DOROTEO TORRE TOLEDANO es en la actualidad investigador post-doctoral asociado en el *Laboratory for Computer Science* del *Massachusetts Institute of Technology,* donde realiza su actividad investigadora en sistemas conversacionales hombre-máquina en el seno del grupo *Spoken Language Systems*. Doroteo es Ingeniero Superior de Telecomunicación (1997) y Doctor Ingeniero de Telecomunicación (2001) por la Universidad Politécnica de Madrid. Su historial académico se ha visto reconocido con numerosos premios, desde el primer Premio Nacional de Bachillerato (1990) hasta la obtención del Número Uno de la 70 promoción de Ingenieros Superiores de Telecomunicación de la Universidad Politécnica de Madrid (1998). Doroteo ha compaginado su actividad académica con su actividad profesional en Telefónica I+D (1994-2001), liderando por parte de dicha empresa diversos proyectos de investigación de ámbito internacional. Muchos de los resultados de su investigación están siendo aplicados en la actualidad en servicios vocales del grupo Telefónica.

RESUMEN DE LA TESIS DOCTORAL

SEGMENTACIÓN Y ETIQUETADO FONÉTICOS AUTOMÁTICOS

Un Enfoque Basado en Modelos Ocultos de Markov y Refinamiento Posterior de las Fronteras Fonéticas

FECHA DE LECTURA: 15 de Febrero de 2001
CALIFICACIÓN: Sobresaliente cum laude por unanimidad del tribunal

AUTOR: Doroteo Torre Toledano

DIRECTOR: Prof. Luis A. Hernández Gómez.
DEPARTAMENTO: Señales, Sistemas y Radiocomunicaciones, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid.

Esta Tesis se realizó en estrecha colaboración entre dicho departamento de la U.P.M. y Telefónica Investigación y Desarrollo.

1. Contexto

El objetivo de este primer apartado del resumen es situar esta Tesis Doctoral en el contexto científico-tecnológico en el que fue concebida y desarrollada.

Esta Tesis Doctoral se desarrolla en un campo de la Inteligencia Artificial que se conoce genéricamente como Tecnologías del Lenguaje Humano (*Human Language Technologies, HLT*). El objetivo genérico de este campo consiste en proveer a las máquinas de los conocimientos necesarios (en forma de algoritmos, modelos, etc.) para hacerlas capaces de comunicarse con los humanos de la forma más natural para los mismos, mediante el lenguaje oral.

Este objetivo ha sido un sueño para los científicos durante años. En la actualidad, sin embargo, los primeros frutos están empezando a cambiar nuestras vidas con multitud de aplicaciones ya prácticas y viables desde un punto de vista económico. El objetivo último de las Tecnologías del Lenguaje Humano está, sin embargo, todavía muy lejos de ser alcanzado: las máquinas están todavía muy lejos de acercarse a la capacidad de los humanos para el procesado del lenguaje oral: no pueden reconocer las palabras con la misma precisión que los humanos, no son capaces de procesar un lenguaje tan rico y variado como los humanos, ni tampoco de generar un lenguaje oral que resulte natural y espontáneo.

Dentro del campo genérico de las Tecnologías del Lenguaje Humano tienen cabida varias tecnologías que muchas veces tienen poco en común, salvo el objetivo último de procesar el lenguaje natural oral. Entre estas tecnologías podemos mencionar las siguientes:

- Reconocimiento de Voz, que se encarga de transformar la señal acústica del lenguaje oral producida por los humanos en una representación textual de lo pronunciado.

- Procesamiento del Lenguaje Natural, que se encarga de transformar una representación textual de una frase en lenguaje natural, en una representación de su contenido semántico apropiada para su procesamiento mediante computadoras, así como de la transformación inversa.

- Gestión del Diálogo, que se encarga de procesar representaciones semánticas obtenidas a partir de frases de un humano y de generar las representaciones semánticas de las respuestas adecuadas, haciendo uso para ello, si es necesario, de información contextual del diálogo.

- Síntesis de Voz, que se encarga de generar una señal acústica que trata de imitar la forma en que un humano pronunciaría una frase representada en forma textual.

- Identificación y Verificación del Locutor, que se encargan, respectivamente, de reconocer a una persona por su voz, y de verificar, mediante la voz de una persona, que esa persona es quien dice ser.

- Codificación de Voz, que se encarga de comprimir la información contenida en la señal acústica producida por los humanos, tratando de alcanzar los mayores ratios de compresión y las mínimas pérdidas.

2. Objetivos

El objetivo principal de esta Tesis es mejorar un aspecto concreto de una de las Tecnologías del Lenguaje Humano mencionadas en el apartado anterior, la síntesis de voz. Para llegar a comprender ese aspecto concreto así como la necesidad de mejorarlo resulta necesaria una breve explicación sobre síntesis de voz, que nos permitirá ya plantear los objetivos y la necesidad de esta Tesis.

Existen varias aproximaciones al problema de la síntesis de voz, la solución más directa y que requiere menos "inteligencia" por parte de las máquinas es la solución basada en mensajes (o fragmentos de mensajes) pregrabados, que se concatenan y reproducen para formar la señal acústica.

