RECONOCIMIENTO ROBUSTO DEL HABLA EN PRESENCIA DE RUIDO DE COCHE

J. Hernando, C. Nadeu, J. Dachs y L. Janer
Departamento de Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña
E.T.S.I. Telecomunicación, Apdo. 30002, 08071 Barcelona

ABSTRACT

The performance of existing speech recognition systems degrades rapidly in the presence of background noise when training and testing cannot be done under the same ambient conditions. The aim of this paper is to report the application of several robust techniques on a system based on the HMM (Hidden Markov Models) and VQ (Vector Quatization) approaches for speech recognition in car environment: optimization of spectral model order and cepstral lifter, parameterizatrion based on the AR modeling of the autocorrelation sequence -proposed by the authors in [1]- , use of dynamic information and multilabeling.

1. INTRODUCCION

El reconocimiento automático del habla en ambientes ruidosos es un problema todavía no resulto, incluso en el caso de palabras aisladas y vocabulario reducido. Por este motivo, en los últimos pocos años se han propuesto algunos métodos y algoritmos en varias etapas del proceso de reconocimiento [2], principalmente en las de extracción de características y medida de similitud.

La técnica de predicción lineal (LPC), equivalente a un modelado autorregresivo de la señal, ha mostrado ser de gran utilidad en reconocimiento del habla [3] y, en particular, 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 tanto en condiciones libres de ruido [4] como ruidosas [5]. También se ha de considerar el hecho de que el aumento del orden de predicción lineal puede robustecer la estimación en presencia de ruido.

Como una representación más robusta de la señal en presencia de ruido que la predicción lineal clásica, en [1] los autores presentaron la técnica de parametrización OSALPC, basada en el modelado autorregresivo de la parte causal de la secuencia de autocorrelación de la señal de voz. Esta técnica, estrechamente relacionada cono la representación SMC [6], es interesante en reconocimiento de habla ruidosa debido a su simplicidad, su eficiencia computacional y su alta tasa de reconocimiento en ambientes ruidosos.

Por otro lado, es un hecho bien conocido que las características dinámicas de la señal de voz juegan un papel importante en la percepción humana de la voz y son más robustas a la variabilidad interlocutor y al entorno que las características instantáneas [7]. Es por ello que en la mayoría de los sistemas actuales de reconocimiento se añade información dinámica a la estática para mejorar las prestaciones.

Aplicando técnicas de múltiple etiquetado [8] sobre ambos tipos de información en un sistema basado en la cuantificación vectorial y los modelos ocultos de Markov, los autores [9] han obtenido notables mejoras de la tasas de reconocimiento en ambientes ruidosos respecto a los resultados proporcionados por la cuantificación vectorial clásica.

El propósito de esta comunicación es realizar un estudio comparativo de las técnicas mencionadas en reconocimiento multilocutor de palabras aisladas en presencia de ruido de coche. En el apartado 2 se revisarán los conceptos fundamentales relacionados con estas técnicas (para mayor información consúltese [10]) y en el apartado 3 se expondrán y comentarán los resultados experimentales obtenidos.

2. TECNICAS ROBUSTAS DE RECONOCIMIENTO

2.1. Optimización del orden de predicción y la ventana cepstral

Cuando se utilizan técnicas de predicción lineal en la etapa de parametrización de la señal de voz enpresencia de ruido de banda ancha es preferible el uso de un orden de predicción relativamente alto debido a quelos coeficientes de autocorrelación de índice bajo están más contaminados por el ruido que los de índice alto.

Este trabajo ha sido financiado por el proyecto TIC 92-1026-C02/02

No obstante, si se utiliza un orden de predicción demasiado alto aparecen picos espurios en el espectro provocados por errores de estimación.

En cuanto a las ventanas de ponderación utilizadas en la distancia euclídea entre vectores cepstrales, en este estudio se han considerado las más usuales: rectangular, seno realzado, inversa de la desviación típica y rampa. Debido a que los coeficientes cepstrales de índice bajo están más contaminados por el ruido de banda ancha que los de índice alto, de forma análoga a lo que ocurre en el caso de los coeficientes de autocorrelación, se espera que se obtendrán mejores resultados con ventanas crecientes, como la inversa de la desviación típica o la rampa, ya que desenfatizan los coeficientes cepstrales de índice bajo.

Como resultado de la ponderación cepstral del modelo LPC, 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.

2.2. Predicción lineal de la parte causal de la autocorrelación

A partir de la secuencia de autocorrelación R(n), se define 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

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

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 [11] 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 [1].

