Ascensión Gallardo Antolín obtuvo el título de Ingeniera de Telecomunicación por la Universidad Politécnica de Madrid en 1993, y de Doctora Ingeniera de Telecomunicación por la misma universidad en 2002. Desde 1994 hasta 1997 estuvo vinculada al Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica (ETSIT) de la Universidad Politécnica de Madrid hasta su incorporación al Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid (1997-2003). En la actualidad es profesora Ayudante Doctor en la Escuela Politécnica Superior de la Universidad Autónoma de Madrid.

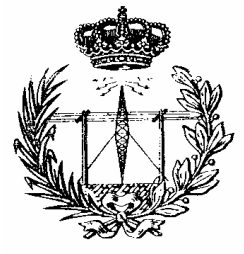
Sus intereses de investigación se centran en el reconocimiento automático del habla en ambientes ruidosos y en redes de comunicaciones (GSM, UMTS, IP), los sistemas de diálogo y la extracción automática de contenidos multimedia. En dichas áreas, ha participado en varios proyectos y contratos de investigación, tanto de ámbito nacional como europeo.

Es coautora de diversas comunicaciones en congresos nacionales e internacionales y artículos en revistas internacionales relacionados con las líneas de investigación anteriores. También es coautora de un capítulo del libro “Signal Processing for Mobile Communications Handbook” editado por CRC Press de próxima aparición.

UNIVERSIDAD POLITÉCNICA DE MADRID

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN



TESIS DOCTORAL

RECONOCIMIENTO DE HABLA ROBUSTO FRENTE A CONDICIONES DE RUIDO ADITIVO Y CONVOLUTIVO

ASCENSIÓN GALLARDO ANTOLÍN
Ingeniera de Telecomunicación

Director de la Tesis
JOSÉ MANUEL PARDO MUÑOZ
Doctor Ingeniero de Telecomunicación

Madrid, 2002

Datos personales de la autora de la Tesis Doctoral

Autora:

ASCENSIÓN GALLARDO ANTOLÍN

Domicilio:

Teléfono:

Dirección de correo electrónico:

Número de asociado:

A05881

Datos de la Tesis Doctoral

Título:

RECONOCIMIENTO DE HABLA ROBUSTO FRENTE A CONDICIONES DE RUIDO ADITIVO Y CONVOLUTIVO

Director:

Dr. JOSÉ MANUEL PARDO MUÑOZ

Departamento:

Departamento de Ingeniería Electrónica de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid.

Fecha de lectura:

29 de Octubre de 2002

Calificación:

Sobresaliente Cum Laude por unanimidad del tribunal

Resumen de la tesis doctoral

1. Contexto de la tesis doctoral

La siguiente tesis está enmarcada en el campo de las tecnologías del habla, concretamente en el de reconocimiento automático del habla (RAH), cuyo objetivo fundamental es la conversión del habla humana en texto por parte de máquinas.

La importancia de la comunicación oral es evidente en numerosos aspectos de la vida de los seres humanos. En todas las sociedades, incluso en las más primitivas, la comunicación oral existe y está basada en mecanismos acústico-sintáctico-semánticos complejos, independientes del nivel tecnológico alcanzado por la sociedad creadora del lenguaje en cuestión.

En las últimas décadas se han realizado grandes esfuerzos en el área del reconocimiento automático del habla. De hecho, una gran cantidad de laboratorios en el ámbito nacional e internacional han concentrado sus esfuerzos en la consecución de sistemas cada vez más complejos y eficientes, que no sólo hacen uso de las características acústicas de la voz, sino también de las particularidades sintácticas e incluso semánticas de cada lengua. Dichos sistemas han alcanzado tasas de reconocimiento aceptables en condiciones "ideales" o de laboratorio. Dichas condiciones se refieren a voz grabada en entornos acústicos limpios (por ejemplo, en cámaras sordas), tanto para la parte de la base de datos dedicada para entrenamiento como para la dedicada a evaluación (mismo micrófono, mismo locutor, etc.). Sin embargo, cuando las condiciones de entrenamiento y evaluación son distintas (lo que es habitual en las aplicaciones reales), los sistemas de reconocimiento automático funcionan considerablemente peor.

A medida que las tecnologías del habla se han ido implicando cada vez más como parte integral de aplicaciones prácticas en escenarios reales (como acceso a bases de datos por línea telefónica, marcado automático de números telefónicos, control de instrumental en coches, máquinas de dictado, etc.), se ha hecho patente la necesidad de desarrollar sistemas automáticos de reconocimiento de habla robustos, es decir que mantengan sus prestaciones dentro de un amplio margen de condiciones de entorno, incluso en el caso en que dichas condiciones varíen de forma rápida.

En general, denominamos condiciones adversas a todas aquellas que degradan el funcionamiento del sistema de reconocimiento de habla. En el esquema de la Figura 1 se resumen los tipos de condiciones adversas usualmente presentes en los escenarios reales de aplicación de los sistemas de reconocimiento automático del habla. Pueden clasificarse en dos grandes grupos: fuentes de distorsión externas y fuentes de distorsión internas o debidas al locutor. Entre las primeras, destacan la aparición de ruido de fondo (también llamado ruido aditivo, puesto que se suele modelar como una suma a la señal de voz de entrada en el dominio del tiempo), la distorsión producida por las características frecuenciales del canal de transmisión, tales como micrófonos o canal telefónico (también llamada ruido convolutivo puesto que los canales de transmisión suelen modelarse como un sistema lineal e invariante), los errores de transmisión

propios de los sistemas de RAH cuando funcionan en entornos de comunicaciones móviles como GSM y UMTS y la pérdida de tramas o paquetes de voz, propios de las aplicaciones de reconocimiento en redes sin cable y VoIP. Las fuentes de distorsión internas son debidas a las diferencias entre las elocuciones pronunciadas por distintos locutores (variabilidad interlocutor) y las diferencias de pronunciación en un mismo locutor debido al estilo de habla utilizado (lectura, habla espontánea), velocidad de elocución y su estado de ánimo (variabilidad intralocutor).