El problema de la concatenación directa es que se producen efectos de discontinuidad acústica y de entonación que casi siempre permiten al oyente detectar los puntos de unión, y que a veces pueden resultar molestos. Estos efectos de discontinuidad se pueden reducir dotando al sistema de algo más de inteligencia. Por ejemplo, utilizando técnicas de procesado de señal para suavizar las uniones, o utilizando técnicas de reconocimiento de patrones para encontrar los fragmentos que mejor encajan entre sí (en caso de que se disponga de múltiples combinaciones de fragmentos de voz para generar la voz correspondiente a la frase en cuestión).

En cualquier caso, la solución anterior tiene el inconveniente de que sólo es capaz de reproducir los textos que se pueden representar como concatenación de los fragmentos de voz pregrabada disponibles. Esto hace necesario generar un inventario de fragmentos de voz para cada aplicación a desarrollar. Además, existen aplicaciones para las que es necesario generar mensajes vocales para textos sin ningún tipo de restricción (por ejemplo, un sistema de acceso vocal al correo electrónico, un sistema de lectura de noticias, etc.). Para resolver este problema se lleva al extremo la técnica de la concatenación de fragmentos de voz, haciendo que los fragmentos de voz no correspondan siquiera con palabras, sino con pequeñas agrupaciones de fonemas a partir de las cuales es posible componer cualquier palabra del idioma. Esta solución tiene el inconveniente de que la calidad de la voz generada es, usualmente, inferior debido al alto grado de concatenación, y a que la voz generada a partir de fragmentos de voz tan pequeños tiene una entonación totalmente plana, lo que hace necesario modelar la entonación (caracterizada principalmente por la energía, frecuencia fundamental y duración de los fonemas) de forma separada. Debido a la capacidad de los sistemas de síntesis así diseñados de producir voz sintética para un texto general, se les suelen denominar sistemas de conversión Texto a Voz (Text-To-Speech systems, TTS).

Aunque existen otras formas de síntesis de voz que no hacen uso de voz pregrabada (como por ejemplo, la síntesis por formantes), la forma de síntesis más extendida en la actualidad es la síntesis por concatenación de unidades, ya sean éstas fragmentos de frases incluyendo varias palabras, o pequeñas secuencias de fonemas como en los sistemas de conversión Texto a Voz. En este último caso se hace necesario también un modelo de entonación para generar voz con un grado de naturalidad aceptable.

Para la generación de sistemas de síntesis de voz por concatenación es necesario, obtener los fragmentos de interés a partir de un conjunto de pronunciaciones, cortando (segmentando) la señal de voz por los puntos adecuados, e identificando los contenidos de cada uno de los segmentos obtenidos (es decir, etiquetando los

segmentos). Aunque el tipo de fragmentos de voz a obtener viene definido por el conjunto (o inventario) de unidades de síntesis a generar, suele ser preferible realizar una segmentación y un etiquetado fonéticos, ya que a partir de esta segmentación y este etiquetado es relativamente sencillo generar la segmentación y el etiquetado para cualquier inventario de unidades. La precisión en la segmentación de la señal de voz resulta un factor esencial para una síntesis de voz de calidad, ya que un pequeño error en la segmentación (incluso tan pequeño como 5 ms.) puede dar lugar a un error audible en la voz sintética (piénsese que si, por ejemplo, un fragmento que supuestamente modela una transición del fonema /a/ al fonema /s/ incluye también 5 milisegundos del fonema siguiente al fonema /s/, que resulta ser un fonema /o/, cada vez que se utilice ese fragmento para la síntesis se oirá un chasquido en el fonema /s/). Además, para los sistemas de conversión Texto a Voz, es necesario entrenar modelos de entonación, midiendo la energía, frecuencia fundamental y duración de los fonemas en un elevado número de frases, para lo cual también es necesaria una muy elevada precisión en la segmentación y el etiquetado fonéticos.

Debido a estos requerimientos de precisión en la segmentación y el etiquetado, tanto para obtener un inventario de unidades como para entrenar modelos prosódicos, en síntesis de voz es usual recurrir a una segmentación y etiquetado manuales. Este procedimiento es muy costoso en tiempo y esfuerzo (piénsese que 20 frases de tamaño medio pueden fácilmente contener más de 600 fonemas y fronteras fonéticas que deben ser identificados y localizados de forma individual, y se necesitan muchas más frases para construir un sistema de síntesis). De todos modos, el coste y el tiempo necesarios son asumibles en tanto se mantengan los inventarios de unidades de síntesis limitados y sólo se produzca voz sintética imitando la voz de un único locutor humano.

El problema es que la competencia tecnológica exige, por un lado, aumentar la calidad de la voz sintética, lo cual se puede conseguir fácilmente aumentando el inventario de unidades de síntesis (y eligiendo las unidades sabiamente); y por otro, cubrir con un mismo sistema diferentes lenguajes, disponer de voces variadas para un mismo lenguaje, e incluso desarrollar voces personalizadas (de forma que, por ejemplo, una compañía pueda disponer de un sistema de síntesis con su propia voz corporativa), todo lo cual pasa por la posibilidad de desarrollar de forma rápida y económica nuevas voces.

