Madrid, 27 de Febrero de 2005

Estimado Sr. Decano-Presidente del Colegio Oficial de Ingenieros de Telecomunicación,

Ruego que considere la Tesis Doctoral titulada “Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano” de la que soy autor como candidata al Premio COIT/AEIT a la Mejor Tesis Doctoral en Comercio Electrónico en la XXV Convocatoria de Premios “Ingenieros de Telecomunicación”.

Sin otro particular, se despide afectuosamente,

Juan Manuel Montero Martínez

UNIVERSIDAD POLITÉCNICA DE MADRID

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TESIS DOCTORAL

ESTRATEGIAS PARA LA MEJORA DE LA NATURALIDAD Y LA INCORPORACIÓN DE VARIEDAD EMOCIONAL A LA CONVERSIÓN TEXTO A VOZ EN CASTELLANO

JUAN MANUEL MONTERO MARTÍNEZ
Ingeniero de Telecomunicación

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

Madrid, 2003

Datos de la Tesis Doctoral

Autora:

JUAN MANUEL MONTERO MARTÍNEZ

Título:

ESTRATEGIAS PARA LA MEJORA DE LA NATURALIDAD Y LA INCORPORACIÓN DE VARIEDAD EMOCIONAL A LA CONVERSIÓN TEXTO A VOZ EN CASTELLANO

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:

14 de Noviembre de 2003

Calificación:

Sobresaliente Cum Laude por unanimidad del tribunal

Número de colegiado/asociado: A09073

Resumen de la tesis doctoral

1. Contexto de la tesis doctoral

La Tesis está enmarcada en el campo de las tecnologías del habla, concretamente en los de la Conversión Texto a Voz (CTV) y el Procesamiento de Lenguaje Natural (PLN), cuyo objetivo fundamental es la conversión de un texto en habla por parte de máquinas, alcanzando una inteligibilidad y una naturalidad que la hagan indistinguible del habla humana.

La importancia de la comunicación oral es evidente en la vida de los seres humanos. En todas las sociedades, incluso en las más primitivas, la comunicación oral existe y está basada en mecanismos acústicos, sintácticos y semánticos complejos, con independencia del nivel tecnológico alcanzado por la sociedad en cuestión.

En las últimas décadas se han realizado grandes esfuerzos en el área de la síntesis 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.

La tecnología actual es capaz de convertir texto en voz con una alta tasa de inteligibilidad, aunque su grado de naturalidad no sea tan alto como desearíamos: no podemos imitar el amplio espectro de cadencias, melodías y cualidades que cubre la voz humana. Por lo general las voces sintéticas podían ser catalogadas como monótonas o incluso aburridas: nuestros ordenadores carecían hasta ahora de capacidad para transmitirnos emociones, para adaptar la voz a diferentes estilos de locución (formales o informales), para engañarnos con su naturalidad.

A medida que las tecnologías del habla se han ido implicando cada vez más como parte integral de aplicaciones prácticas en escenarios reales (como servicios de obtención de información por línea telefónica [5] [4] [3] [18] [19] [35] [36], control de dispositivos e instrumental en coches, asistentes personales (PDAs), robots o agentes virtuales animados dotados de personalidad propia [11] [16], etc.), se ha hecho patente la necesidad de desarrollar sistemas automáticos de conversión texto-habla naturales y dotados de la variedad que nos caracteriza a los seres humanos.

Vivimos una época en la que se está dando un gran auge en los estudios teórico-prácticos de la llamada inteligencia social o inteligencia emotiva, la que se encarga de controlar con inteligencia las propias emociones, reconocer las emociones de los demás y reaccionar empáticamente a las mismas. No es descabellado pensar en que esa misma inteligencia social debería gobernar las futuras aplicaciones de intercomunicación hombre-máquina, haciéndolas más y más amigables (*user-friendly*), tanto si son presenciales como si son telefónicas o telemáticas [23] [31] [33]. Para ello, deberíamos de dotar a los sintetizadores de una voz más diversa y humana: un usuario habitual del sistema o un usuario con problemas que dialoga con el sistema, y tras repetidos intentos no consigue acceder a la información que precisa, deben ser tratado de un modo especial, como lo haría un experto humano.