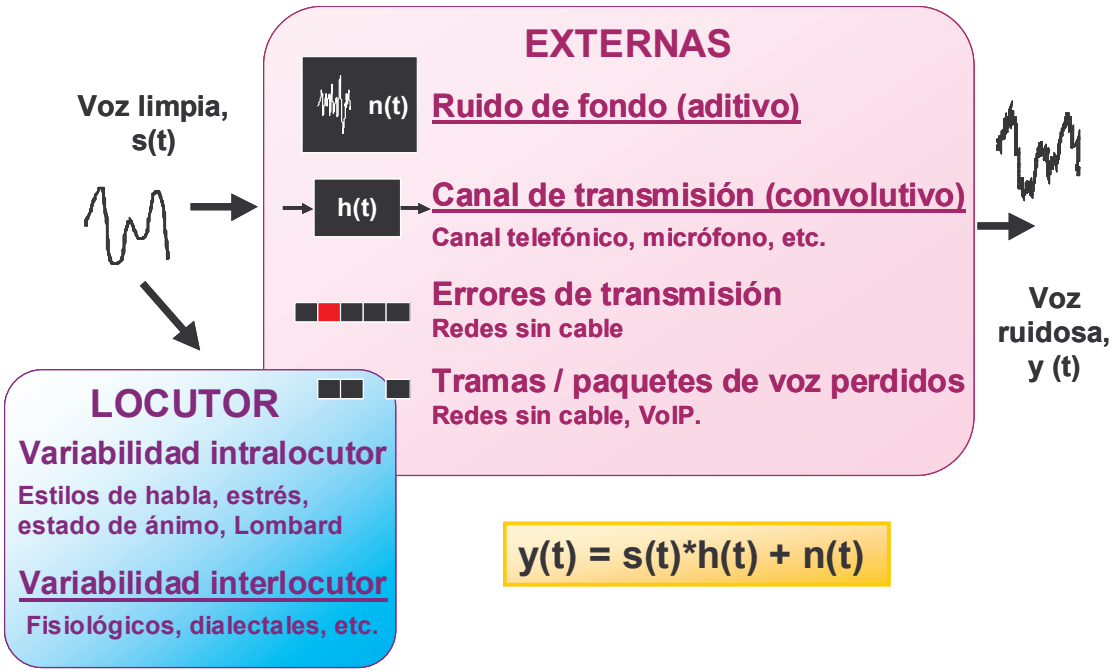


Figura 1. *Esquema de las condiciones acústicas que degradan el funcionamiento de los sistemas de reconocimiento automático del habla.*

2. Objetivos de la tesis doctoral

El objetivo de la tesis doctoral ha sido profundizar en el estudio de métodos de reconocimiento robusto en las condiciones anteriormente mencionadas, haciendo especial énfasis en la variabilidad interlocutor, canal telefónico y ruido aditivo. En concreto, se abordó el problema desde tres perspectivas distintas (pero no excluyentes): en primer lugar, se trabajó en la integración de modelado acústico múltiple en un sistema de reconocimiento; en segundo lugar, en el estudio de representaciones más robustas de la señal de voz; y en tercer lugar, en el estudio de la transformación de parámetros y de modelos acústicos de acuerdo con las condiciones del entorno acústico.

El sistema de reconocimiento sobre el que realizaremos este estudio está basado en modelos ocultos de Markov y ha sido desarrollado por el Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica. En la actualidad, dicho sistema funciona en aplicaciones en tiempo real (ver referencias [21], [22]y [23]). Este es el motivo por el cual, a lo largo de toda la tesis, se ha optado por trabajar con técnicas que no requieren de un incremento prohibitivo de la carga computacional y de memoria del sistema, de modo que pudieran ser incorporadas al sistema final para permitir su funcionamiento en aplicaciones reales.

3. Robustez frente a la variabilidad interlocutor

Una de las causas más comunes que provoca la aparición de errores de reconocimiento en las aplicaciones reales basadas en sistemas de RAH independientes del locutor es la distorsión interlocutor, es decir, las variaciones de pronunciación entre locutores diferentes. Son debidas a las características propias de cada locutor, que dependen, entre otros factores, de la longitud de su tracto vocal y de sus cuerdas vocales.

En este tesis hemos abordado este problema utilizando las denominadas técnicas de multimodelado, que consisten en la división del universo de locutores posibles en diversos grupos con características homogéneas, de modo que el sistema pueda generar y utilizar modelos (acústicos, sintácticos, semánticos, etc) específicos para cada uno de dichos grupos. En esta tesis, hemos investigado en la aplicación de esta estrategia a un sistema de reconocimiento con arquitectura no integrada (multimodular), en concreto, al módulo de generación de hipótesis de un sistema de RAH de palabras aisladas en entorno telefónico, independiente del locutor y de gran vocabulario (5000 y 10000 palabras), en el que se ha hecho especial énfasis en dos aspectos fundamentales: mejorar la tasa de inclusión y no incrementar de forma desproporcionada los requerimientos del sistema en cuanto a su carga computacional y memoria.

Debido justamente a la arquitectura modular del sistema (formado básicamente por dos submódulos: generación de cadenas fonéticas usando modelos ocultos de Markov seguido por un proceso de acceso léxico guiado por el diccionario de la tarea en cuestión), se presentan diversas alternativas para la incorporación tanto de modelado acústico múltiple (en el primer submódulo) como de información léxica múltiple (en el segundo submódulo).