La segmentación y el etiquetado manuales no pueden responder a estas nuevas demandas, por ello es necesario buscar procedimientos alternativos, tan automáticos como sea posible para generar una segmentación y un etiquetado fonéticos con una precisión lo más próxima a la precisión de la segmentación y el etiquetado manuales. Éste es, sin más, el objetivo principal de esta Tesis Doctoral.

3. Resumen

El problema que tratamos de resolver en esta Tesis Doctoral es el problema de la segmentación y el etiquetado fonéticos automáticos. Este problema se puede definir de forma muy resumida como el problema de obtener, a partir de la señal de voz y de la transcripción ortográfica de la misma, la secuencia de fonemas pronunciada, así como las posiciones de las fronteras entre los mismos.

Nótese que consideramos que la transcripción ortográfica (secuencia de palabras) es conocida *a priori*. Ésta es una suposición de orden práctico, ya que, por un lado, realizar una segmentación y un etiquetado fonético a partir de la voz exclusivamente es un problema todavía demasiado complejo como para ser realizado de forma automática con una precisión aceptable, y por otro, es muy frecuente disponer de la transcripción ortográfica de las pronunciaciones (y caso de no ser así, es relativamente poco costoso generarlas).

El análisis del Estado del Arte en segmentación y etiquetado fonéticos automáticos lleva a la conclusión de que existen dos aproximaciones principales al problema:

- La aproximación más ampliamente utilizada consiste en **reutilizar técnicas y características de reconocimiento de voz**. La ventaja principal de este método es que se beneficia de la madurez de las técnicas de reconocimiento de voz, así como de la amplia infraestructura de que se dispone. Este método supone normalmente realizar pequeñas modificaciones a un reconocedor de voz (que normalmente utilizan como características Mel-Frequency Cepstral Coefficients, MFCCs), y como modelos Modelos Ocultos de Markov (*Hidden Markov Models, HMMs*). Las modificaciones van dirigidas a que el reconocedor haga uso de la transcripción ortográfica de la frase a segmentar para restringir la búsqueda de posibles transcripciones fonéticas y de posibles posiciones para las fronteras fonéticas. También es necesario hacer que el reconocedor produzca la información de las posiciones de las fronteras fonéticas (un subproducto de los reconocedores de voz que normalmente se descarta). Existen algunas adaptaciones adicionales, ampliamente aceptadas, que no vamos a detallar aquí. Esta aproximación tiende a dar buenos resultados de etiquetado fonético automático y también de segmentación, en lo que a errores de bulto ser refiere, pero, debido por un lado a la poca resolución temporal de las características utilizadas y a que los Modelos Ocultos de Markov no permiten modelar de forma muy detallada las transiciones, la precisión que alcanzan en la segmentación fonética no es muy alta (en otras palabras, producen muchos errores pequeños de segmentación, significando aquí pequeño "menor que la duración de un fonema").

- Debido a la poca precisión obtenida en la segmentación con técnicas de reconocimiento de voz, otros autores han experimentado con **otras características** (contorno de energía, energía en distintas bandas de frecuencia, tasa de cruces por cero, funciones de variación espectral, etc.) habitualmente con más resolución temporal, así como con **otras técnicas** (pre-segmentaciones en segmentos estables seguidos de alineación con fonemas, detectores de transiciones basados en reglas, redes neuronales, alineamiento de la voz a segmentar y una versión sintética de la frase, etc.) más orientadas a la detección de transiciones. En general estas características y técnicas conducen a una inferior precisión en el etiquetado

fonético, así como a un mayor número de errores de bulto de segmentación, pero pueden llegar a situar fronteras fonéticas de forma muy precisa cuando no se producen errores de bulto.

En esta Tesis se trata de fusionar estas dos aproximaciones al problema de la segmentación y el etiquetado fonéticos automáticos, proponiendo un novedoso enfoque basado en dos etapas (Figura 1). En primer lugar, se utilizan técnicas de reconocimiento de voz para producir el etiquetado fonético y una segmentación fonética intermedia con pocos errores de bulto pero poco precisa. En segundo lugar, se refina localmente la segmentación fonética intermedia aumentando la precisión de la misma sin por ello aumentar el número de errores de bulto. Para ello se hace uso de otras técnicas y características similares a las que utilizan los autores que no reutilizan técnicas y características de reconocimiento de voz.