Tres son las grandes líneas que se deben abordar para avanzar hacia la consecución del ambicioso objetivo final, dos de las cuales se abordan en esta tesis:

- Mejorar el procesamiento automático lingüístico-prosódico que, a partir de texto, debe recoger la información sintáctica y semántica del mismo y aprovecharla para

generar una voz más natural desde el punto de vista de su ritmo y su entonación. Colateralmente, las técnicas empleadas son de gran utilidad en comprensión automática de habla en sistemas de diálogo [30] [28] [6].

- Analizar y generar variantes de una voz artificial, en especial incorporar emociones y actitudes que la doten de personalidad y empatía [1] [13] [29]. Esto resulta de utilidad en estudios forenses [32] [21].
- Mejorar la capacidad de imitación del timbre de cada uno de los locutores apropiados para las aplicaciones más habituales.

En esta tesis se ha tratado una parte importante de los dos primeros puntos.

2. Objetivos de la tesis doctoral

El objetivo de la tesis doctoral ha sido profundizar en el estudio de diversas estrategias para incorporar naturalidad y variedad emocional en la conversión texto a voz en castellano, haciendo especial énfasis en el procesado lingüístico orientado al modelado de la prosodia, en el modelado de la frecuencia fundamental dentro de un dominio restringido y en el análisis, modelado y síntesis de voz con emociones.

Los sistemas de síntesis de habla sobre los que se ha realizado este estudio comprenden las principales tecnologías del estado del arte, comenzando con el sistema multilingüe de síntesis por formantes de la empresa sueca Telia-Infovox [15] hasta el sistema registrado Boris de síntesis por concatenación, desarrollado por el autor de la tesis y otros miembros del Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la Universidad Politécnica de Madrid. En la actualidad, ambos sistemas funcionan en aplicaciones en tiempo real, motivo por el cual, a lo largo de toda la tesis, se ha optado por trabajar con técnicas que no requieran 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. Procesamiento de Lenguaje Natural

En el capítulo dedicado a las investigaciones en procesado lingüístico del texto se comienza describiendo en detalle los corpora empleados en la experimentación (que incluyen 54 millones de palabras procedentes de 2 años completos de artículos del periódico El Mundo), tanto en normalización como en etiquetado. La técnica desarrollada en normalización emplea reglas de experto, con muy buenos resultados tanto en precisión como en cobertura (>85%), destacando el empleo de reglas de silabicación para la detección precisa de palabras extranjeras (>99%). La cobertura del sistema desarrollado alcanza un nivel casi insuperable (>99,8%), lo cual confirma la calidad del trabajo llevado a cabo.

Al afrontar la desambiguación gramatical, se comparan tres técnicas: reglas de experto (tasa inferior al 98%), aprendizaje automático de reglas (tasa por debajo del 99%) y modelado estocástico (por encima del 99%), obteniéndose los mejores resultados con esta última técnica, debido a su capacidad de procesar más adecuadamente textos fuera del dominio de entrenamiento.

Finalmente se aborda el análisis sintáctico por medio de gramática de contexto libre como un proceso en dos fases: una primera sintagmática (*shallow parsing*) y una segunda relacional básica, a fin de maximizar la cobertura del análisis. Para la resolución de las ambigüedades que nos permiten alcanzar gran cobertura se adapta el principio de mínima longitud de descripción con notables resultados. Con las gramáticas desarrolladas se alcanza una tasa de cobertura y precisión superior al 96% en

el caso del análisis sintagmático y superiores al 87% en el sintáctico, los mejores resultados publicados en castellano para unas tareas tan complejas.

4. Modelado prosódico en dominio restringido

Para el modelado de F0 en un dominio restringido se emplean perceptrones multicapa. En una primera etapa se describe y evalúa una nueva técnica de diseño de base de datos basada en un algoritmo voraz moderado mediante subobjetivos intermedios. Esta novedosa combinación de voracidad y moderación permite alcanzar precisiones superiores al 95% para los numerosos rasgos que relacionados con la predicción de la prosodia [10] [9].