De los diversos experimentos realizados, hemos podido concluir que la incorporación de modelado acústico múltiple produce mejoras significativas en la tasa de inclusión del módulo de preselección, y que, sin embargo, la utilización de información léxica múltiple (varias matrices de confusión con los costes de inserción, sustitución y borrado asociados a cada conjunto de modelos) no produce ninguna mejora y complica el entrenamiento de dichos costes. De hecho, con la utilización de modelos acústicos dobles (este es el caso particular de modelado múltiple en el que el conjunto de locutores se divide en dos grupos homogéneos desde el punto de vista acústico), en una tarea de gran vocabulario (5000 y 10000 palabras) y la base de datos VESTEL, hemos obtenido mejoras relativas del error de inclusión entre un 23 % (con el diccionario de 5000 palabras) y un 39 % (con el de 10000) para una reducción del vocabulario activo del 90 %, con respecto a los resultados conseguidos con modelado simple. Para este mismo caso, hemos comprobado que el incremento computacional es permisible.

4. Robustez frente al ruido convolutivo. Parametrizaciones robustas

Los sistemas de reconocimiento automático del habla con arquitectura integrada suelen constar de dos módulos básicos: el módulo de parametrización y el módulo de clasificación o reconocedor propiamente dicho. El primero se encarga de la extracción de una serie de parámetros o rasgos acústicos que son una representación compacta de la señal de voz. El segundo compara dichos parámetros acústicos con los modelos acústicos de cada sonido (alófonos, en nuestro caso) y decide la palabra o frase

reconocida con mayor probabilidad. Ambos módulos son susceptibles de ser modificados para aumentar la robustez del sistema completo a las diferentes distorsiones antes mencionadas.

Las parametrizaciones robustas son, por tanto, el conjunto de técnicas que se aplican sobre el módulo de parametrización del sistema de RAH, para conseguir que las prestaciones del sistema no se degraden en presencia de diversos tipos de distorsiones. En nuestro caso, hemos optado por investigar en este grupo de técnicas para paliar los efectos de la distorsión introducida por el canal telefónico (ruido convolutivo), que aparecen en todo tipo de aplicaciones reconocimiento de voz sobre línea telefónica. Esta elección ha sido motivada por el hecho de que el módulo de parametrización es el que presenta menos requerimientos de memoria y tiempo de cómputo, de modo que su aplicación a sistemas de RAH en aplicaciones reales es factible.

En este ámbito, hemos analizado el funcionamiento de algunas técnicas ya conocidas y hemos propuesto una serie de parametrizaciones alternativas a las convencionales. En resumen, hemos trabajado en las siguientes líneas:

- Técnicas "clásicas" de extracción de parámetros. En particular, hemos evaluado las prestaciones de las parametrizaciones convencionales basadas en el análisis de Fourier tanto en el dominio cepstral (parámetros mel-cepstrum o MFCC), como en el dominio log-espectral (parámetros log-energías filtradas o FFLFBE).
- Técnicas de normalización de parámetros acústicos. Esta normalización consiste básicamente en la sustracción de la media y división por la varianza de los parámetros anteriores. En este caso, hemos hecho especial énfasis en trabajar con versiones adaptativas de estos algoritmos para reducir el retardo introducido por los mismos (para calcular las medias y varianzas de normalización, es preciso, en principio, disponer de toda la elocución lo que en un sistema práctico supone un retardo en el tiempo de proceso que puede ser inaceptable para el usuario).
- Parámetros derivados de la transformada ondicular. Hemos comprobado que la transformada ondicular, muy utilizada en el ámbito de la codificación de imágenes, es una parametrización alternativa interesante a las clásicas basadas en la transformada de Fourier, puesto que proporciona un mejor compromiso entre la resolución temporal y en frecuencia. También hemos trabajado en su combinación con las técnicas de normalización de parámetros antes mencionadas.
- Combinación de parámetros acústicos de distinta naturaleza. En este contexto, hemos propuesto la utilización conjunta en un solo vector de rasgos acústicos de los parámetros derivados de la transformada de Fourier y la transformada ondicular, planteando diversos tipos de combinaciones.
- Extracción discriminativa de rasgos (DFE): Aunque habitualmente los módulos de parametrización y clasificación de un sistema de RAH son diseñados de forma independiente, en la tesis, hemos abordado el problema de su optimización conjunta mediante la aplicación de técnicas de extracción discriminativa de rasgos (DFE), de modo que los parámetros extraídos sean los más discriminativos posibles desde el punto de vista del clasificador, es decir, contribuyan a la disminución de la tasa de error del sistema global. De este modo, es posible mejorar de forma significativa la tasa de reconocimiento del sistema final. En concreto, hemos aplicado este método con éxito para el ajuste

de las longitudes temporales de las ventanas de análisis en las parametrizaciones basadas en la transformada ondicular.

La Figura 2 es un resumen de los mejores resultados obtenidos con la utilización de las parametrizaciones diseñadas para combatir la distorsión producida por el canal telefónico. Los experimentos se realizaron sobre la base de datos por línea telefónica SpeechDat. Como puede observarse, las parametrizaciones propuestas basadas en la transformada ondicular en el dominio ceptral (MWCC) o en el dominio log-espectral (WFLFBE), ambas con normalización de parámetros y optimizadas con DFE, proporcionan tasas de reconocimiento superiores a las parametrizaciones convencionales correspondientes basadas en el análisis de Fourier y sin la extracción discriminativa de rasgos (MFCC y FFLFBE, respectivamente). En este experimento, hemos utilizado unos modelos acústicos muy sencillos, para poner de manifiesto la eficacia de estas parametrizaciones en situaciones en las que la reducción del consumo de memoria y tiempo de cómputo es un requerimiento, como es el caso de reconocedores de voz funcionando de forma local en teléfonos móviles u otros dispositivos portables (como "Personal Digital Assistant", PDAs).