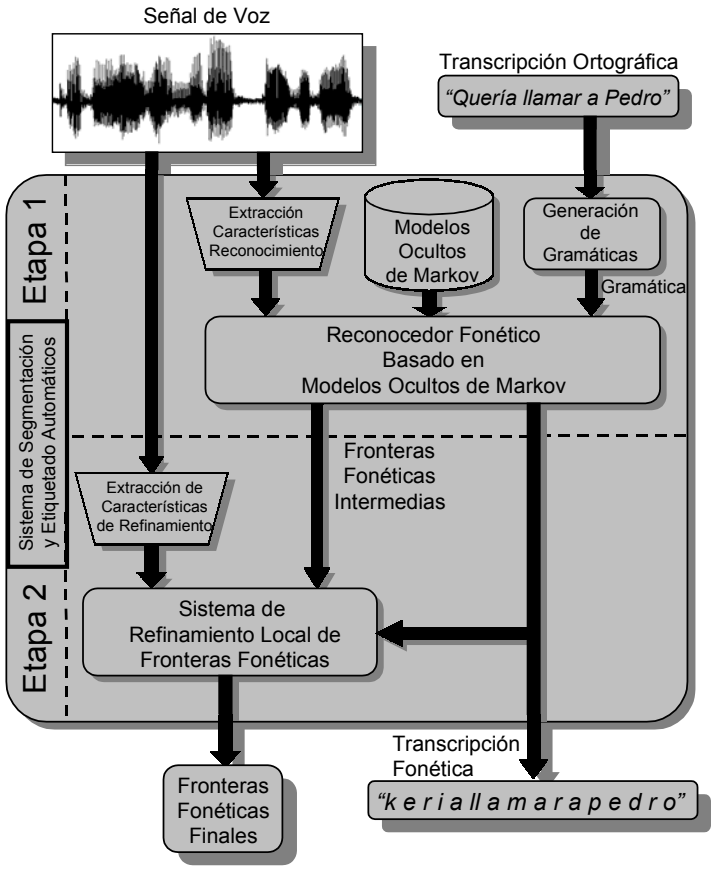

Figura 1: Enfoque propuesto en esta Tesis para afrontar el problema de la segmentación y el etiquetado fonéticos automáticos.

En esta Tesis se ha utilizado este enfoque como guía para analizar a un nivel de detalle con el que no se había hecho anteriormente, en primer lugar el problema de la segmentación y el etiquetado fonéticos utilizando técnicas de reconocimiento de voz (HMMs y MFCCs), y en segundo lugar el problema del refinamiento de esa segmentación.

En cuanto al análisis de la reutilización de técnicas de reconocimiento de voz, se ha comenzado comparando el funcionamiento de los HMMs dependientes e independientes del contexto. Los HMMs dependientes del contexto tienen más capacidad para modelar los efectos de coarticulación en las transiciones espectrales, y por ello consiguen una precisión mucho mayor en reconocimiento de voz. En segmentación fonética automática, en cambio, la mayor parte de los autores utilizan HMMs independientes del contexto, ya que en numerosos experimentos se ha encontrado que obtienen una mayor precisión en la segmentación. En esta tesis hemos encontrado estos mismos resultados, pero los hemos detallado un poco más. En concreto, hemos encontrado que, dependiendo del umbral de error considerado (por debajo del cual se considera que no hay error de segmentación), se obtienen mejores resultados con los HMMs dependientes del contexto o independientes del contexto. Para los umbrales de error más usados en la literatura, también en esta Tesis se ha encontrado que los HMMs independientes del contexto producen mejores resultados. En cuanto a las razones para este paradójico comportamiento, un investigador había argumentado que se debe a que los HMMs dependientes del contexto pierden la correspondencia con los fonemas que tratan de modelar, en otras palabras, que el HMM en realidad modela el fonema y parte de adyacentes o sólo una parte del fonema. En esta Tesis se ha planteado por primera vez la hipótesis de que, de ser correcta esta explicación, los errores producidos por esta causa deberían ser errores sistemáticos, y por tanto modelables estadísticamente y parcialmente cancelables. No sólo se ha propuesto esta hipótesis, sino que se ha llevado a la práctica, con un nuevo sistema de corrección estadística de los errores sistemáticos de segmentación producidos por HMMs dependientes del contexto. Este sistema ha demostrado ser capaz de incrementar enormemente la precisión de la segmentación obtenida con los HMMs dependientes del contexto, tal como se muestra en la Figura 2, hasta el punto de superar ampliamente la precisión obtenida con los HMMs independientes del contexto para cualquier umbral de error considerado. Estos resultados constituyen una demostración empírica de la hipótesis planteada como explicación de la pobre precisión de los HMMs dependientes del contexto en tareas de segmentación. A la vez constituye un nuevo mecanismo para incrementar la precisión de la segmentación fonética.

