Predicción Lineal de la Parte Causal de la Autocorrelación para la Identificación del Locutor en Ambientes Ruidosos

C. Villagrasa, J. Hernando, C. Nadeu, Josep M. Salavedra
Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Ap. 30002, 08071 Barcelona
Telf. (93) 4016440, Fax: (93) 4016447, E-mail: javier@tsc.upc.es

Abstract.- Recently, a new parametrization technique based on the AR modelling of the one-sided autocorrelation sequence (OSALPC) has shown to be attractive for speech recognition because of its simplicity and its high recognition perfomance in noisy conditions. In this paper, that new parametrization technique is proposed to speaker identification in noisy enviroment. Experimental results obtained with a new speaker identification system based on the statistics of the cepstrals vectors show that OSALPC also achieves much better results than standard parametrization techniques.

1.- Introducción

El comportamiento de los sistemas actuales de reconocimento del habla y del locutor se degrada rápidamente en presencia de ruido cuando el entrenamiento y el test no han podido ser realizados bajo las mismas condiciones. Por este motivo, en los últimos años se han propuesto algunos métodos y algoritmos en varias etapas del proceso de reconocimento. Sin embargo el problema del reconocimento del habla y del locutor en ambientes ruidosos aún no ha sido resuelto.

Uno de los intentos para combatir el problema del ruido consiste en nuevas parametrizaciones del habla que sean invariantes o resistentes a la corrupción ruidosa, para substituir las técnicas de parametrización convencionales, tales como LPC o mel-cepstrum, de las cuales es sabido que son muy sensibles a la presencia de ruido.

Recientemente, Hernando y Nadeu propusieron una técnica de parametrización alternativa llamada Predicción Lineal de la Parte Causal de la Autocorrelación (OSALPC) [1] para el reconocimiento del habla en presencia de ruido. Esta técnica, muy relacionada con la representación de la señal de voz llamada Coherencia Modificada a Corto Plazo (SMC,"Short-Time Modified Coherence"), es esencialmente un modelado AR de la parte causal de la secuencia de autocorrelación y su uso en reconocimiento del habla con ruido es atractivo por su simplicidad y su alta tasa de reconocimiento con respecto al LPC standard en condiciones severas de ruido blanco aditivo [1] y ruido de automóvil [3].

El propósito de esta comunicación es hacer un estudio comparativo de las técnicas clásicas de parametrización de la señal de voz frente a la nueva técnica OSALPC en el problema de la identificación de locutor en presencia de ruido.

Las pruebas experimentales se han realizado utilizando un sistema de identificación

basado en la distancia Aritmético-Armónica de Esfericidad [4] sobre matrices de covarianza de los parámetros cepstrales.

Esta comunicación esta organizada del siguiente modo: en el apartado 2 se revisa la técnica de parametrización OSALPC, estableciéndose su relación con las técnicas LPC y SMC; en el apartado 3 se proporciona una descripción del sistema de identificación empleado, de la base de datos y de los resultados obtenidos. Por último, en el apartado 4 se recogen las principales conclusiones del trabajo .

2.- Predicción Lineal de la Parte Causal de la Autocorrelación

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

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

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 [5] como

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

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. 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 -llamada OSALPC- equivale a suponer un modelo todo polo para $E^2(\omega)$. Ello da lugar a que esta técnica sólo realize una deconvolución parcial de la señal de voz [1].

La relación de esta técnica con la LPC standard es obvia: la técnica consiste en aplicar el algoritmo LPC standard sobre $R^{+}(n)$. En cuanto a su relación con la técnica SMC [2], las principales diferencias son que esta última utiliza :1) el estimador coherencia para calcular la

secuencia de autocorrelación, mientras que OSALPC utiliza 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 desarrolo teórico de la técnica OSALPC ni en los resultados de reconocimiento [1].

3.- Resultados Experimentales

Este apartado muestra la aplicación de la técnica descrita a la identificación de locutor en presencia de ruido blanco aditivo.