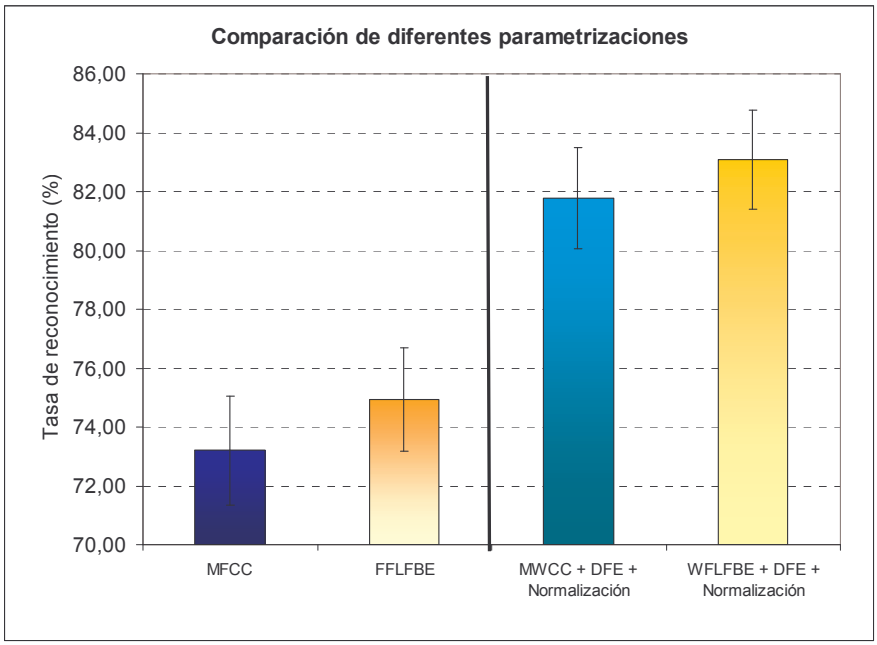


Figura 2. *Resultados más significativos con las parametrizaciones robustas propuestas y su comparación con los métodos clásicos.*

5. Robustez frente al ruido aditivo. Transformación de parámetros y modelos acústicos

En el contexto de las distorsiones provocadas por la presencia de ruido aditivo o ruido de fondo, muy habituales en situaciones prácticas, hemos enfocado el problema desde dos perspectivas diferentes: técnicas de transformación de parámetros y técnicas de transformación de modelos acústicos. Ambos conjuntos de técnicas tienen un alto grado de semejanza, en cuanto que se basan en la definición de una función de transformación (o función de entorno) que relaciona los parámetros/modelos de voz sin contaminar con

los de voz contaminada y viceversa. La principal diferencia estriba en que en el primer caso, el objeto del método es minimizar la presencia del ruido aditivo de los vectores de rasgos acústicos que representan la señal de voz y se aplica sobre el módulo de parametrización, mientras que en el segundo caso, el método trata de contaminar los modelos acústicos generados a partir de voz limpia de modo que sus características se adecuen lo más posible a las de la señal ruidosa de entrada al sistema y se aplican sobre el módulo de clasificación.

Para ambos casos, hemos propuesto la definición matemática de nuevas funciones de entorno que representaran más fielmente el proceso de contaminación por ruido de la señal de voz y que además fueran lo suficientemente sencillas como para permitir una mayor tratabilidad matemática y su aplicación en sistemas de RAH reales sin incrementar de forma considerable los requerimientos de memoria y cómputo.

En particular, se ha trabajado en funciones de entorno aplicables al dominio de las energías en banda que incorporan como novedad, respecto a las funciones de transformación comúnmente utilizadas, la estimación de términos cruzados entre la densidad espectral de la señal ruidosa y el ruido, que están relacionados con la correlación existente entre el habla ruidosa y el ruido.

La Figura 3 muestra un resumen de los resultados obtenidos con voz contaminada con ruido de coche a diferentes relaciones señal a ruido (SNR). La figura de la izquierda contiene las tasas de reconocimiento para las técnicas de transformación de parámetros y los siguientes casos: sin transformación (base), función de transformación clásicas sin término de correlación cruzada (STC) y la función de entorno propuesta con término de correlación cruzada (CTC) y la de la derecha representa los mismos casos cuando la transformación se realiza directamente sobre los modelos acústicos.

Como puede observarse, para el caso de transformación de parámetros, la inclusión del término de correlación en la función de entorno (CTC) mejora sensiblemente las tasas de reconocimiento del sistema en condiciones de SNR bajas con respecto al experimento base (base) y el experimento sin incluir dicho término (STC). Para relaciones señal a ruido más elevadas, las técnicas propuestas ofrecen resultados similares a los obtenidos sin utilizar ningún tipo de compensación. Sin embargo, son mejores que los del método clásico. Este resultado nos sugiere que son técnicas más estables con respecto a la variación en el grado de contaminación de la señal.

Con respecto a la transformación de modelos, puede observarse que el método clásico (STC) incrementa drásticamente la tasa de reconocimiento del sistema respecto al experimento base. Por otro lado, la incorporación del término de correlación cruzada (CTC) entre la señal limpia y el ruido produce sólo pequeñas mejoras respecto al método clásico. Esto sugiere que las técnicas de transformación de modelos son más insensibles al refinamiento propuesto de la función de entorno.