La exhaustiva experimentación con los diversos parámetros de predicción, la configuración de la red y las subdivisiones de la base de datos ocupa la mayor parte del capítulo, destacando la aportación de un parámetro específico del dominio restringido (el número de la frase portadora del texto que sintetizar) junto a otros más clásicos (la acentuación, el tipo de grupo fónico y la posición en el mismo), además del análisis sobre cómo agrupar las grabaciones para obtener el mejor modelado posible [8] [2] [22] [27].

5. Análisis y síntesis de voz con emociones

El capítulo dedicado a la voz emotiva comienza detallando el proceso de creación de una nueva voz castellana masculina en síntesis por formantes con modelo mejorado de fuente (reglas y metodología), evaluando las posibilidades de personalización de voz que ofrece. La voz desarrollada, por su calidad, fue adoptada por el sintetizador multilingüe de Infovox en castellano [14].

Para trabajar con voz con emociones se diseña, graba y etiqueta una base de datos de voz en la que un actor simula tristeza, alegría, sorpresa, enfado y también una voz neutra. Por medio de técnicas paramétricas (modelo de picos y valles en tono, y multiplicativo en las duraciones) se analiza prosódicamente la base de datos y se establece una primera caracterización de la voz en las distintas emociones. Empleando como base la voz personalizable se desarrolla el sistema completo de conversión texto a voz con emociones y se evalúa, destacando la rápida adaptación de los usuarios en cuanto a la identificación de la emoción expresada [12]. Finalmente se experimenta con síntesis por concatenación y síntesis por copia, llegando a las siguientes conclusiones: la voz sorprendida se identifica prosódicamente, las características segmentales son las que caracterizan al enfado en frío; y, finalmente, la tristeza y la alegría son de naturaleza mixta, hallazgo pionero que fue posteriormente confirmado por posteriores investigadores en diversos países y diversas lenguas [7].

6. Conclusiones

Procesado lingüístico automático

Se han probado tres técnicas de desambiguación contextual gramatical:

1. una basada en reglas manuales (cuyo coste de desarrollo no se ve compensado con una tasa adecuadamente elevada);

2. otra basada en reglas inferidas automáticamente (que supere la dificultad de generar manualmente las reglas). En esta técnica de aprendizaje de reglas los resultados han sido excelentes, aunque bastante dependientes del dominio de entrenamiento (pudiéndose reducir la tasa desde un 99% a menos de un 96%), debido al sobre entrenamiento y al aprendizaje de reglas no generales;
3. otra basada en modelado estocástico, constatándose el superior comportamiento de esta última técnica al realizar ensayos fuera de dominio. Al emplear la técnica estocástica con diccionarios no adaptados a un dominio concreto sin probabilidades, se ha alcanzado una cobertura del 99,89% en textos de un dominio distinto al dominio de entrenamiento, comparable a los mejores sistemas en castellano y que (a pesar de no tener probabilidades) supera significativamente en precisión los resultados de un sistema léxico basado en probabilidades, aunque a costa de una mayor ambigüedad media (>96% en el primer candidato). En la desambiguación, resulta también significativa la mejora debida al tratamiento de las locuciones, dado que los modelos probabilísticas basados en categorías no son capaces de modelar bien contextos amplios y con pocos ejemplos.

Se ha adaptado un sistema de análisis por medio de gramáticas de contexto libre, desarrollando y evaluando con éxito una gramática robusta de dominio general en dos niveles, uno sintagmático y otro relacional. Cabe destacar el empleo de reglas de corte para reducir el número de análisis posibles (con sólo un 0,35% de imprecisión), la aplicación de reglas de concordancia como filtrado posterior al análisis y el uso de un criterio muy simple de número mínimo de segmentos para elegir el mejor análisis (sin necesidad de información probabilística). En el primer nivel sintagmático los resultados han sido excelentes (cobertura y precisión superiores al 96%, comparables a los mejores resultados en castellano, aunque sobre distinto corpus), a pesar de que haya un 1% de errores debidos al etiquetado previo. En el segundo nivel relacional los resultados son prometedores (cobertura y precisión superiores al 87%).

