COMPORTAMIENTO DE LA TRANSFORMACION BILINEAL DE FRECUENCIAS EN RECONOCIMIENTO DE HABLA RUIDOSA

J. Hernando, C. Nadeu y D. Riu

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

1. INTRODUCCION

Los actuales sistemas de reconocimiento automático del habla experimentan una degradación importante cuando han de desenvolverse en ambientes ruidosos. En los últimos pocos años se han propuesto técnicas para aliviar el problema [1]. Sin embargo, el reconocimiento de habla ruidosa no ha encontrado una solución satisfactoria incluso en el caso de palabras aisladas y vocabularios reducidos.

Una posible aproximación al problema consiste en encontrar representaciones de la señal de voz que sean robustas en sí mismas a los ambientes ruidosos. Ello está motivado por el hecho de que la técnica de predicción lineal (LPC), ampliamente usada en reconocimiento [2] y procesado de habla en general, ha mostrado ser muy sensible a la presencia de ruido aditivo. En esta aproximación se pueden distinguir dos tipos de técnicas: uno que intenta realizar un análisis espectral robusto de la señal de voz desde el punto de vista de procesado de la señal y otro que trata de emular la capacidad auditiva humana, basándose en el hecho bien conocido de que nuestro oído parece percibir la voz mejor que cualquier máquina en presencia de ruido interferente sin un conocimiento previo ni de la voz ni del ruido.

Dentro de este último tipo de técnicas, una forma de emular la capacidad auditiva humana es realizar una transformación de la escala de frecuencias que aproxime la sensibilidad frecuencial logarítmica del oído humano, lo cual puede implementarse fácilmente con una transformación bilineal en el plano de las frecuencias complejas. El propósito de esta comunicación es el estudio del comportamiento de la transformación bilineal en el reconocimiento del habla ruidosa. En el siguiente apartado, se concretará la definición y la implementación de dicha transformación. Seguidamente, en el apartado 3 se expondrán los resultados de su aplicación al reconocimiento de palabras aisladas mediante modelos ocultos de Markov y se discutirán los resultados.

2. LA TRANSFORMACION BILINEAL

Es ampliamente conocida la escala Mel como una aproximación a la escala logarítmica de percepción del oído humano, dada por la relación:

$$m = 6 \log \left[(f/600) + \sqrt{1 + (f/600)^2} \right] \qquad (1)$$

donde f está en Hertzios y m en Barks. La obtención del espectro de una señal en la escala Mel puede realizarse a través de un banco de filtros paso-banda con anchos de banda distribuidos uniformemente en la escala Mel sobre el rango de frecuencias deseado [3].

Una alternativa a este proceso fue propuesta en [4], basándose en la transformación bilineal para aproximar la escala de frecuencias Mel. La transformación bilineal es una transformación definida sobre el plano complejo z, que realiza una transformación no lineal sobre el eje de frecuencias, según las relaciones

$$Z^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \qquad |\alpha| < 1 \qquad (2.a)$$

$$\operatorname{tg} \omega = \frac{(1 - \alpha^2) \operatorname{sen} \phi}{-2\alpha + (1 + \alpha^2) \cos \phi} \qquad (2.b)$$

donde $z = e^{j\phi}$ y $Z = e^{j\omega}$ son las circunferencias de radio unidad en el plano original y transformado, respectivamente, y α es un parámetro que controla la transformación. Valores positivos de α producen una expansión de la zona baja de frecuencias y una compresión de la zona alta. Puede comprobarse que para frecuencia de muestreo 8 KHz. la transformación bilineal con $\alpha = 0.4$ corresponde aproximadamente a la escala Mel.

La ventaja de la utilización de la transformación bilineal en lugar de la transformación directa expuesta anteriormente es que permite obtener una expresión matricial, cuyos elementos pueden calcularse recursivamente, para la transformación de los coeficientes cepstrales. Esto es muy importante ya que últimamente se ha puesto de manifiesto que los coeficientes cepstrales correspondientes al espectro LPC, convenientemente ponderados y usando la distancia euclídea tradicional, ofrecen en general mejores prestaciones

que cualquier otro tipo de parámetros asociados al modelo LPC [5].