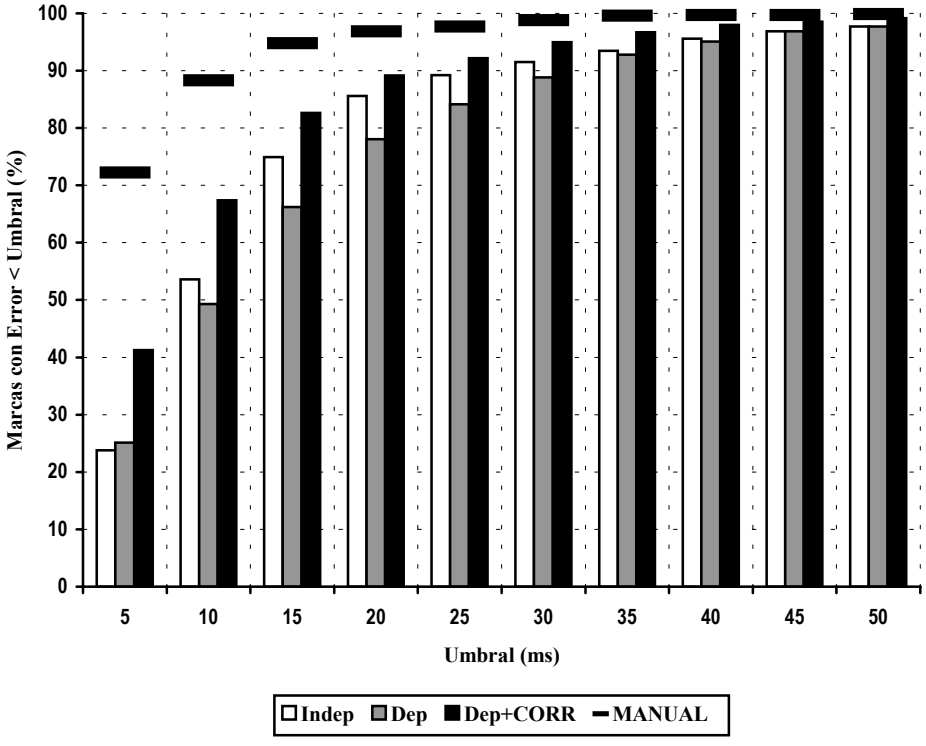

Figura 2: Comparación de los resultados de segmentación obtenidos con HMMs independientes del contexto, dependientes del contexto, y dependientes del contexto con corrección de errores sistemáticos.

Posteriormente se ha analizado el comportamiento de las técnicas de adaptación al locutor para segmentación fonética. Las técnicas de adaptación al locutor son técnicas bastante maduras capaces de incrementar la precisión del reconocimiento de voz. Sin embargo, todavía no estaba muy claro si eran capaces de mejorar la segmentación automática de voz, y en qué medida. En esta Tesis hemos encontrado que las técnicas de adaptación al locutor más ampliamente utilizadas en reconocimiento de voz son, en efecto, capaces de aumentar la precisión de la segmentación automática de voz, pero en mucha menor medida que la nueva técnica de corrección de los errores sistemáticos de segmentación producidos por los HMMs dependientes del contexto introducida en esta Tesis. De todos modos, hemos comprobado que los tipos de error que es capaz de corregir la adaptación al locutor y nuestra nueva técnica son distintos. Nuestra técnica corrige fundamentalmente pequeños errores de segmentación, mientras que la adaptación al locutor corrige fundamentalmente errores de bulto. De hecho, hemos comprobado empíricamente que las mejoras obtenidas por separado con adaptación al locutor y con nuestra técnica de corrección de errores sistemáticos de segmentación prácticamente se suman cuando se aplican las dos técnicas de forma conjunta. Por ello, se ha decidido aplicar estas dos técnicas de forma conjunta utilizando HMMs dependientes del contexto en la primera etapa de segmentación que reutiliza técnicas de reconocimiento de voz (ver Figura 1).

En cuanto al etiquetado fonético, hemos comprobado que es más preciso cuando se permiten varias transcripciones alternativas por cada palabra, permitiendo a los modelos acústicos decidir qué sonidos son los que en realidad se han producido para pronunciar cada palabra. La detección de silencios entre palabras presenta una problemática distinta, que también ha sido analizada. Se han permitido silencios opcionales entre palabras, para dejar a los modelos acústicos la última decisión. Sin embargo, se ha observado que los modelos acústicos tendían a introducir demasiados silencios entre palabras. Como muchos de los silencios que se introducían eran de duraciones anormalmente cortas se ha utilizado un filtrado de los silencios intermedios basado en su duración. Con esto se obtenían las mayores precisiones de etiquetado fonético reportadas en la literatura analizada (por encima del 99% de precisión en el etiquetado de fonemas, y por encima del 92% de precisión en el etiquetado de silencios entre palabras).

Con todo, las mejoras obtenidas en esta Tesis en cuanto a la utilización de técnicas de reconocimiento de voz para el etiquetado y la segmentación fonéticos de voz quedan resumidas en la Figura 3. Como se puede observar, las mejoras conseguidas son importantes, pero la precisión de la segmentación está todavía muy alejada de la precisión de una segmentación manual, especialmente cuando se toman en consideración los pequeños errores de segmentación, errores que resultan ser muy importantes para la generación de inventarios de unidades y modelos para síntesis de voz, el objetivo primordial de esta Tesis. Para conseguir reducir el número de errores pequeños de segmentación, sin aumentar el número de errores de bulto de segmentación, se introduce la segunda etapa (de refinamiento local de la segmentación) en el esquema de la Figura 1.