En el nivel léxico, se ha experimentado con diccionarios adaptados al dominio, con diccionarios generales y con diccionarios extranjeros, destacando la aportación de los dos primeros tipos. También se ha constatado la necesidad de incluir reglas robustas para etiquetar palabras fuera de vocabulario, experimentándose con mejoras significativas el empleo de reglas de experto basadas en las terminaciones de las palabras. En este sentido se ha ensayado el empleo de varios conjuntos de reglas manuales de experto procedentes de otros proyectos (completadas con algunas nuevas reglas), aunque la inclusión de los diccionarios ha obligado a filtrarlas y adaptarlas, incrementando su precisión de un 77,5% a un 98,88% (aplicada a un 24,8% de las palabras desconocidas). Se ha incorporado un conjugador verbal de gran cobertura basado en un paradigma sencillo pero efectivo; a pesar de la sobregeneración de este módulo, no se han producido importantes incrementos en el número de etiquetas por palabra.

Trabajando en el nivel de palabra, se han estudiado los distintos tipos de palabras no estándar y la manera de detectar su presencia en un texto, de manera que sea posible procesar no sólo un corpus convenientemente preparado, sino textos obtenidos directamente del dominio sin supervisión. Se ha creado y evaluado un normalizador de texto basado en diccionarios genéricos y diccionarios especializados y en reglas de experto empotradas. La precisión global alcanzada fue del 98,41%, mientras que la precisión sobre las palabras no estándar fue siempre superior al 85% sobre un corpus de evaluación de textos periodísticos, destacando el 96% en nombres propios simples. Para la detección de palabras y nombres propios extranjeros se ha probado un sencillo

método basado en las reglas de silabicación del castellano, cuya precisión supera el 99,5%, aunque cubre pocos casos.

Modelado de F0 en dominio restringido

Se ha estudiado el modelado mediante perceptrones multicapa, destacando la significativa importancia que adquieren la información sobre "la frase portadora" y el "signo de puntuación final del grupo fónico". El parámetro "número de frase portadora" introduce mejoras significativas; a pesar de que en las condiciones de grabación se intentó aislar el elemento variable de su frase portadora por medio de pausas obligatorias. El parámetro más importante, el "signo de puntuación final del grupo fónico", es muy relevante porque permite distinguir elementos variables con cadencias y elementos variables que generalmente presentan anticadencias en las grabaciones.

Se ha constatado la importancia de codificar la información enventanada sobre el acento de cada sílaba, así como su situación inicial o final en el grupo fónico. Se han ensayado varias codificaciones alternativas para la misma información sin conseguir superar la tasa. El tamaño óptimo de la ventana es dependiente de la tarea, aunque tiene relación con el tamaño de los grupos fónicos y los datos disponibles.

Se ha desarrollado y evaluado un nuevo método de diseño de bases de datos: por medio de un nuevo algoritmo voraz moderado por medio de subobjetivos parciales, se ha conseguido resumir una gran base de datos con una precisión superior al 95 %, de acuerdo con amplio espectro de vectores prosódico y segmentales.

Es igualmente importante estudiar cómo agrupar las frases en subdominios, (realizar una correcta agrupación de las grabaciones de acuerdo con su prosodia, proponiendo un modelado individual para algunas grabaciones), aunque las diferencias encontradas no han sido significativas.

Parámetros secundarios a la hora de modelar han resultado ser el tamaño del grupo fónico en sílabas o en palabras, la pertenencia de la sílaba a una palabra función o su situación en posición final de palabra. Apenas aportan mejoras; y si las aportan nunca es significativamente ni en todos los subdominios.

Se ha empleado una estrategia de experimentación no exhaustiva, que al ser comparada con la búsqueda exhaustiva del óptimo, ha mostrado su validez.