3.1.- Sistema de Identificación y base de datos

Se entrenaron los modelos de los locutores con señal libre de ruido. La señal de voz ruidosa se simuló añadiendo ruido blanco gausiano de media cero a la señal límpia de manera que se obtuviese la SNR deseada.

En la etapa de parametrización del sistema de identificación, la señal se dividió en tramas de 25 ms de duración con un desplazamiento de 10 ms y cada trama se caracterizó con 20 parámetros cepstrales estimados mediante la técnica LPC clásica, mel-cepstrum o OSALPC-cepstrum. En los tres casos se utilizó un orden de análisis de 20.

Posteriormente en la fase de entrenamiento del sistema se calculó la matriz de covarianza de la secuencia de vectores cepstrales formada por la concatenación de las frases de entrenamiento de cada locutor, obteniendose una matriz por cada locutor. Las matrices constituyeron los modelos de cada locutor. Una vez obtenidos todos los modelos, en la fase de test se calculó la matriz de covarianza de la secuencia de vectores cepstrales formada por la concatenación de los vectores de las frases de test del locutor. Se calculó la distancia de la matriz de test a todas las matrices-modelo de los locutores de la base de entrenamiento. La distancia entre matrices utilizada fue la propuesta por Bimbot [4] llamada distancia Aritmético-Armónica de Esfericidad basada en el hecho de que cuanto más semejantes son dos matrices X e Y más cercanos a 1 son los valores propios de $X \cdot Y^{-1}$. La regla de decisión empleada fue la de mínima distancia.

Se empleó la base de datos TIMIT, utilizando subconjuntos de 100 o 200 locutores. Cada locutor consta de 10 frases distintas entre sí y cada frase tiene una duración de aproximadamente 3 s cada una. Para entrenar el modelo de cada locutor se utilizaron las 5 frases etiquetadas en la base de datos como "TI". Las 5 frases restantes se consideraron como el conjunto de frases de test y se obtuvieron 2 señales de test formadas de la concatenación de 2 frases cada una. A las señales de test de les añadió ruido blanco gausiano de media cero de manera que se obtuviese la SNR deseada. Debido a que las frases empleadas en el entrenamiento son distintas de las empleadas en el test, el sistema de identificación de locutor es independiente del texto.

3.2.- Resultados

Los experimentos llevados a cabo consistieron en utilizar los parametros LPC-cepstrum y

mel-cepstrum a distintas SNR, comparandolos con los parametros OSALPC-cepstrum. En las tablas I y II se muestran las tasas de identificación correspondientes a cada parametrización en términos de la SNR de ruido blanco aditivo gausiano para un conjunto de 100 locutores y 200 locutores. En la gráfica I se comparan las tasas de identificación en función de la SNR de la tabla I junto a los resultados obtenidos al utilizar la parametrización LPC cuando se contaminaron las señales de entrenamiento y de test de manera que resultase la misma SNR.

Tabla I : Tasas de Identificación (100 locutores)

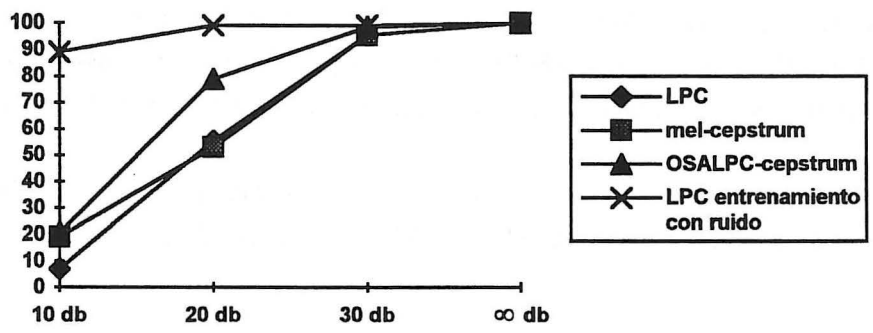
SNR	sin ruido	30 db	20 db	10 db
LPC-cepstrum	100	95.5	55.5	7.0
mel-cepstrum	100	95.5	53.0	19.0
OSALPC	100	98.5	79.0	20.5