Además, comparando ambas figuras, puede comprobarse que las prestaciones de las técnicas de transformación de parámetros son inferiores al caso de transformación de modelos. Sin embargo, desde un punto de vista práctico, las técnicas que modifican los parámetros tienen utilidad en un mayor número de aplicaciones y presentan la ventaja adicional de permitir una rápida adaptación a condiciones cambiantes del ruido de fondo y la posibilidad de ser aplicadas a un mayor número de parametrizaciones distintas que en el caso de las técnicas que trabajan directamente sobre los modelos. Además, la cantidad de memoria y capacidad computacional que necesitan las técnicas basadas en transformación de modelos es muy superior a las de transformación de parámetros, lo

que puede limitar su utilización en aplicaciones sobre teléfonos móviles y otros dispositivos portables como PDAs.

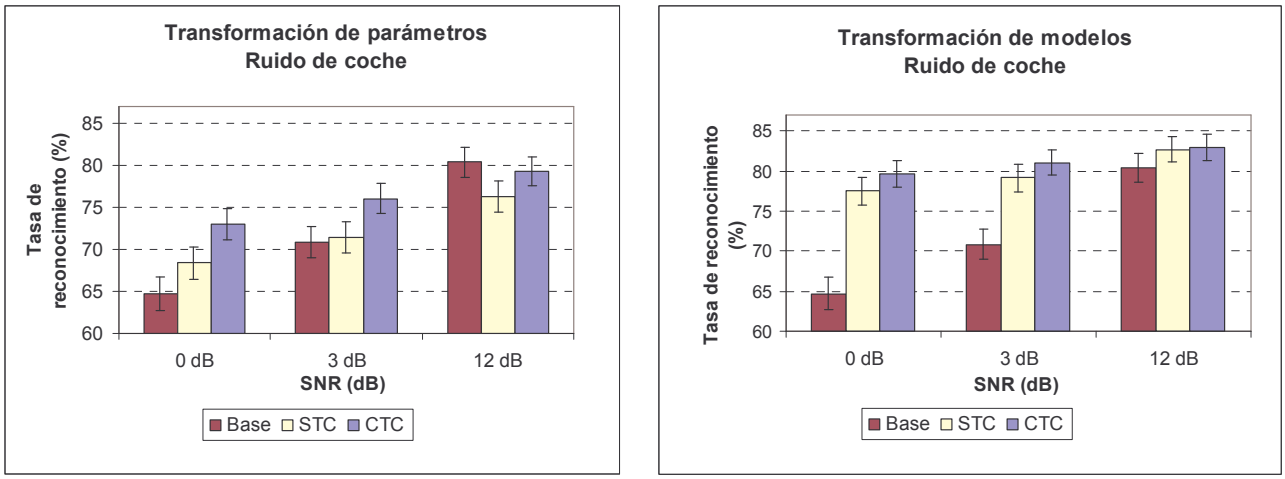


Figura 3. *Resultados con las técnicas de transformación de parámetros (izquierda) y modelos (derecha) con ruido aditivo de coche.*

6. Aplicación práctica e interés industrial

Como hemos mencionado en el apartado dedicado al contexto de esta tesis doctoral, la utilización de los sistemas de reconocimiento de habla en aplicaciones reales depende de que sean capaces de mantener sus prestaciones en una gran variedad de escenarios en los que están presentes diversos tipos de ruidos y distorsiones acústicas. Recientemente, en una conferencia internacional sobre la robustez de los sistemas de RAH, y tras una serie de encuestas informales realizadas entre investigadores y empresas relacionadas con las tecnologías de reconocimiento de habla, se llegó a la conclusión de que la limitación más significativa de la tecnología existente era precisamente la falta de robustez frente a estas condiciones adversas.

Desde este punto de vista, las diversas técnicas presentadas en esta tesis doctoral, pueden ayudar a mejorar la tasa de reconocimiento de los sistemas de RAH funcionando en condiciones acústicas adversas, que son las habitualmente presentes en aplicaciones reales. Además, en esta tesis se ha hecho un énfasis especial en la investigación y desarrollo de técnicas con demandas de memoria y cálculo computacional bajas o moderadas, lo que hace que su utilización sea perfectamente factible en muy diversas aplicaciones y en muy diversos escenarios, algunos tan novedosos como el reconocimiento de voz sobre teléfonos móviles o sobre otros dispositivos portables como PDAs, cuyo uso se está extendiendo considerablemente en los últimos años.

7. Relación cronológica de las publicaciones de la autora

Publicaciones en revistas internacionales:

[1] C. Peláez, A. Gallardo Antolín y F. Díaz de María., "Recognizing Voice over IP: A New Front-End for Speech Recognition on the World Wide Web", *IEEE Transactions on Multimedia*, vol. 3, nº 2, pp. 209-218, Junio 2001.

[2] A. Gallardo-Antolín, C. Peláez y F. Díaz-de-María, "Recognizing from GSM Digital Speech", *IEEE Transactions on Speech and Audio Processing*, aceptado para su publicación, 2004.

Capítulos de libro:

[3] C. Peláez, A. Gallardo Antolín y F. Díaz de María, "Recognizing Speech over IP: Towards Spoken Language Interfaces for E-business", *E-Business: Key Issues, Applications and Technologies*, Ed. Brian Stanford-Smith and Paul T. Kidd (IOS Press and Ohmsha), pp. 1065-1071, 2001.

[4] C. Peláez, E. Parrado, A. Gallardo-Antolín, A. Zambrano y F. Díaz de María, "An Application of SVM to Lost Packets Reconstruction in Voice-Enabled Services, *Lecture notes in computer science; Artificial Neural Networks (ICANN'02)*, Ed. Springer-Verlag, vol. 2415, pp. 1174-1179, 2002.

[5] F. Díaz-de-María, A. Gallardo-Antolín y C. Peláez.Moreno, "Voice over IP over Wireless: Principles and Challenges", *Signal Processing for Mobile Communications Handbook*, Ed. CRC Press, prevista su aparición para mediados de 2004.