Análisis y síntesis de voz con emociones

Se han realizado experimentos para determinar la naturaleza segmental o prosódica de las emociones simuladas: en una evaluación pionera confirmada por posteriores experimentos de otros grupos de investigación, hemos mezclado los difonemas y la prosodia de diferentes emociones para concluir la naturaleza segmental del enfado amenazante del actor de nuestra base de datos y la naturaleza prosódica en el caso de la sorpresa; para la alegría y la tristeza se ha revelado como de naturaleza mixta, en parte segmental en parte prosódica.

Se ha creado un sistema completo de conversión texto a voz en castellano, con una nueva voz configurable con emociones: para ello se ha empleado síntesis basada en formantes en castellano, con capacidad de personalización evaluada con usuarios. Los parámetros de personalización han sido elegidos de manera que permitan implementar las emociones como un caso particular de personalización dinámica. Por lo que hemos podido ver, los resultados globales resultan prometedores, puesto que los humanos parecen adaptarse con rapidez a la voz sintética con emociones, y el periodo de

adaptación podría resultar breve y por lo tanto satisfactorio, especialmente en el caso de la tristeza (siendo peor para el enfado). Es cierto que hubo problemas de inteligibilidad en algunos contextos fonéticos, pero los resultados son perfectamente aceptables, incluso si la voz evaluada no resulta totalmente natural. De acuerdo con los resultados de evaluación, la voz sintética con emociones desarrollada en el proyecto VAESS es comparable al estado de la cuestión en otros idiomas a nivel mundial. Aunque en los planteamientos iniciales del proyecto VAESS se consideró que eran independientes los módulos prosódico y segmental del sintetizador, la conclusión de nuestro trabajo dista mucho de ser esta: debemos señalar que las trayectorias de los formantes pueden provocar, al incrementarse la velocidad de elocución, ruidos de naturaleza pseudo-oclusiva, que es necesario limar uno a uno, modificando la reglas segmentales.

Creación de la primera base de datos de habla emotiva simulada en castellano: Está orientada a síntesis prosódica y al análisis de la prosodia en párrafos y frase cortas mediante técnicas paramétricas, y dio lugar a un modelado diferencial de cada emoción respecto a la voz neutra y su evaluación en experimentos de copy-synthesis.

7. Aplicación práctica e interés industrial

A medida que las tecnologías del habla se han ido implicando, cada vez más, como parte integral de aplicaciones prácticas en escenarios reales, se ha hecho patente la necesidad de desarrollar sistemas de conversión texto a voz dotados de gran naturalidad.

Los resultados de esta Tesis han sido exitosamente aplicados en productos como:

- el robot del proyecto nacional Urbano/Ivanhoe (destinado a hacer de guía en el Museo de las Ciencias de Valencia),
- el sintetizador comercial multilingüe de la empresa Telia-Infovox (proyecto europeo VAESS),
- el sintetizador del GTH (registrado y comercializado por la UPM bajo el nombre de Boris, y vendido a empresas nacionales e internaciones como Digra o Ayllón),
- los sistemas de atención telefónica automática de la empresa Natural Vox (proyecto de mejora de su voz femenina).

Relación cronológica de las publicaciones del autor

ARTÍCULOS EN REVISTAS CON ÍNDICE DE IMPACTO EN JCR

1. *"Sesgos cognitivos en el reconocimiento de expresiones emocionales de voz sintética en la alexitimia"* (F. Martínez-Sánchez, J.M. Montero, J. de la Cerra) en *Psicothema 14(2), pp. 344-349 (ISSN: 0214-9915)* 2002 (**Impact Factor en "JCR Social Sciences Edition 2002": *1,098***)

2. "*Selection of the Most Significant Parameters for Duration Modeling in a Spanish Text-To-Speech System Using Neural Networks"* (R. De Córdoba, J.M. Montero, J. Gutiérrez-Arriola, J.A. Vallejo, E. Enríquez, J.M. Pardo) en *Computer Speech & Language Vol 16 Number 2 pp. 183-203 (ISSN: 0885-2308)* April 2002 (**Impact Factor en "JCR Science Edition 2003": *0,541***).

