CURRICULUM RUBÉN SAN SEGUNDO:

"Rubén San Segundo es ingeniero de Telecomunicación por la Universidad Politécnica de Madrid (1997) y Doctor Ingeniero por esta misma universidad (2002). Ha realizado su tesis relacionada con el diseño de sistemas de diálogo por teléfono con reconocimiento y síntesis de voz.
Rubén ha realizado dos estancias de investigación en el centro CSLR (Center of Spoken Language Research) de la Universidad de Colorado.
Desde septiembre de 2001 hasta febrero de 2003, Rubén ha trabajado en la División de Tecnología del Habla de Telefónica I+D donde ha aportado gran cantidad de la investigación realizada en su tesis doctoral.
Desde febrero de 2003 hasta la actualidad, Rubén es profesor de la Universidad Politécnica de Madrid: inicialmente en el departamento de Sistemas Electrónicos y Control (EUITT) y actualmente en el departamento de Ingeniería Electrónica de la ETSIT, dentro del Grupo de Tecnología del Habla."

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR

DE INGENIEROS DE TELECOMUNICACIÓN

TESIS DOCTORAL

MEJORA DE SERVICIOS AUTOMÁTICOS POR TELÉFONO CON RECONOCIMIENTO DE HABLA: NUEVA GENERACIÓN DE SERVIDORES VOCALES INTERACTIVOS

RESUMEN

Autor

Rubén San Segundo Hernández

Director de Tesis

Dr. Ingeniero José Manuel Pardo Muñoz

2002

Índice de contenidos

1. INTRODUCCIÓN

En los últimos años, las Tecnologías del Habla han constituido un campo importante de investigación. En la actualidad, estas tecnologías están pasando de ser un objetivo meramente científico a ser un objetivo comercial. Esta tendencia se ha puesto de manifiesto en las importantes inversiones que se están haciendo en este sector por parte de las grandes empresas de telecomunicaciones. En este cambio hacia la comercialización de estas tecnologías, tienen un protagonismo relevante los *Servidores Vocales Interactivos (SVIs).* Un SVI no es más que un sistema capaz de proporcionar un servicio de adquisición y/o difusión de información a través de la línea telefónica, utilizando síntesis y reconocimiento de voz. Para ofrecer este servicio, el sistema entabla un diálogo con el usuario que finalmente lleva a éste a conseguir la información solicitada o a realizar las operaciones deseadas. Tradicionalmente, las empresas han venido dando este tipo de servicios a través de operadores humanos que atendían personalmente las llamadas. La automatización que se puede conseguir en gran parte de estos servicios mediante la introducción de las Tecnologías del Habla, y la reducción de costes asociada, está despertando un gran interés.

En esta tesis se ha realizado la mayor cantidad de trabajo en los módulos resaltados en la figura 1: en el reconocimiento de habla, la obtención de medidas de confianza y en la gestión del diálogo. En el reconocimiento de habla se pretende el análisis del fenómeno de deletreo en castellano y se describe el diseño e implementación de un sistema de reconocimiento de nombres deletreados completo. En este módulo también se detalla la implementación de un sistema de reconocimiento de habla continua para dominios restringidos con vocabularios medios (400 palabras). El dominio elegido ha sido el de fechas y horas. En cuanto a las medidas de confianza, se presentarán los resultados del trabajo realizado sobre el sistema CU Communicator. Por otro lado, también se han realizado análisis de medidas de confianza sobre los reconocedores desarrollados en la presente tesis con el fin de aplicarlas en la gestión del diálogo. En cuanto al gestor de diálogo, se propone y describe una metodología para su diseño en SVIs. Dicha metodología, ha sido aplicada para el desarrollo del gestor de diálogo en un servicio de información y reserva de billetes de tren.