Publicaciones en congresos internacionales:

[6] J. Macías, A. Gallardo, J. Ferreiros y J. M. Pardo y L. Villarrubia, "Initial Evaluation of a Preselection System for a Flexible, Large Vocabulary Speech Recognition System in Telephone Environment", *4th International Conference on Spoken Language Processing (ICSLP'96)*, vol. II, pp. 1343-1346, Philadelphia, USA, 1996.

[7] J. Ferreiros, J. Macías, A. Gallardo-Antolín, R. Córdoba, J. M. Pardo y L. Villarrubia, "Recent Work on a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment", *5th International Conference on Spoken Language Processing (ICSLP'98)*, vol. 2, pp. 321-324, Sydney, Australia, 1998.

[8] A. Gallardo Antolín, F. Díaz de María y F. Valverde, "Recognition from GSM Digital Speech", *5th International Conference on Spoken Language Processing (ICSLP'98)*, vol. 4, pp. 1443-1446, Sydney, Australia, 1998.

[9] A. Gallardo Antolín, F. Díaz de María y F. Valverde, "Avoiding Distortions due to Speech Coding and Transmission Errors in GSM ASR Tasks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, vol. 1, pp. 277-280, Phoenix, US, 1999.

[10] J. Macías, J. Ferreiros, A. Gallardo, R. San Segundo, J. M. Pardo y L. Villarrubia, "A Variable Preselection List Length Estimation Using Neural Networks in a Telephone Speech Hyphotesis-Verification System", *6th European Conference on*

Speech Communication and Technology (EUROSPEECH'99), vol. 1, pp. 295-298, Budapest, Hungría, 1999.

[11] A. Gallardo Antolín, M. Vázquez de Castro, F. Díaz de María, F. Valverde y F. Pérez Fontán, BER Performance Assessment of the Land Mobile GSM Channel with Application to Automatic Speech Recognition Tasks", *5th Bayona Workshop on Emerging Technologies in Telecommunications*, vol. 1, pp. 212-216, Bayona, España, 1999.

[12] A. Gallardo Antolín, J. Ferreiros, J. Macías Guarasa, R. Córdoba, J. Colás y J. M. Pardo, "Incorporating Multiple-HMM Acoustic Modeling in a Modular Large Vocabulary Speech Recognition System in Telephone Environment", *6th International Conference on Spoken Language Processing (ICSLP'00)*, vol. 2, pp. 827-830, Pekín, China, 2000.

[13] J. Macías Guarasa, J. Ferreiros, J. Colás, A. Gallardo Antolín y J. M. Pardo, "Improved Variable Preselection List Length Estimation Using NNs in a Large Vocabulary Telephone Speech Recognition System", *6th International Conference on Spoken Language Processing (ICSLP'00)*, vol. 2, pp. 823-826, Pekín, China, 2000.

[14] A. Gallardo Antolín, C. Peláez y F. Díaz de María, "A Robust Front-End for ASR over GSM and IP Networks: an Integrated Scenario", *7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, pp. 1103-1106, Aalborg, Dinamarca, 2001.

[15] C. Peláez, A. Gallardo-Antolín, E. Parrado y F. Díaz de María, "SVM-Based Lost Packets Concealment for ASR Applications over IP", *XI European Signal Processing Conference (EUSIPCO'02)*, vol. 3, pp. 529-532, Toulouse, Francia, 2002.

[16] C. Peláez, A. Gallardo-Antolín, J. Vicente-Peña y F. Díaz de María, "Filtering the Spectral Parameters to Mitigate the Influence of Transmission Errors on ASR Systems", *7th International Conference on Spoken Language Processing (ICSLP'02)*, pp. 2217-2220, Denver, USA, 2002.

[17] F. Díaz de María, J. Vicente-Peña, A. Gallardo-Antolín y C. Peláez, "Linear Equalization of the Modulation Spectra: A Novel Approach for Noisy Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)*, vol. II, pp. 141-144, Hong-Kong, China, 2003.

[18] A. Gallardo-Antolín, J. Macías-Guarasa, J. Ferreiros, R. Córdoba, J. M. Montero, R. San-Segundo y J. M. Pardo Muñoz, "A Comparison of Several Approaches to the Feature Extractor Design for ASR Tasks in Telephone Environment", *15th International Congress of Phonetic Sciences (ICPhS'03)*, pp. 1345-1348, Barcelona, España, 2003.

Publicaciones en congresos nacionales:

[19] A. Gallardo Antolín, F. Díaz de María, F. Valverde y R. Bravo, "Reconocimiento de Voz Procedente de Teléfonos Móviles Digitales", *VIII Jornadas de I+D en Telecomunicaciones (TELECOM I+D'98)*, vol. I, pp. 379-387, Madrid, España, 1998.

[20] C. Peláez, A. Zambrano Miranda, A. Gallardo Antolín y F. Díaz de María, "Reconocimiento de Habla en Internet: una Aproximación Eficiente", *IX*

JORNADAS de I+D en Telecomunicaciones (TELECOM I+D'99), Madrid y Barcelona, España, 1999.

[21] R. San-Segundo, J. Colás, J. M. Montero, R. Córdoba, J. Ferreiros. J. Macías-Guarasa, A. Gallardo Antolín, J. M. Gutiérrez, J. Pastor y J. M. Pardo, "Servidores Vocales Interactivos: Desarrollo de un servicio de páginas blancas por teléfono con reconocimiento de voz (Proyecto IDAS: Interactive telephone-based Directory Assistance Service)", *IX JORNADAS de I+D en Telecomunicaciones (TELECOM I+D'99)*, Madrid y Barcelona, España, 1999.