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 polos para $E^2(\omega)$. Ello da lugar a que esta última técnica sólo realice una desconvolución parcial de la señal de voz, lo cual da lugar a estimaciones espectrales pobres de la señal de voz en ausencia de ruido.

2.3. Incorporación de información dinámica

Debido a que la diferencia finita de primer orden es intrínsecamente ruidosa, la estimación de las características dinámicas de la señal de voz suele realizarse aplicando análisis de regresión sobre la evolución temporal de cada coeficiente cepstral (delta-cepstrum) o de la energía (delta-energía) en un intervalo de duración adecuado.

En cuanto a la incorporación de estas informaciones a un sistema de reconocimiento basado en los modelos ocultos de Markov con cuantificación vectorial, se suelen utilizar las dos siguientes alternativas:

- Distancia compuesta, que consiste en construir un supervector concatenando con una ponderación adecuada las informaciones que se desean utilizar y obtener un único símbolo aplicando distancia euclídea en el proceso de cuantificación vectorial.

- Diccionarios múltiples, en la que se cuantifican por separado cada una de las informaciones y se considera independencia estadística de las mismas en el entorno de los modelos ocultos de Markov.

2.4. Múltiple etiquetado

En la cuantificación vectorial clásica, utilizada por los modelos ocultos de Markov discretos, se elige únicamente la palabra-código del diccionario que dista menos al vector a cuantificar y se descarta la información sobre el grado en que dicho vector se ajusta a otras palabras-código. Esta información puede ser importante en el caso de habla ruidosa, ya que la decisión tomada por el cuantificador del modelo discreto puede ser fácilmente modificada por el ruido añadido a la señal.

Sin embargo, en los modelos semicontinuos y de múltiple etiquetado el cuantificador vectorial proporciona información sobre el grado de ajuste a las palabras-código más cercanas mediante unas funciones de peso, lo cual disminuye en gran medida el error de cuantificación y robustece el sistema frente al ruido aditivo. La única diferencia entre los modelos semicontinuos y los de múltiple etiquetado es que los primeros utilizan como funciones de peso funciones de densidad de probabilidad, mientras que los segundos se basan exclusivamente en el concepto de distancia. Ambos tipos de modelos proporcionan prestaciones similares y notablemente superiores a los de los modelos discretos. Sin embargo, los modelos de múltiple etiquetado son mucho más eficientes desde el punto de vista computacional, de ahí que sean los utilizados en los experimentos siguientes.

3. RESULTADOS EXPERIMENTALES

3.1. Base de datos y pruebas de reconocimiento

La base de datos empleada procede del proyecto ESPRIT-ARS y está formada por 4 locutores, 2 hombres y 2 mujeres, que pronuncian 25 veces los dígitos italianos en el interior de un coche (Fiar Tipo) y en diferentes condiciones de ruido: 5 repeticiones con el motor y el ventilador parados y 20 más con el motor en marcha y diferentes potencias del ventilador, de las cuales 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 se entrenó con las señales pronunciadas con el motor y el ventilador parados, es decir, en condiciones virtualmente libres de ruido, y en la fase de reconocimiento se utilizaron el resto de las señales. En ambas etapas se hizo uso de detección manual de voz.

En la etapa de parametrización, la señal de voz, una vez muestreada a 8 kHz, cuantificada con 12 bits por muestra y preenfatizada, se dividió en tramas de 30 ms. de duración con un desplazamiento de 15 ms. y cada trama se caracterizó por sus parámetros cepstrales estimados bien mediante la técnica clásica LPC, bien mediante la nueva técnica OSALPC, y en su caso la energía y los parámetros regresivos correspondientes. Posteriormente, se utilizaron diccionarios de 64 palabras-código, entrenados mediante el algoritmo de Lloyd, para cuantificar por separado cada una de las informaciones. La elección del tamaño del diccionario y de la estrategia de múltiples diccionarios se realizó en base a resultados de reconocimiento efectuados en pruebas preliminares.

Cada palabra se caracterizó por un modelo de Markov 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 llevó a considerar modelos de izquierda a derecha de 10 estados sin posibilidad de transición entre estados no consecutivos.

3.2. Resultados de reconocimiento

Los primeros experimentos llevados a cabo consistieron en optimizar empíricamente el orden de predicción y el tipo de ponderación cepstral utilizando sólo información espectral instantánea y cuantificación vectorial clásica. Los mejores resultados se obtuvieron con orden de predicción igual a 16 y ventana cepstral inversa de la desviación típica en el caso de parametrización LPC y rampa en el caso de la técnica OSALPC. En la figura 1 se comparan las tasas de reconocimiento obtenidas en función de la velocidad del coche utilizando el orden de predicción y la ventana cepstral más usuales (orden 8 y ventana seno realzado) y los óptimos.