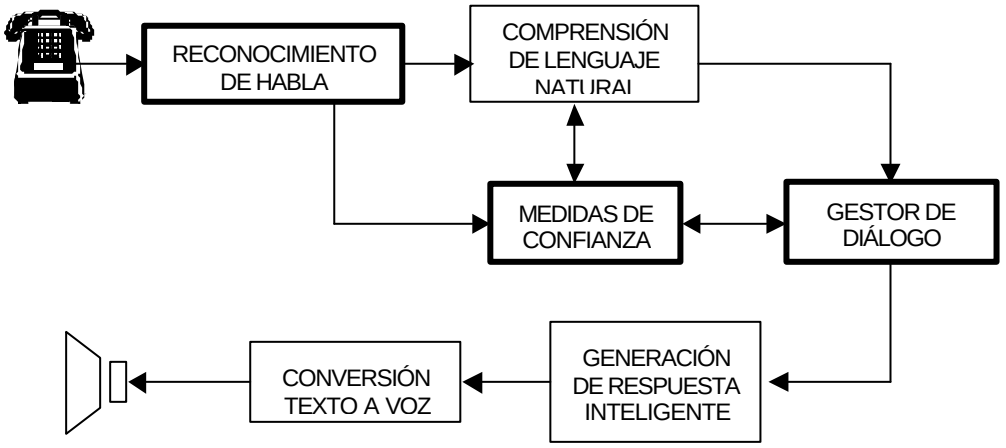

Figura 1: *Diagrama de bloques de un Servidor Vocal Interactivo.*

En esta tesis trabajaremos fundamentalmente con tres sistemas diferentes: un sistema de información y reserva de billetes de avión, hotel y coches de alquiler en inglés, un servicio de reserva de billetes de tren y otro servicio de información de números de teléfono ambos en español. En la figura 1, se muestra un esquema modular genérico de un SVI.

2. SISTEMA DE RECONOCIMIENTO DE NOMBRES DELETREADOS

2.1 Tarea de deletreo en castellano

El funcionamiento de un sistema de reconocimiento no sólo depende del tamaño o perplejidad del vocabulario de reconocimiento sino también del grado de similitud existente entre las palabras que forman dicho vocabulario. En la tabla 1 presentamos las transcripciones de las pronunciaciones estándar en castellano de las letras. Para ello hemos utilizado el alfabeto fonético internacional (IPA: International Phonetic Alphabet). En inglés, la mayor dificultad de esta tarea reside en el reconocimiento del conjunto de letras denominado E-set = {B, C, D, E, G, P, T, V, Z}. Analizando la tabla 1 podemos identificar el conjunto E-set para el caso del castellano = {B, C, CH, D, E, G, P, T}. En este conjunto, los problemas de confusión son muy parecidos a los existentes en inglés: las transcripciones de sus pronunciaciones tienen una estructura muy similar formada por una consonante y el fonema **e.**

La única diferencia reside en la consonante que acompaña a la vocal. En castellano debemos considerar otro conjunto de letras también de gran confusión que denominaremos ExE-set = {F, L, LL, M, N, Ñ, R, S}. En este conjunto, las transcripciones de las letras forman también la misma estructura fonética: **'e _ e**. Estas letras tienen únicamente un fonema diferente (el fonema central), por lo que las diferencias acústicas entre dichas letras son también muy pequeñas.

Transcripciones de las letras en castellano (IPA)									
A	a	**F**	'e f e	**L**	'e l e	**P**	p e	**V**	'u b e
B	b e	**G**	g e	**LL**	'e ʔ e	**Q**	k u	**W**	u b e 'd o b l e
C	θ e	**H**	'a t ʔ e	**M**	'e m e	**R**	'e r̄ e	**X**	'e k i s
Ch	t ʔ e	**I**	i	**N**	'e n e	**S**	'e s e	**Y**	'i g r j e g a
D	d e	**J**	'x o t a	**Ñ**	'e ɲ e	**T**	t e	**Z**	'θ e t a
E	e	**K**	K a	**O**	o	**U**	u		

Tabla 1: *Transcripciones de las pronunciaciones estándar de las letras en castellano.*

Cuando se trabaja con habla continua (sin pausas explícitas entre las palabras que forman la frase), una fuente importante de errores de reconocimiento es la coarticulación existente entre las palabras, en nuestro caso letras. Este efecto es más peligroso cuando las palabras del vocabulario son más cortas y con pronunciaciones similares como es nuestro caso.