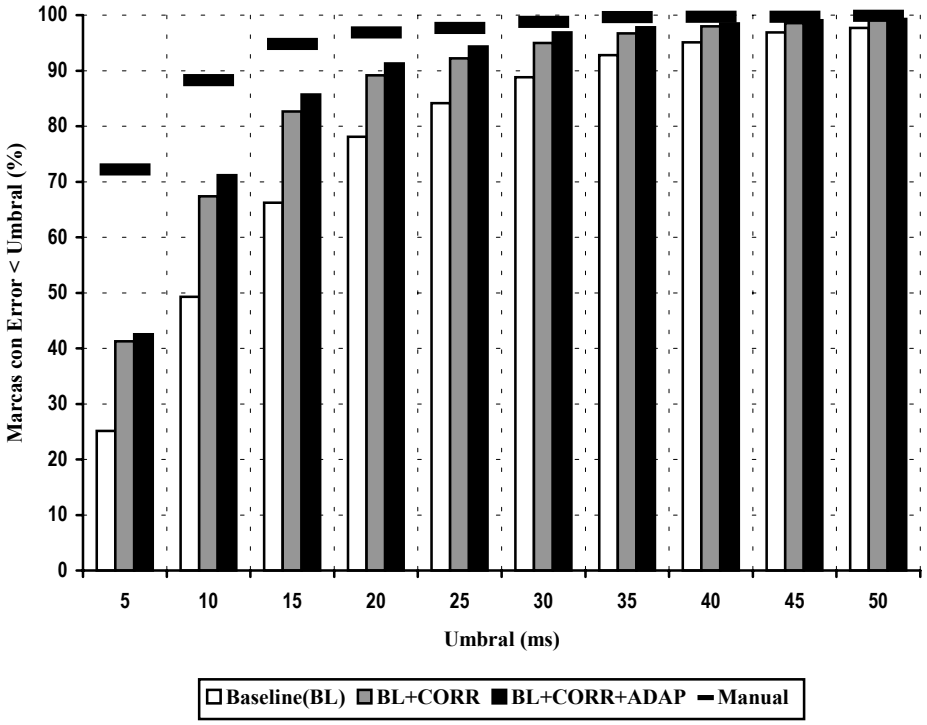

Figura 3: Mejora obtenida al aplicar la nueva técnica de corrección estadística de errores sistemáticos de segmentación (CORR), y esta técnica más adaptación al locutor (CORR+ADAP).

Para la etapa de refinamiento local esta Tesis (ver Figura 1) ha introducido un marco genérico novedoso en el que tiene cabida la utilización de distintos vectores de características de refinamiento, así como de distintas técnicas de modelado. El

marco genérico define cómo se toma en consideración la información aportada por la primera etapa (la basada en técnicas de reconocimiento) y cómo se combina, a través de las distintas técnicas de modelado, con la información aportada por el vector de características de refinamiento.

Utilizando ese novedoso marco genérico de refinamiento local se ha experimentado con tres vectores de características distintos y tres tipos de técnicas de modelado: sistemas de reglas borrosas (fuzzy logic rules), redes neuronales, y modelos de mezclas de gaussianas. Una vez optimizados los modelos para cada una de las técnicas empleadas, los resultados del refinamiento local quedan recogidos en la Figura 4. Como puede apreciarse, el refinamiento local consigue incrementar enormemente la precisión de la segmentación, sin por ello aumentar el número de errores de bulto de segmentación. Se puede comprobar además, que los resultados de segmentación obtenidos tras el refinamiento, particularmente con redes neuronales, son prácticamente idénticos a los resultados de una segmentación fonética manual (excepto para errores muy pequeños, de tan sólo 5 ms.). No tenemos conocimiento hasta el momento de ningún trabajo de investigación que presente resultados de segmentación fonética automática superiores a los obtenidos en esta Tesis.

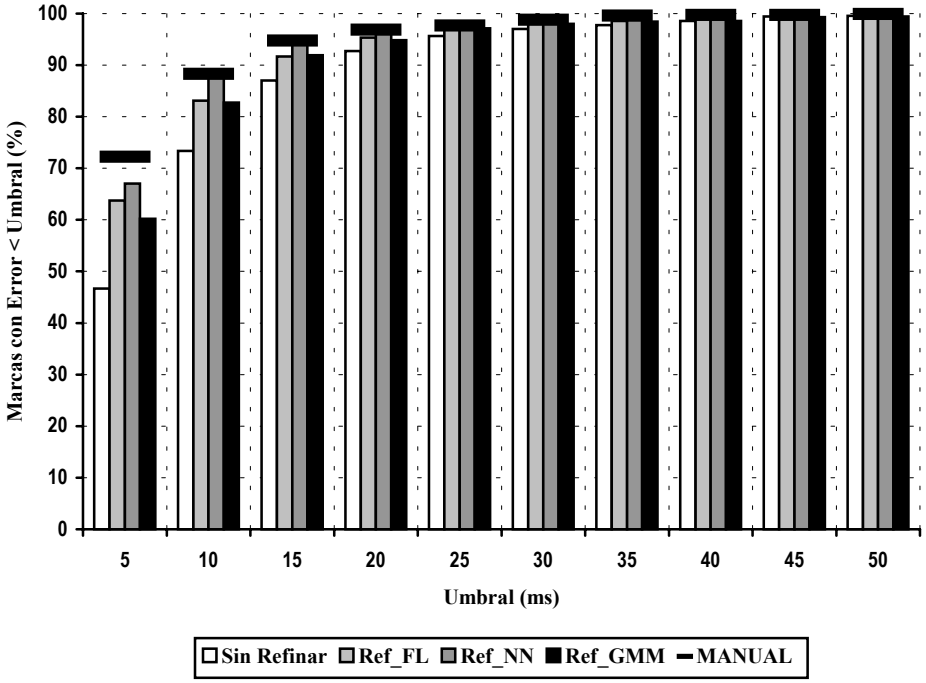

Figura 4: Mejora obtenida sobre los resultados de la primera etapa de segmentación basada en técnicas de reconocimiento (Sin Refinar), al aplicar refinamiento local con reglas borrosas (Ref_FL), con redes neuronales (Ref_NN) y con modelos de mezcla de gaussianas (Ref_GMM).