En esta figura se observa una gran sensibilidad de los resultados a ambos factores, siendo aconsejables órdenes de predicción relativamente altos y ventanas cepstrales crecientes. También se observa que, utilizando el orden de predicción y la ventana cepstral óptimos, la parametrización OSALPC supera claramente a la técnica LPC en condiciones severas de ruido, pero sus prestaciones son algo peores que las de dicha técnica en condiciones poco ruidosas.

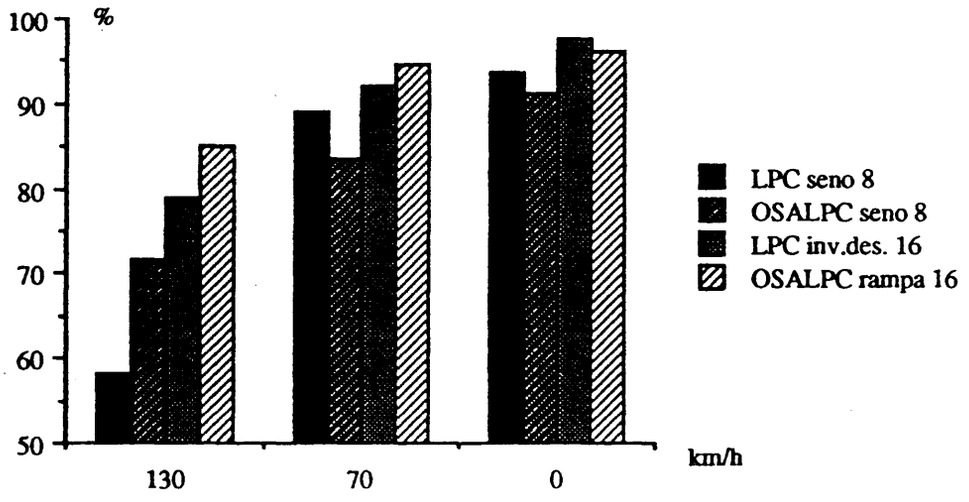
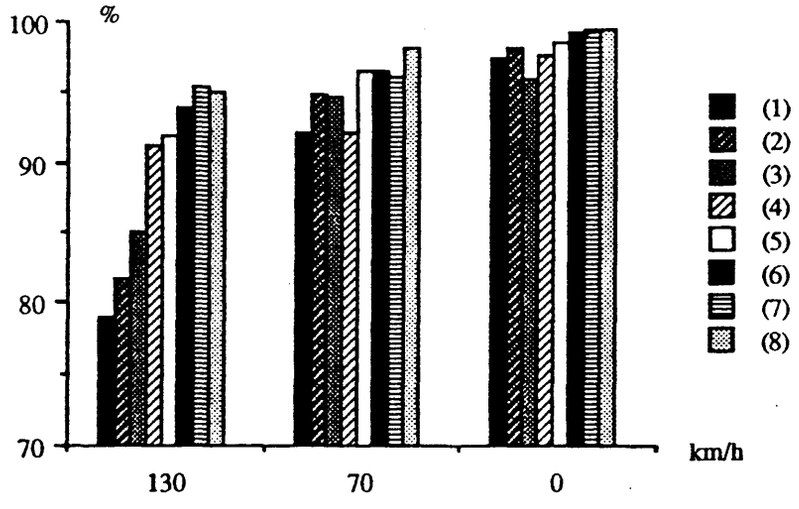


Figura 1. Influencia del orden de predicción y la ventana cepstral en LPC y OSALPC

En cuanto a la utilización de energía y parámetros regresivos, para el caso de la parametrización LPC se obtiene una notable mejora de resultados incorporando la información de delta-cepstrum y delta-energía y en el caso de la parametrización OSALPC utilizando la información de delta-cepstrum. Las mejores resultados se obtienen para una una longitud del intervalo de estimación de los parámetros regresivos de 240 ms.

También se logra mejorar ostensiblemente los resultados sustituyendo la cuantificación vectorial clásica por el múltiple etiquetado. El compromiso coste computacional-tasa de reconocimiento ha llevado a considerar únicamente la información correspondiente a las 5 palabras-código más cercanas al vector a cuantificar.