[22] R. San Segundo, J.M. Montero, J. Colás, J. Ferreiros, R. de Córdoba, A. Gallardo Antolín, J. Macías-Guarasa, J.M. Gutiérrez, J. Pastor y J. M. Pardo, "Entorno para el desarrollo de aplicaciones multimedia con síntesis y reconocimiento de voz", *X JORNADAS de I+D en Telecomunicaciones (TELECOM I+D'00)*, Madrid y Barcelona, España, 2000.

[23] R. Córdoba, R. San-Segundo, J. Colás, J.M. Montero, J. Ferreiros. J. Macías-Guarasa, A. Gallardo, J.M. Gutiérrez y J.M. Pardo, "Optimización de un servicio automático de páginas blancas por teléfono: proyecto IDAS), *X JORNADAS de I+D en Telecomunicaciones (TELECOM I+D'00)*, Madrid y Barcelona, España, 2000.

[24] A. Gallardo Antolín, J. Macías Guarasa, R. San Segundo, J. Ferreiros, R. Córdoba y J. M. Pardo Muñoz, "Comparación de diversas parametrizaciones para reconocimiento de habla robusto en entorno telefónico", *II Jornadas en Tecnologías del Habla*, Granada, España, 2002.

Otras publicaciones:

[25] Gallardo Antolín, I. Mayoral y J.M. Pardo, "Automatic Speech Recognition in Additive Noise (NOISEX-92)", *Research Report GTH-DIE-ETSIT-UPM 1/97*, Noviembre 1997.

[26] A. Gallardo Antolín, I. Mayoral y J.M. Pardo, "Automatic Speech Recognition Under Simulated Stress Conditions", *Research Report GTH-DIE-ETSIT-UPM 2/97,* Noviembre 1997.

[27] F. Díaz de María, A. Gallardo Antolín y F. J. Valverde, "Speaker-independent Isolated-word Speech Recognition for GSM Mobile Communications. First Progress Report", *Research Report*, 1998.

8. Otros méritos relacionados con la tesis doctoral

Referencias al trabajo por parte de otros investigadores:

- Se referencia: *"Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment", J. Macias-Guarasa, A. Gallardo, J. Ferreiros, J. M. Pardo y L. Villarrubia, 5th International Conference on Spoken Language Processing (ICSLP'96), vol. 1, pp. 30-33, Philadelphia, USA, 1996.*
    - En: "Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation", M. J. Caraty, C. Montacié and F. Lefèvre, 5th European Conference on Speech Communication and Technology (EUROSPEECH'97), vol. 5 pp. 2691-2694, Rodas, Grecia, 1997.

- Se referencia: *"Automatic Speech Condition Under Simulated Stress Conditions", A. Gallardo Antolín, I. Mayoral y J. M. Pardo, Research Report GTH-DIE-ETSIT-UPM2/97, Universidad Politécnica de Madrid, 1997.*
    - En: "Speech Under Stress Conditions: Overview of the Effect on Speech Production and on System Performance", H. K. Steeneken y J.H.L. Hansen, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99), vol. 4, pp. 2079-2082, 1999.
- Se referencia: *"Recent Work on a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment", J. Ferreiros, J. Macías-Guarasa, A. Gallardo, J. Colás, R. de Córdoba, J.M. Pardo and L. Villarrubia, 5th International Conference on Spoken Language Processing (ICSLP'98), vol. 1,.pp. 321-324, Sydney, Australia, ISBN: 1-876346-17-5, 1998.*
    - En: "Utterance Verification Based Speech Recognition System", B. T. Tan, Y. Gu, y T. Thomas., 6th International Conference on Spoken Language Processing (ICSLP'00), vol. 2, pp. 899-902, Pekín, China, ISBN: 7-80150-114-4/G.18, 2000.
- Se referencia: *"Recognition from GSM Digital Speech", A. Gallardo Antolín, F. Díaz de María and F. Valverde, 5th International Conference on Spoken Language Processing (ICSLP'98), vol. 4, pp. 1443-1446, Sydney, Australia, ISBN: 1-876346-17-5, 1998.*
    - En: "Speech recognition in mobile environments", J. M. Huerta, Tesis doctoral. Carnegie Mellon University, 2000.
    - En: "Bistream-based feature extraction for wireless speech recognition", H. K. Kim and R. V. Cox , IEEE International Conference on Speech and Audio Processing (ICASSP'00), Istambul (Turkey), 2000.
    - En: "Distributed Speech Recognition with Codec Parameters", B. Raj, J. Migdal and R. Singh, Proceedings of Automatic speech recognition and understanding workshop (ASRU 2001), Italy, 2001.
    - En: "A bitstream-based front-end for wireless speech recognition on IS–136 communications system", H. K. Kim and R. V. Cox, IEEE Transactions on Speech and Audio Processing vol 9, no5, pp. 558- 568. julio 2001.
    - En: "Distortion-class modeling for robust speech recognition under GSM RPE-LTP coding", J. M. Huerta and Richard M. Stern, Speech Communication, 34, pp. 213-225, 2001.
    - En: "Reconocimiento de voz en el entorno de las nuevas redes de comunicación UMTS e Internet", L. Villarrubia Grande, I. Cortázar Mújica, L. Hernández Gómez y E. López Gonzalo, Comunicaciones de Telefónica I+D, vol. 5 (nº 2), pp.: 3-27, 2001.
    - http://www.tid.es/presencia/publicaciones/comsid/esp/23/08.pdf
    - En: "Speech Coding and Transmission for Improved Automatic Recognition", X. Zhong, J. Arrowood y M. Clements, Int. Conf. Spoken Language Processing (ICLSP'02), 2002. Disponible en http://users.ece.gatech.edu/~gt4517b/pub.html.
- Se referencia: *"Avoiding Distortions due to Speech Coding and Transmission Errors in GSM ASR Tasks", A. Gallardo Antolín, F. Díaz de María y F. Valverde, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99), vol. 1, pp. 277-280, 1999.*

