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 1

INTRODUCCION

1.1. OBJETIVOS GENERALES

En el planteamiento general del desarrollo de los ordenadores de las próximas generaciones se prevé la comunicación hombre-máquina mediante mensajes orales, esto es, un dispositivo privilegiado de entrada del mundo exterior al ordenador será un sistema de reconocimiento del habla. Sin embargo, con las técnicas actuales, sólo en el caso de palabras aisladas, cuando el vocabulario es reducido y en situaciones acústico-fonéticas (variabilidad fonética, tipo de locutores, ruido y distorsiones,...) poco dificultosas existen soluciones satisfactorias. Una mejora en los métodos de extracción de características de la señal de voz es necesaria para avanzar en este problema.

El comportamiento de los sistemas actuales de reconocimiento automático del habla se degrada rápidamente en presencia de ruido de fondo cuando las etapas de entrenamiento y de test no se llevan a cabo en las mismas condiciones ambientales.

El problema requiere soluciones más elaboradas que el simple uso de patrones de referencia entrenados en las mismas condiciones ambientales en que se va a llevar a cabo el reconocimiento, pues en la mayoría de las aplicaciones esta estrategia no es factible.

En los últimos años se han propuesto una gran variedad de métodos y algoritmos en la dirección de desarrollar sistemas de reconocimiento robustos frente al ruido.

Fundamentalmente, estas técnicas se basan en el perfeccionamiento de los sensores, la aplicación de métodos de mejora de la señal de voz para suprimir el ruido antes de aplicar el algoritmo de reconocimiento, el uso de nuevas representaciones de la señal de voz, la aplicación de medidas de distancia robustas entre espectros de voz, la utilización de modelos estocásticos y cuantificadores vectoriales adaptativos, etc. Sin embargo, el reconocimiento de habla ruidosa no ha encontrado todavía una solución satisfactoria incluso en el caso de palabras aisladas y vocabularios reducidos.

El objetivo fundamental de esta tesis es el estudio y la aplicación al reconocimiento automático del habla en entornos ruidosos de representaciones alternativas de la señal de voz que sean robustas por sí mismas al ruido y computacionalmente eficientes. La principal motivación que ha conducido a esta aproximación al problema del reconocimiento robusto del habla es la conocida sensibilidad al ruido aditivo que presenta la técnica de predicción lineal, ampliamente usada en reconocimiento y procesado de habla en general.

En la búsqueda de nuevas representaciones de la señal de voz se pueden distinguir a su vez dos enfoques: 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. Ambos enfoques son considerados en este trabajo.

Respecto al primer enfoque, pueden encontrarse en la literatura gran cantidad de trabajos de estimación espectral robusta AR y ARMA relacionados con el problema del ruido. La inmensa mayoría de estos métodos conducen a algoritmos iterativos, con los consiguientes problemas de convergencia y coste de cálculo, o exigen un estimación previa del ruido. En su lugar, en esta tesis se estudian modificaciones simples y eficientes de la técnica de predicción lineal clásica basadas en el modelado autorregresivo en el dominio de la autocorrelación.

En cuanto al segundo enfoque, se han hecho intentos importantes en los últimos años de representar el patrón de excitación del nervio auditivo mediante modelos computacionales. Sin embargo, tales modelos son demasiado costosos desde el punto de vista de cálculo y no todas las funciones que incorporan son significativas para combatir el ruido. Por ello, en esta tesis se ha optado por realizar la estimación básica del espectro de la señal de voz según el enfoque anterior e incorporar sobre esta

estimación básica aquellas evidencias auditivas que puedan ayudar de una forma eficiente a la representación robusta de la señal de voz.

De forma natural, la estrecha relación existente entre el tipo de representación de la señal y la medida de distancia idónea para confrontar los vectores de parámetros ha conducido también al estudio de medidas de distancias en esta tesis.

Ultimamente se ha puesto de manifiesto que la utilizazción de la distancia euclídea sobre los coeficientes cepstrales del modelo de predicción lineal, convenientemente ponderados con una ventana, ofrece en general mejores prestaciones que cualquier otro tipo de distancia asociada al modelo. Como resultado de la ponderación cepstral se obtiene una versión suavizada del espectro que depende tanto de la forma y longitud de la ventana de ponderación como del orden del modelo de estimación espectral. Uno de los objetivos de este trabajo será, por tanto, obtener el grado óptimo de suavizado en condiciones ruidosas para el caso de la predicción lineal clásica y las parametrizaciones alternativas propuestas.

Además, se considera en este trabajo el estudio de distancias alternativas a la distancia euclídea cepstral, como por ejemplo, las distancias de proyección cepstral, utilizando distintos tipos de ponderación y órdenes de del modelo de estimación espectral.