4. Aplicación Industrial y Otras Aplicaciones

Los resultados obtenidos en esta Tesis han superado incluso nuestras mejores expectativas, consiguiendo aproximarse claramente a los resultados obtenidos con una segmentación y un etiquetado fonéticos manuales. Comparando la Figura 2 y la Figura 4 se aprecia la mejora obtenida en la precisión de la segmentación. Esta mejora ha sido en su mayor parte debida a la introducción de dos técnicas novedosas: la corrección estadística de los errores sistemáticos de segmentación de los HMMs dependientes del contexto, y el refinamiento de las fronteras fonéticas utilizando un nuevo marco genérico de refinamiento local y redes neuronales.

El sistema resultante de esta Tesis (o versiones anteriores del mismo) está en explotación industrial desde 1998 en Telefónica Investigación y Desarrollo, que lo usa desde entonces para la creación de todas las nuevas voces sintéticas para el Conversor Texto-Voz Multilingüe de Telefónica. La mejora conseguida en la precisión de la segmentación fonética (así como en el etiquetado fonético) permite emplear segmentación automática para generar una nueva voz para un sintetizador (incluyendo inventario de unidades de síntesis y modelos de prosodia) de forma totalmente automática. La versión completamente automática de la voz sintética (que se puede conseguir en tan solo una semana) no es completamente perfecta, y suele contener errores audibles. Telefónica Investigación y Desarrollo ha estimado que corregir esos errores requiere alrededor de una persona-mes, mientras que la creación de una nueva voz sintética de forma totalmente manual requería alrededor de una persona-año, para obtener una calidad similar de la voz sintética.

Aunque el sistema resultante de esta Tesis ha sido desarrollado para el castellano, en Telefónica Investigación y Desarrollo se ha aplicado también con éxito para el desarrollo de voces sintéticas hispanoamericanas, e incluso de voces sintéticas en varios idiomas del Estado Español (catalán, gallego y euskara), utilizando para ello un mapeado de los fonemas del idioma considerado sobre los del castellano.

Debemos mencionar también que las aplicaciones del sistema resultante de esta Tesis no se limitan únicamente al principal objetivo de esta Tesis, la generación semiautomática de voces sintéticas, sino que incluyen también la segmentación y el etiquetado fonéticos automáticos de corpus de voz. En este sentido, Telefónica ha firmado un acuerdo de colaboración con la Real Academia Española (RAE) que, entre otras cosas, establece que Telefónica colaborará con la RAE en la anotación del Corpus de Referencia del Español Actual (CREA) que incluye grabaciones de voz que serán segmentadas y etiquetadas fonéticamente. Para esta tarea se está considerando utilizar el sistema de segmentación y etiquetado fonéticos automáticos presentado en esta Tesis Doctoral, al menos para producir una primera segmentación y etiquetado fonéticos, que posteriormente se refinará manualmente.

Por último, hay autores que tratan de emplear técnicas de segmentación fonética para mejorar los resultados del reconocimiento de voz. Algunos autores han comprobado que el empleo de voz segmentada y etiquetada fonéticamente para entrenar reconocedores de voz hace que estos obtengan mejores resultados que los entrenados con voz sin segmentar ni etiquetar fonéticamente (lo más normal en la actualidad). También han mostrado que introduciendo en los algoritmos de reconocimiento información sobre las fronteras fonéticas se pueden mejorar las tasas de reconocimiento. Uno de los centros de investigación internacionales donde más hincapié se hace en la utilización de segmentación en el reconocimiento de voz es el Spoken Language Systems Group del Laboratory for Computer Science del M.I.T., por ello elegí este centro para continuar mi labor investigadora en esa línea.

5. Publicaciones

- Capítulos de libros:

  - Toledano DT, Rodríguez MA, Escalada JG and Hernández LA, *Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-Correction Rules*, to be published in Van Santen JPH et al. (eds.) *Progress in Speech Synthesis*.

- Artículos en revistas internacionales:

  - Toledano DT and Hernández LA, *Automatic Phonetic Segmentation*, Submitted for publication to the IEEE Transactions on Speech and Audio Processing.

- Artículos en revistas nacionales:

  - Rodríguez MA, Escalada JG and Toledano DT, *Conversor Texto-Voz Multilingüe para Español, Catalán, Gallego y Euskera*, In Procesamiento del Lenguaje Natural, September 1998, pp 16-23.