- En: “Speech recognition in mobile environments”, J. M. Huerta, Tesis doctoral. Carnegie Mellon University, 2000.
- En: “Low bit rate speech compression for playback in speech recognition systems", D. Chazan, G. Cohen, R. Hoory and M. Zibulski, Proc. European Signal Processing Conference, (EUSIPCO 2000), Tampere Finland, Sept. 2000.
- En: “Distortion-class modeling for robust speech recognition under GSM RPE-LTP coding”, J. M. Huerta y R. M. Stern, Speech Communication, 34, pp. 213-225, 2001.
- En: “Feature Enhancement for a bitstream-based front-end in wireless speech recognition”, H. K. Kim and R. V. Cox, IEEE International Conference on Speech and Audio Processing, ICASSP’01, Salt-Lake City (USA), 2001.
- En: “Soft-Feature Decoding for Speech Recognition over Wireless Channels”, A. Potamianos and V. Weerackod, IEEE International Conference on Speech and Audio Processing, ICASSP’01, Salt-Lake City (USA), 2001.
- En: "Source and channel coding for remote speech recognition over error-prone channel”, A. Bernard and A. Alwan, IEEE International Conference on Speech and Audio Processing (ICASSP’01), Vol. 4, pp. 2613-2616, Salt-Lake City, (USA), 2001.
- En: “Joint Channel Decoding – Viterbi Recognition for Wireless Applications”, A. Bernard and A. Alwan, Eurospeech 2001, Vol. 4, pp. 2703-2706, Aalborg, Denmark, 2001.
- En: “Graceful Degradation of Speech Recognition Performance over Lossy Packet Networks”, C. Boulis, M. Ostendorf, E. A. Riskin, and S. Otterson, UWEE Technical Report Number UWEETR-2001-0003 Department of Electric Engineering, University of Washington, October 2001. Disponible en https://www.ee.washington.edu/techsite/papers/titles2001.html.
- En: “Low-Bitrate Distributed Speech Recognition for Packet-Based and Wireless Communication”, A. Bernard y A. Alwan, IEEE Transactions on Speech and Audio Processing, vol. 10, (nº 8), pp. 570-579, 2002.
- En: “Graceful Degradation on Speech Recognition Performance Over Packet-Erasure Networks”, C. Boulis, M. Ostendorf, E. A. Riskin y S. Otterson, IEEE Transactions on Speech and Audio Processing, vol. 10, (nº 8), pp. 580-590, 2002.
- En: “Performance Improvement of Bitstream-Based Front-End for Wireless Speech Recognition in Adverse Enviroments”, H. K. Kim, R. V. Cox y R. C. Rose, IEEE Transactions on Speech and Audio Processing, vol. 10, (nº 8), pp. 591-604, 2002.
- En: “A simulative study of distributed speech recognition over internet protocol networks”, D. Quercia, Master of Science in Electrical and Computer Engineering in the Graduate College of the University of Ilinois at Chicago, http://www.stud.uni-karlsruhe.de/~uzyg/thesis/thesis.pdf, 2002.

- Se referencia: *“Recognizing Voice over IP: A New Front-End for Speech Recognition on the World Wide Web”, C. Peláez Moreno, A. Gallardo Antolín y F. Díaz de María, IEEE Transactions on Multimedia, vol. 3, (nº 2), pp. 209-218, 2001.*
  - En: “Multiple Description Coding for Recognizing Voice over IP”, X. Zhong, J. Arrowood, A. Moreno y M. Clements, 10th IEEE DSP Workshop, 2002. Disponible en http://users.ece.gatech.edu/~gt4517b/pub.html.

- En: “Graceful Degradation on Speech Recognition Performance Over Packet-Erasure Networks”, C. Boulis, M. Ostendorf, E. A. Riskin y S. Otterson, IEEE Transactions on Speech and Audio Processing, vol. 10, (nº 8), pp. 580-590, 2002.
- En: “A simulative study of distributed speech recognition over internet protocol networks”, D. Quercia, Master of Science in Electrical and Computer Engineering in the Graduate College of the University of Ilinois at Chicago, http://www.stud.uni-karlsruhe.de/~uzyg/thesis/thesis.pdf, 2002.
- En: "Overview of compression and packet loss effects in speech biometrics", L. Besacier, P. Mayorga, J. F. Bonastre, C. Fredouille and S. Meignier, IEE Proceedings – Vision, Image and Signal Processing, vol. 150 (nº 6), pp. 372-376, December 2003.

Proyectos a los que se asocia la tesis:

- “ONOMASTICA Multingual Pronunciation Dictionaries of Proper and Plain Names” (“ONOMÁSTICA: Diccionario de pronunciación de nombres propios y toponímicos”). Proyecto Acción Especial CICYT TIC 1994-1525-CE.
- "Speech under Stress Conditions ". Proyecto del grupo experto de la NATO: RTO Information Systems Technology Panel (IST). RTO-TR-10, ETSIT - Universidad Politécnica de Madrid y otros grupos de investigación europeos.
- “Proyecto VOZ: Sistema automático de información telefónica”. Financiado por el Rectorado de la Universidad Politécnica de Madrid.
- “Módulo de búsqueda rápida para un sistema de reconocimiento de gran vocabulario sobre línea telefónica”. Financiado por Telefónica I+D.
- “Sistema SERVIVOX. Sistema para la automatización de servicios telefónicos”. Financiado por Hewlett Packard Española.