En castellano existe una correspondencia directa entre la escritura de una palabra y su pronunciación (con la excepción de los hiatos que no están acentuados y no forman diptongo). Los castellano–hablantes no estamos acostumbrados a deletrear puesto que habitualmente no lo necesitamos para conocer la escritura de una palabra. Por esta razón, en el proceso de deletreo en castellano aparecen con cierta frecuencia los siguientes efectos que se describen a continuación:

- Existencia de pausas grandes entre las pronunciaciones de las letras, y con duración muy variable.
- Abundancia de ruidos cometidos por el locutor: tos, respiración fuerte, dentelleo y pausas rellenas como uh, um, er, mm.
- Gran cantidad de errores cometidos por el locutor.

2.2 Arquitectura de reconocimiento propuesta

La arquitectura de reconocimiento propuesta se muestra en la figura 2. En una primera fase se realiza un análisis o parametrización de la señal de voz. Este proceso se va realizando cada 10 ms con ventanas de análisis de 25 ms. Las muestras de voz que corresponden a cada una de las ventanas analizadas se denominan tramas de voz. En bs experimentos presentados, consideraremos como parámetros de cada trama 10 coeficientes cepstrales, la energía local de la trama y sus respectivas derivadas, tanto de la energía como de los coeficientes cepstrales (en total 22 parámetros para caracterizar cada trama de voz). La parametrización utilizada es la RASTA-PLP.

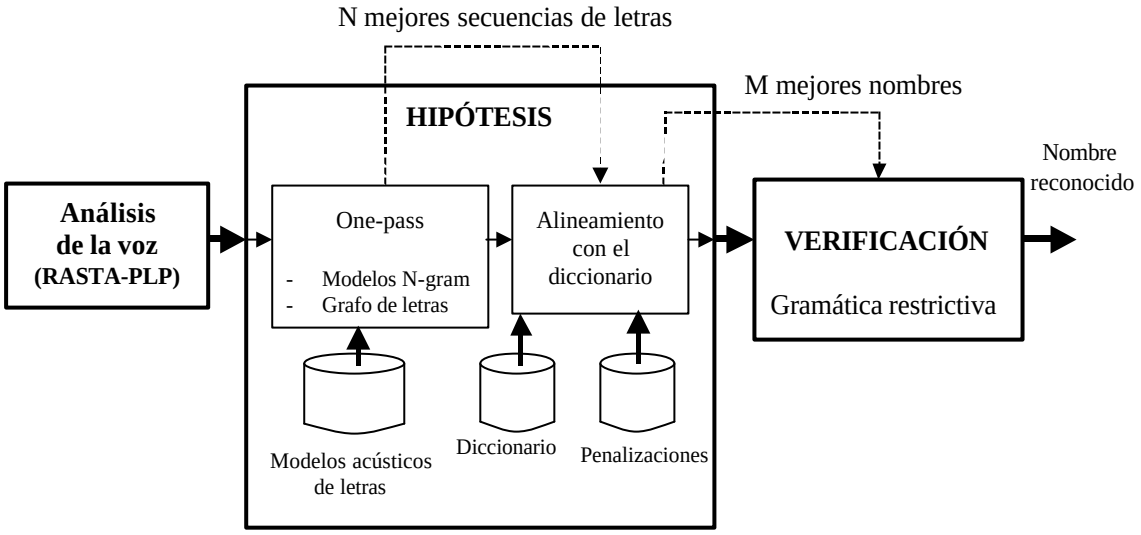

Figura 2: Diagrama de bloques del reconocedor de nombres deletreados.

En la etapa de hipótesis consideraremos dos fases:

- En la primera fase se aplica el algoritmo de One-pass para obtener la secuencia de letras que mejor se ajusta acústicamente al nombre pronunciado. En esta etapa, consideraremos una nueva topología con modelos de silencios contextuales, la incorporación de modelos acústicos para los ruidos, la utilización de modelos de lenguaje 2 ó 3-gram, y la obtención de las N mejores cadenas de letras.
- Una vez obtenidas las N mejores secuencias de letras, el objetivo de la segunda fase es obtener los M mejores nombres del directorio. Para ello, las N secuencias de letras se comparan con todos los nombres del diccionario mediante un algoritmo de programación dinámica. Este algoritmo aplica diferentes penalizaciones para las posibles sustituciones, borrados e inserciones de letras en la cadena. En esta comparación se calcula el mejor camino de alineamiento en el espacio de búsqueda entre secuencia y nombre, así como el coste asociado con dicho camino. Dada una secuencia de letras, se seleccionan los nombres del diccionario con menor coste de alineamiento.