También relacionado con el tema de la parametrizaciones y las distancias robustas de la señal de voz, se ha observado que las condiciones adversas afectan más a las representaciones espectrales instantáneas de la señal de voz que a las representaciones dinámicas. Por ello, paralelamente al desarrollo de nuevas representaciones espectrales, otro objetivo de este trabajo es profundizar en la implementación de los parámetros regresivos: algoritmo e intervalo de estimación, número óptimo de parámetros,... y se ha considerado la posibilidad de una generalización de esta técnica hacia un filtrado cepstral. Por último, también se ha estudiado la incorporación de la información de la energía de la señal de voz para robustecer el sistema de reconocimiento, tanto mediante representaciones instantáneas como dinámicas.

Para la evaluación de las técnicas propuestas se ha utilizado un sistema de reconocimiento de palabras aisladas mediante modelos ocultos de Markov. El hecho de que el sistema sea de palabras aisladas permite prescindir de las implicaciones de los niveles de conocimiento superiores al acústico: sintáctico, semántico, pragmático,...

Por otro lado, los modelos ocultos de Markov son los que en estos momentos proporcionan unas mejores prestaciones en todos los sistemas en desarrollo.

La cuantificación vectorial inherente a distintos tipos de modelos ocultos de Markov, discretos, semicontinuos y de múltiple etiquetado, puede tener repercusiones importantes en las tasas de reconocimiento en presencia de ruido. Este problema también ha sido objeto de estudio en esta tesis. Para ello, se han comparado las prestaciones de los diferentes tipos de modelos en reconocimiento de habla ruidosa, variando los parámetros del cuantificador vectorial: tamaño del diccionario, número de palabras-código a considerar, funciones de ponderación, etc.

1.2. ESTRUCTURA DE LA TESIS

Tras este primer capítulo de introducción, el capítulo 2 de esta tesis está dedicado al estado del arte en el reconocimiento automático del habla en entornos adversos. En primer lugar, se comentan las ventajas de la comunicación oral hombre-máquina y las dificultades que conlleva el reconocimiento automático del habla en general derivadas de su carácter multiinteractivo, variable, continuo y redundante, que obligan a imponer restricciones al problema global del reconocimiento en cuanto al tipo de habla, la talla del léxico, la gramática del lenguaje, el número de locutores y las condiciones ambientales. También se describen las principales técnicas que se aplican al reconocimiento automático del habla: comparación de patrones, modelos ocultos de Markov, redes neuronales y métodos basados en el conocimiento. En segundo lugar, se aborda el problema concreto del reconocimiento del habla en entornos adversos. Después de unas ideas sobre la naturaleza y dimensión del problema, se revisan los principales fenómenos físicos que provocan entornos adversos para el reconocimiento automático del habla y las principales técnicas de reconocimiento robusto que se han propuesto en la literatura: transductores especiales, nuevas representaciones de la señal, preprocesado de mejora de la voz, medidas de distorsión robustas, enmascaramiento, modelos adaptativos, etc.

El capítulo 3 está dedicado a las técnicas de representación de la señal de voz. Para empezar, se revisan los modelos digitales de producción de la señal de voz, basados en los principios fisiológicos y en las características temporales y frecuenciales de la misma y se resumen los principales características de la predicción lineal clásica. Seguidamente, se aborda el tema de la sensibilidad al ruido de las técnicas de predicción

lineal clásicas y las principales variaciones que se han propuesto para combatir el problema. A continuación, se expone una nueva interpretación de las técnicas anteriores desde el punto de vista de la señal de autocorrelación, que dará pie a la introducción de la técnica de predicción lineal de la parte causal de la autocorrelación como parametrización robusta de la señal de voz en presencia de ruido. Finalmente, se describe una transformación eficiente de la escala de frecuencias para emular la sensibilidad logarítmica en frecuencia del oído.

En el capítulo 4 se describen las medidas de distorsión utilizadas en reconocimiento robusto del habla. Respecto a la distancia euclídea cepstral, se aborda el tema de la ponderación óptima de los coeficientes cepstrales en presencia de ruido. También se considera el uso de distancias alternativas, como la distancia de proyección cepstral. Finalmente, se incluye en este capítulo el tema de la utilización de la energía y las características dinámicas de la señal de voz.

El capítulo 5 está dedicado a la revisión de la teoría correspondiente a los modelos ocultos de Markov. En primer lugar, se presenta la estructura común a todos los tipos de modelos. Seguidamente, se revisan los algoritmos para la evaluación, decodificación y entrenamiento de la aproximación básica discreta, se tratan aspectos prácticos de su implementación y se estudia su aplicación al reconocimiento automático del habla. Por último, se extienden estos resultados a las aproximaciones continua, semicontinua y con múltiple etiquetado, estudiando las posibles ventajas de estas en reconocimiento de habla en entornos adversos.

Los resultados experimentales se recogen en el capítulo 6. Mediante pruebas de reconocimiento de palabras aisladas en presencia de ruido blanco y ruido real de coche se muestran las prestaciones de las técnicas de parametrización y comparación propuestas en esta memoria y se comparan con las correspondientes a las técnicas clásicas. Por último, en el capítulo 7 se resumen las conclusiones más importantes que se han derivado de este trabajo.