Tabla II : Tasas de Identificación (200 locutores)

SNR	sin ruido	30 db	20 db	10 db
LPC-cepstrum	99.75	91.75	43.5	8.25
OSALPC	99.5	97.5	73.0	16.0

En la tabla I se observa que la tasa del sistema de identificación se degrada rápidamente cuando la SNR disminuye. Openshaw y Sun [6] ya habían informado sobre este comportamiento de los sistemas de identificación de locutor. En cuanto a la técnica de parametrización, las tasas de identificación obtenidas usando LPC y mel-cepstrum son muy similares, pero OSALPC resulta ser mucho mejor a menos que la SNR sea demasiado baja.

Debido a que en las pruebas anteriores no hubo ningún error en condiciones libres de ruido, se realizaron pruebas con 200 locutores para poder comparar las técnicas de parametrización en estas condiciones. En la tabla II se puede observar que para condiciones libres de ruido la tasa de identificación de la técnica OSALPC es ligeramente inferior a la tasa de la técnica LPC clásica. Esto es debido a que la técnica de predicción lineal OSALPC sólo realiza una deconvolución parcial de la señal de voz, tal como se explica en el apartado 2.



Gráfica I

Se observa de la gráfica anterior que la tasa de identificación utilizando la técnica LPC, para el caso en que las señales utilizadas en las fases de test y de entrenamiento tengan la misma SNR, disminuye más lentamente al disminuir la SNR que en el caso de utilizar referencias libres de ruido. Por tanto, si conocieramos las condiciones de SNR en que se lleva a cabo el test del sistema de identificación de locutor, podríamos entrenar los modelos del sistema con la misma SNR consiguiendo un sistema de identificación muy robusto al ruido. Sin embargo, en la mayoría de las situaciones prácticas el nivel de ruido en las señales de test es desconocida

4.- Conclusiones

En esta comunicación hemos presentado una técnica de parametrización robusta de la señal de voz aplicada a la identificación del locutor en presencia de ruido, 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-cepstrum y mel-cepstrum, se ha concluido que en identificación de locutor en presencia de ruido, en el caso de ruido blanco aditivo:

-cuando se usa la parametrización LPC-cepstrum o mel-cepstrum la tasa de identificación se degrada rápidamente al disminuir la SNR

-la nueva parametrización OSALPC obtiene mucho mejores resultados que las parametrizaciones clásicas a menos que la SNR sea demasiado baja.

Referencias

[1] J. Hernando, C. Nadeu, E. Lleida, "On the AR Modelling of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition", Proc. ICSLP'92, Banff (Canada), pp. 1593-1596, Octubre 1992.

[2] D. Mansour, B. H. Juang, "The Short-Time Modified Coherence Representation and its Application for Noisy Speech Recognition", IEEE Trans. on ASSP-37, nº 6, pp. 795-804, Junio 1989.

[3] J. Hernando, C. Nadeu, "Speech Recognition in Noisy Car Enviroment Based on OSALPC Representation and Robust Similarity Measuring Techniques", Proc. ICASSP'94, Adelaida (Australia), Abril 1994.

[4] F. Bimbot, L. Mathan, "Text-Free Speaker Recognition Using an Arithmetic-Harmonic Sphericity Measure", Proc. EUROSPEECH'93, Berlin, pp. 169-172, Septiembre 1993.

[5] M. A. Lagunas, M. Armengual, "Non-Linear Spectral Estimation", Proc. ICASP'87, Dallas, pp. 2035-2038.

[6] J. P. Openshaw, Z. P. Sun, J. S. Mason, "A Comparison of Composite Features Under Degraded Speech in Speaker Recognition", Proc. ICASSP'93, pp. 371-374.