ARTÍCULOS EN OTRAS REVISTAS

3. "*Knowledge Combining Methodology for Dialogue Design in Spoken Language Systems*" (Rubén San-Segundo, Juan M. Montero, Javier Macías, Javier Ferreiros y José M. Pardo) en *International Journal of Speech Technology* Vol. 8(1) pp. 45-66 enero 2005 (ISSN: 1381-2416).

4. "*Medidas de confianza en sistemas de diálogo*" (R. San-Segundo, J. Macías, J.M. Montero, J. Ferreiros, R. Córdoba, J.M. Pardo) en *Procesamiento del Lenguaje Natural*, nº 33 pp. 95-102 (ISSN: 1135-5948) Sept. 2004.

5. "*Plataforma de generación semiautomática de sistemas de diálogo multimodales y multilingües: Proyecto GEMINI*" (L.F. D'Haro, R. Córdoba, I. Ibarz, R. San-Segundo, J.M. Montero, J. Macias-Guarasa, J. Ferreiros, J.M. Pardo) en *Procesamiento del Lenguaje Natural*, nº 33 pp. 103-110 (ISSN: 1135-5948) Sept. 2004

6. "*Sistema de comprensión de comunicaciones habladas para el control de tráfico aéreo del proyecto INVOCA*" (V. Sama Rojo, F. Fernández Martínez, J. Ferreiros López, J. Macias-Guarasa, R. De Córdoba, J. M. Montero Martínez, J. Colas Pasamontes, E. Campos Palarea, J. M. Pardo Muñoz) en *Procesamiento del Lenguaje Natural*, nº 31 pp.337-338 (ISSN: 1135-5948) Sept. 2003.

CAPÍTULOS DE LIBROS INTERNACIONALES

7. *"The role of pitch and tempo in Spanish emotional speech: towards concatenative synthesis" (*Juan Manuel Montero, Juana Gutiérrez-Arriola, Ricardo de Córdoba, Emilia Enríquez, José Manuel Pardo*) incluido en "Improvements in speech synthesis*" de los editores Eric Keller y Gerard Bailey, A. Monahan, J. Terken, M. Huckvale (ISBN 0-471-49985-4) pp. 246-251 editado por John Wiley & Sons, Ltd. en el año 2002.

8. *"Application of neural networks to duration modelling in a Spanish text-to-speech system" (*Ricardo de Córdoba, Juan Manuel Montero, José Manuel Pardo*) incluido en "Advances in Systems Engineering, Signal Processing and Communications*" pp. 244-247 editado por WSEAS Press (ISBN 960 8052 696). en el año 2002.

OTRAS PUBLICACIONES INTERNACIONALES REVISADAS

9. *"Parameter Selection for Prosodic Modeling in a Restricted-Domain Spanish Text-to-Speech System"* (J. M. Montero, J. Macias-Guarasa, R. De Cordoba, J. Gutierrez-Arriola, J. M. Pardo, R. San Segundo) en *Proceedings of World Automation Congress (WAC 2004)* pp. 155-160 (ISBN: 1-889335-23-1) Sevilla.

10. *"ANN F0 Modelling for Female-Voice Synthesis in Spanish: restricted and non-restricted domains"* (J.M. Montero, L.F. d´Haro, R. de Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, J.M. Pardo) en *Proceedings of the XVth International Congress of Phonetic Sciences* pp. 563-566. Agosto de 2003, Barcelona. (ISBN: 1-876346-48-5) editado en el año 2003.

11. "*ANESTTE: a writer's assistant for a specific purpose language"* (J.M Montero, M.M. Duque) en *University centre for Computer corpus REserarch on Language Technical Papers Volumen 16 - Special Issue: Proceedings of Corpus Linguistics 2003 Conference* de los editores Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. pp. 544-551. Marzo de 2003, Lancaster (ISBN: 1-86220-131-5) editado en el año 2003.