La combinación de estas técnicas proporciona mejoras todavía mayores a las obtenidas mediante su aplicación de forma aislada. En la figura 2 se comparan las tasas de reconocimiento obtenidas, utilizando el orden de predicción y la ventana cepstral óptimos, en función de la parametrización -LPC u OSALPC- y la cuantificación vectorial -clásica o múltiple etiquetado- empleadas y la utilización o no de parámetros regresivos. El orden en que se han etiquetado las diferentes combinaciones de técnicas ha sido el de prestaciones crecientes en condiciones severas de ruido.



(1) Parametrización LPC, cuantificación vectorial clásica, sólo cepstrum
(2) Parametrización LPC, etiquetado múltiple, sólo cepstrum
(3) Parametrización OSALPC, cuantificación vectorial clásica, sólo cepstrum

(4) Parametrización OSALPC, múltiple etiquetado, sólo cepstrum
(5) Parametrización LPC, cuantificación vectorial clásica, varias informaciones
(6) Parametrización LPC, múltiple etiquetado, varias informaciones
(7) Parametrización OSALPC, cuantificación vectorial clásica, varias informaciones
(8) Parametrización OSALPC, múltiple etiquetado, varias informaciones

Figura 2. Comparación y combinación de técnicas

Puede observarse en esta figura que la técnica OSALPC sin utilizar delta-cepstrum obtiene resultados excelentes en condiciones severas de ruido, pero sus prestaciones son algo peores que los de la técnica clásica LPC en condiciones poco ruidosas. Sin embargo, utilizando delta-cepstrum la parametrización OSALPC es superior a la LPC en todas las condiciones consideradas. Por otro lado, se observa que el múltiple etiquetado proporciona excelentes resultados combinado con la utilización de parámetros regresivos. Los mejores resultados se obtienen utilizando parametrización OSALPC, múltiple etiquetado y delta-cepstrum.

4. CONCLUSIONES

En esta comunicación se han revisado diversas técnicas de reconocimiento robusto del habla en ambientes ruidosos relacionadas con las etapas de parametrización de la señal de voz y comparación de vectores de características. A partir de un estudio comparativo de estas técnicas en una aplicación monolocutor de palabras aisladas en el caso de ruido de coche, utilizando un sistema de reconocimiento basado en la cuantificación vectorial y los modelos ocultos de Markov, se han extraído las siguientes conclusiones:

a) Cuando se utilizan técnicas de predicción lineal en la etapa de parametrización es preferible el uso de un orden de predicción relativamente alto.

b) Se obtiene una importante mejora de resultados utilizando ventanas cepstrales crecientes.

c) La representación cepstral basada en la técnica de predicción lineal de la parte causal de la secuencia de autocorrelación (OSALPC), propuesta por los autores [1], alcanza excelentes resultados en condiciones severas de ruido; en condiciones poco ruidosas, se produce un ligero empeoramiento con respecto a la parametrización clásica LPC que puede ser corregido utilizando esta técnica combinada con parámetros regresivos.

d) Resulta de gran utilidad el uso de parámetros dinámicos de las características de la señal de voz para el reconocimiento de habla ruidosa.

e) El múltiple etiquetado proporciona una clara mejora de resultados en reconocimiento de habla ruidosa con respecto a la cuantificación vectorial clásica.

f) La combinación de estas técnicas proporciona mejoras todavía mayores a las obtenidas mediante su aplicación de forma aislada.

5. REFERENCIAS

[1] J. Hernando, C. Nadeu y D. Riu, "Técnicas de modelado AR robusto de la señal de voz para el reconocimiento del habla en ambientes ruidosos", Proc. URSI'92, Málaga, pp. 134-138.

[2] B.H. Juang, "Speech recognition in adverse conditions", Computer Speech and Language, 1991, vol. 5, pp. 275-94.

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

[4] 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. 947-54.

[5] 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. 968-73.

[6] 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.

[7] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum". IEEE Trans. ASSP, vol. 34, n° 1, pp. 52-59, 1986.

[8] M. Nishimura y K. Toshioka, "HMM-based speech recognition using multi-dimensional multi-labeling", Proc. ICASSP-87, Dallas, pp. 1163-66.

[9] J. Hernando, J.B. Mariño y C.Nadeu, "Multiple multilabeling to improve HMM-based speech recognition in noise", EUROSPEECH'93, Berlín.

[10] J. Hernando, "Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos", Tesis Docotoral, Dpto. Teoría de la Señal y Comunicaciones (UPC), Mayo 1993.

[11] M.A. Lagunas y M. Amengual, "Non-linear spectral estimation", Proc. ICASSP-87, Dallas, pp. 2035-38.