Para evaluar el comportamiento de la transformación bilineal en reconocimiento de habla ruidosa, compararemos en el siguiente apartado los resultados de reconocimiento al usar directamente los coeficientes cepstrales del modelo LPC con los obtenidos al aplicar la transformación bilineal sobre estos mismos coeficientes.

3. RESULTADOS EXPERIMENTALES

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. Con ella, se han realizado pruebas multilocutor de forma que las señales a reconocer no tomaran parte en el entrenamiento.

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. No se realizó preénfasis. 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, la señal se dividió en tramas de 30 ms. con un desplazamiento de 15 ms. y para cada trama se calculó su cepstrum según un modelo LPC de orden 12. Posteriormente, se aplicó o no la transformación bilineal sobre el cepstrum, se procedió a su enventanado y, finalmente, se cuantificó mediante un VQ de 64 símbolos. Cada palabra se caracterizó por un modelo de Markov discreto de 10 estados, sin posibilidad de transición entre estados no consecutivos, y las fases de entrenamiento y test se realizaron mediante los algoritmos de Baum-Welch y Viterbi, respectivamente. Ver más detalles sobre el sistema en [6].

En la tabla I se muestran los tasas de reconocimiento en tanto por ciento obtenidas al usar directamente los coeficientes cepstrales LPC con diversas ventanas de ponderación cepstral y en la tabla II las obtenidas utilizando la transformación bilineal con $\alpha = 0.4$ (rect.= rectangular, in.desv. = inversa de la desviación típica).

Tabla I

vent.	∞dB	20dB	10dB	0dB
rect.	99.8	66.1	34.0	22.8
seno	99.7	96.2	73.7	29.0
in.desv.	99.7	97.8	84.0	41.8
rampa	99.8	98.9	89.5	54.2

Tabla II

vent.	∞dB	20dB	10dB	0dB
rect.	99.9	96.8	78.6	38.4
seno	100.0	97.5	74.8	24.6
in.desv.	99.7	98.3	84.8	42.5
rampa	99.9	69.1	34.3	20.4

A la vista de estos resultados, podemos observar que en ausencia de ponderación cepstral (ventana rectangular), la transformación bilineal robustece al cepstrum frente al ruido blanco aditivo. Ello es debido a que la transformación bilineal expande la zona de bajas frecuencias que es donde la señal de voz tiene más energía y, por lo tanto, es más robusta a este tipo de ruido. Sin embargo, cuando se utilizan las ponderaciones cepstrales usuales -seno e inversa de la desviación típica- la transformación bilineal no parece ayudar al reconocimiento de habla ruidosa. Por último, en el caso de la ventana rampa los resultados empeoran con la transformación bilineal. Ello puede ser debido a que la ventana rampa ha sido diseñada para el reconocimiento de habla ruidosa teniendo en cuenta la sensibilidad al ruido relativa de los diferentes coeficientes cepstrales del modelo LPC en ausencia de transformación bilineal.

Por otro lado, utilizando la ventana inversa de la desviación típica de cada coeficiente, hemos intentado optimizar la tasa de reconocimiento en función de α. Como puede observarse en la tabla III, no se han conseguido superar los resultados obtenidos con la ventana rampa y sin transformación bilineal.

Tabla III

α	∞dB	20dB	10dB	0dB
0.00	99.7	97.8	84.0	41.8
0.05	99.8	98.4	87.1	51.0
0.10	99.8	98.5	87.6	50.4
0.20	99.8	98.8	84.0	43.3
0.30	99.8	97.5	85.9	43.2
0.40	99.7	98.3	84.8	42.5
0.50	99.9	96.8	74.3	33.9

REFERENCIAS

[1] B.H. Juang, Computer Speech and Language, 1991, vol. 5, pp. 275-94.
[2] F. Itakura, IEEE Trans. ASSP, vol. 23, 1975, pp. 67-72.
[3] S. B. Davis y P. Mermelstein, IEEE Trans. ASSP, vol. 24, 1980, pp. 357-366.
[4] K. F. Lee, PhD Thesis, Computer Science Department, Carnegie Mellon University, 1988.
[5] B. H. Juang, L.R. Rabiner y J. G. Wilpon, IEEE Trans. ASSP, vol. 35, 1987, pp. 947-954.
[6] J. Hernando y C. Nadeu , Proc. EUROSPEECH-91, Génova, pp. 91-94.