resultados de reconocimiento obtenidos utilizando la parametrizaciones SMC y OSALPC y se comparan con los correspondientes a la parametrización LPC, usando orden del modelo y ponderación cepstral óptimos en la parametrización LPC. Como puede verse, en condiciones severas de ruido las técnicas SMC y OSALPC obtienen resultados superiores a los obtenidos con la parametrización LPC. En especial, a pesar de su sencillez, la técnica propuesta en esta comunicación, la parametrización OSALPC, obtiene excelentes resultados.

Tabla III

param.	∞ dB	20 dB	10 dB	0 dB
LPC	99.8	98.9	89.5	54.2
SMC	99.0	97.0	89.2	67.5
OSALPC	98.6	97.7	93.7	75.9

5. 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). Después de un estudio comparativo de esta nueva técnica con la parametrización LPC standard y de diferentes órdenes de modelo y ponderaciones cepstrales, se ha concluido que en reconocimiento de habla ruidosa, en el caso de ruido blanco aditivo:

- cuando se usa la parametrización LPC standard, son preferibles un orden de modelo relativamente alto y una ventana cepstral de tipo creciente, debido a que los coeficientes de autocorrelación y del cepstrum de orden alto son más robusto que los de orden bajo;

- la representación cepstral basada en el modelado autorregresivo de la parte causal de la autocorrelación alcanza excelentes resultados en condiciones severas de ruido, utilizando orden de modelo y ponderación cepstral óptimos en la parametrización standard LPC.

6. REFERENCIAS

[1] B.H. Juang, "Speech recognition in adverse conditions", Computer Speech and Language, 1991, vol. 5, pp. 275-294.

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

[3] B. H. Juang, L.R. Rabiner y J. G. Wilpon, "On the use of bandpass liftering in speech recognition", IEEE Trans. ASSP, vol. 35, 1987, pp. 947-954.

[4] B.A. Hanson y H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", IEEE Trans. ASSP, vol. 35, 1987, pp. 968-973.

[5] D. Mansour y B.H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", IEEE Trans. ASSP, vol. 37, 1989, pp. 795-804.

[6] M.A. Lagunas y M. Amengual, "Non-linear spectral estimation", Proc. ICASSP-87, Dallas, pp. 2035-2038.

[7] J. Hernando y C. Nadeu, "A comparative study of parameters and distances for noisy speech recognition", Proc. EUROSPEECH-91, Génova, pp. 91-94.