UNIVERSIDAD POLITECNICA DE CATALUÑA

Departamento de Teoria de la señal y comunicaciones

TECNICAS DE PROCESADO Y REPRESENTACION DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS

Autor: Francisco Javier Hernando Pericas
Director: Climent Nadeu i Camprubi

Barcelona, mayo 1993

Capítulo 7

CONCLUSIONES

En esta tesis se han presentado y 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 un aplicación monolocutor de palabras aisladas, utilizando un sistema de reconocimiento basado en la cuantificación vectorial y los modelos ocultos de Markov, se han extraído las siguientes conclusiones fundamentales para el caso de ruido blanco y ruido real de coche [Her93b]:

- En el caso de que la señal de voz esté perturbada por la presencia de ruido blanco, el efecto del preénfasis no es deseable. El ruido blanco es espectralmente plano, mientras que la señal de voz tiene concentrada su energía en las bajas frecuencias. En consecuencia, si se aplica un filtro paso-alto resulta que se realza la zona del espectro en que el ruido presenta mayor potencia relativa respecto a la señal.

- Cuando se utilizan técnicas de predicción lineal en la etapa de parametrización de la señal de voz, es preferible el uso de un orden de predicción relativamente alto. Así, por ejemplo, utilizando una frecuencia de muestreo de 8 kHz el orden de predicción usual es 8 y se han obtenido importantes mejoras de los resultado utilizando ordenes 12 (en las pruebas con ruido blanco) y 16 (en las pruebas con ruido real de

coche). La conveniencia de este orden relativamente alto es debido al hecho de que los coeficientes de autocorrelación de orden bajo están más contaminados que los coeficientes de orden alto, en el caso de los tipos de ruido considerados.

- Se obtiene una importante mejora de resultados utilizando ventanas cepstrales crecientes, como la ventana rampa o la inversa de la desviación típica de cada coeficiente cepstral. Con ello se consigue desenfatizar los coeficientes cepstrales de orden inferior, que son los más contaminados por estos tipos de ruido.

- La representación cepstral basada en la técnica de predicción lineal de la parte causal de la autocorrelación (OSALPC), propuesta en esta tesis, alcanza excelentes resultados en condiciones severas de ruido y es menos sensible a los factores anteriores que la predicción lineal clásica. Esta técnica se ha derivado a partir de la interpretación de la predicción lineal clásica y del uso de un sistema sobredeterminado de ecuaciones de Yule-Walker como técnicas de predicción lineal en el dominio de la autocorrelación, en lugar de sobre la señal misma. Su uso en reconocimiento de habla ruidosa es muy interesante debido a su simplicidad, su eficiencia computacional y sus altas tasas de acierto en condiciones severas de ruido, que superan ampliamente a las técnicas mencionadas y a la representación SMC, con la cual está estrechamente relacionada. En ausencia de ruido, sin embargo, se produce un ligero empeoramiento de las prestaciones con respecto a la predicción lineal clásica debido a que no realiza una deconvolución completa de la señal de voz.

- Utilizando técnicas de predicción lineal y en ausencia de ponderación cepstral, 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 usuales sobre el cepstrum derivado del modelo de predicción lineal la transformación bilineal no ayuda al reconocimiento de habla ruidosa.

- Las representaciones instantáneas de la señal son menos robustas que las dinámicas frente al ruido, por lo cual resulta de gran utilidad el uso de parámetros dinámicos del espectro y de la energía. No sólo han proporcionado buenos resultados los parámetros dinámicos de primer orden, calculados como el coeficiente de regresión lineal de los parámetros instantáneos en un intervalo finito, sino también los parámetros dinámicos de orden superior n, estimados como el coeficiente de regresión de orden n-ésimo o como el coeficiente de regresión lineal del parámetro dinámico de

orden n-1. Aunque la mejora proporcionada por la adición de un nuevo parámetro de orden superior disminuye con el orden n, en las prueba experimentales se han conseguido mejoras apreciables utilizando hasta parámetros dinámicos de tercer orden. Se ha observado, además, que el intervalo de estimación de los parámetros dinámicos se ha de ajustar en cada aplicación, ya que depende de las condiciones de ruido, el algoritmo de estimación y el orden del parámetro. En cuanto a la forma de incorporar estos parámetros a un sistema de reconocimiento basado en la cuantificación vectorial y los modelos ocultos de Markov, la estrategia que ha proporcionado mejores resultados es la utilización de cuantificadores independientes para cada información, en lugar de distancia compuesta.

- La distancia de proyección supera a la distancia euclídea en condiciones muy severas de ruido, mientras que en ausencia de ruido y relaciones señal-ruido moderadas el comportamiento de las dos distancias es muy similar.

- En condiciones severas de ruido es conveniente utilizar un cuantificador vectorial con un diccionario de pocas palabras-código, lo cual puede justificarse si se tiene en cuenta que en este caso se producirán menores errores de cuantificación debidos al ruido cuanto mayor sea la región del espacio de características asignada a cada palabra-código.

- Los modelos semicontinuos y de múltiple etiquetado superan notablemente en prestaciones a los modelos discretos en condiciones ruidosas y en ausencia de ruido debido a una disminución considerable de los errores de cuantificación. En el proceso de cuantificación vectorial de los modelos ocultos de Markov discretos se elige únicamente la palabra-código que dista menos del 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 especialmente 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 la distancia relativa a las palabras-código más cercanas y, por tanto, se conserva parte de esta información. Los resultados de reconocimiento correspondientes a los modelos semicontinuos y a los de múltiple etiquetado son muy similares. Sin embargo, estos últimos son mucho más eficientes desde el punto de vista computacional.

La utilización combinada de varias de estas técnicas puede proporcionar excelentes resultados en reconocimiento del habla en ambientes ruidosos. La

interrelación existente entre estas técnicas forma parte de un trabajo futuro de investigación. Sin embargo, a partir de los resultados experimentales de esta tesis, se puede observar claramente que el múltiple etiquetado hace superflua la sustitución de la distancia de proyección por la euclídea. También se concluye que la utilización de parámetros dinámicos es siempre beneficiosa, exceptuando cuando se usa la parametrización OSALPC en ausencia de ruido. Este último resultado es debido a que la estimación espectral proporcionada por esta parametrización no es suficientemente precisa y, por tanto, la acentuación de sus cambios temporales puede no ser conveniente.