- Artículos en congresos internacionales:

  - Toledano DT and Hernández LA, *Local Refinement of Phonetic Boundaries: A General Framework and its Application Using Different Transition Models*, In Proceedings EUROSPEECH 2001, Aalborg (Denmark), 2-7 September 2001.

  - Toledano DT, *Neural Network Boundary Refining for Automatic Speech Segmentation*, In Proceedings of the International Conference on Acoustics Speech and Signal Processing 2000, Istambul (Turkey), June 2000.

  - Toledano DT, Rodríguez MA and Escalada JG, *Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-Correction Rules*, In Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 1998 Jenolan Caves (Australia), pp. 207-212.

  - Toledano DT and Hernández LA, *HMMs for Automatic Phonetic Segmentation*, Submitted for publication to the 3rd Language Resources Conference (LREC2002).

  - Toledano DT, Villarubia L, Hernández LA and Elvira JM, *Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers*, In Proceedings of the Interational Conference on Acoustics Speech and Signal Processing 1997, pp. 1463-1466.

6. Otras Publicaciones

- M.I.T. Lab. for Computer Science Annual Research Summary, 2002.

    - Toledano DT, Vargas A, Oliver A, Polifroni J, Hazen TJ, Seneff S, *EL TIEMPO: A Conversational System in Spanish for Accessing Weather Information over the Phone*. To be published in M.I.T. Lab. for Computer Science Annual Research Summary, 2002.

    - Seneff S, Cowan B, Polifroni J, Toledano DT, Wang C., *Multilingual Phrasebook*, To be published in M.I.T. Lab for Computer Science Annual Research Summary, 2002.

- Proyecto Europeo SIRIDUS (Specification, Interaction and Reconfiguration in Dialogue Understanding Systems)

    - Toledano DT, Gabriel J, Quesada JF, García C, *User requirements on a Natural Command Language Dialogue System*, Proyecto Europeo SIDIDUS (IST-1999-10516) D3.1, September 2000.

    - Quesada JF, Toledano DT, Gabriel Amores J, *Design of a Natural Command Language Dialogue System*, Proyecto Europeo SIRIDUS (IST-1999-10516) D3.2, December 2000.

    - Berman A, Cooper R, Ericsson S, Hieronymus J, Jonson R, Larsson S, Milward D, Toledano DT, *Implemented SIRIDUS System Architecture (Baseline)*, Proyecto Europeo SIRIDUS (IST-1999-10516) D6.2, December 2000.

    - Quesada JF, Gabriel J, Toledano DT, Tapias D, *Installation of Current Trindi Software at Telefónica and the University of Seville*, Proyecto Europeo SIRIDUS (IST-1999-10516) D7.1, July 2000.

7. Otros Méritos

- Desde Abril de 2001 soy un Postdoctoral Research Associate en el Spoken Language Systems Group del Laboratory for Computer Science del M.I.T., donde continúo mi investigación sobre las aplicaciones en reconocimiento de voz de las técnicas de segmentación fonética desarrolladas en mi Tesis Doctoral, así como mi investigación en diálogo automático hombre-máquina, gracias a una beca postdoctoral concedida por el M.I.T.

- Seminario "Automatic Phonetic Segmentation and Labelling", impartido en el Laboratory for Computer Science del M.I.T., el 15 de Septiembre de 2000.

- Seminario "Barge-in in Dialogue Systems", impartido en el Laboratory for Computer Science del M.I.T., el 17 de Octubre de 2001.

- Número uno de la 70 promoción (1997) de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid.

- Premio al Mejor Expediente Académico hasta 3° de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid.

- Primer Premio Nacional de Bachillerato.

- Colaboración continuada con la división de Tecnología del Habla de Telefónica I+D, primero como becario (1995-97) y posteriormente contratado (1997-2001), y llegando a dirigir la participación de Telefónica I+D en algunos de los proyectos internacionales más avanzados en el campo de las Tecnologías del Lenguaje Humano:

  - Proyecto Europeo SIRIDUS (Specification, Interaction and Reconfiguration in Dialogue Understanding Systems), que trata de tomar un motor de gestión del diálogo en modo texto, resultado del Proyecto Europeo TRINDI, y crear una arquitectura software que permita utilizarlo como el núcleo central de un sistema de diálogo vocal, tratando de que esta arquitectura contemple los retos de investigación más importantes en el área de los sistemas automáticos de diálogo, y generar así una herramienta muy útil para los investigadores en el campo.

  - Proyecto I3S (Intuitive Interfaces to Information Systems), promovido por la compañía MCC de Texas y con la participación de Texas Instruments, Nortel Networks y algunas de las compañías telefónicas de Estados Unidos, que trataba de partir del sistema de diálogo TRAINS de la Universidad de Rochester y extenderlo aumentando sus capacidades de portabilidad a otros dominios.

  Participación en varios proyectos de Telefónica I+D para el desarrollo de servicios reales avanzados haciendo uso de Tecnologías del Lenguaje Humano, entre ellos la creación de una agenda de marcación vocal modificable por voz, haciendo uso de resultados de mi proyecto fin de carrera.