En la fase de verificación utilizamos los M nombres del diccionario que más se parecen a las secuencias de letras obtenidas, y construimos con ellos una gramática en forma de árbol sobre la que se ejecuta el algoritmo de reconocimiento One-pass obteniendo finalmente el nombre reconocido.

2.3 Resultados de reconocimiento finales

En la tabla 2, se presentan las Tasas de Acierto al nivel de nombre para la etapa de hipótesis (seleccionando el nombre con el mejor alineamiento con las cadenas de letras obtenidas) y para la etapa de verificación que es la que define el comportamiento global del sistema. Se muestran resultados con diccionarios de 1.000, 5.000 y 10.000 palabras.

Tamaño del diccionario	TAN (Hipótesis)	TAN (Verificación)	M	TP (xRT)
1.000 (0,2)	94,2%	96,3%	10	2,8
5.000 (0,5)	88,7%	92,8%	20	3,4
10.000 (0,9)	86,2%	90,3%	50	4,7

Tabla 2: Resultados para varios diccionarios: Tasa de Acierto de nombre de las etapas de hipótesis y verificación, M y Tiempo de Proceso (TP en unidades de Tiempo Real RT).

Finalmente, se demuestra la utilidad del sistema desarrollado incorporándole en un servicio de información telefónica con la finalidad de recuperar errores ocurridos en el reconocedor de nombres propios. En la evaluación de campo realizada, se comprobó que gracias a este sistema se pudieron recuperar gran cantidad de errores de reconocimiento de nombres y apellidos, aumentando un 49% la tasa de llamadas atendidas automáticamente.

3. SISTEMA DE RECONOCIMIENTO DE HABLA CONTINUA EN DOMINIOS RESTRINGIDOS: FECHAS Y HORAS

3.1 Arquitectura de reconocimiento

El sistema de reconocimiento desarrollado responde al siguiente diagrama de bloques (figura 3):

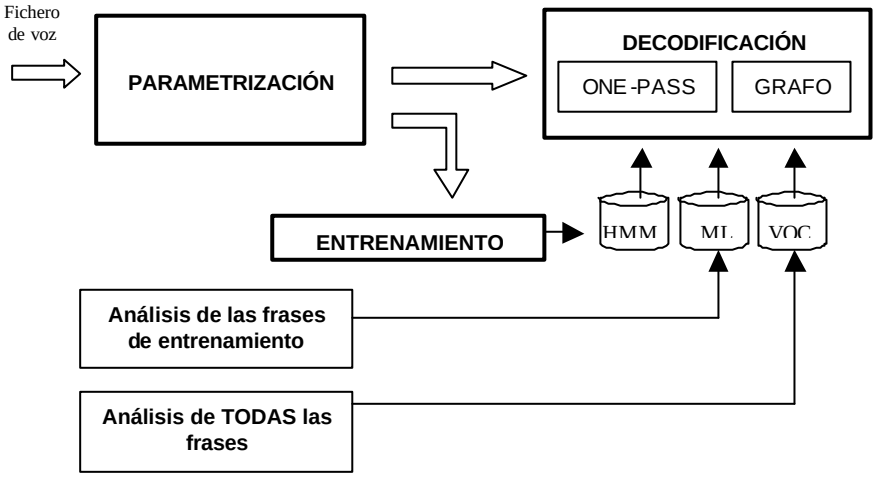

Figura 3: Diagrama del sistema de reconocimiento para fechas y horas.

En el proceso de parametrización, al igual que para el sistema de reconocimiento de nombres deletreados, se utilizará la técnica RASTA-PLP con ventanas de análisis de 25ms. En este sistema utilizamos también 10 coeficientes cepstral, la energía local de la trama y la primera derivada tanto de los coeficientes cepstral como de la energía (en total 22 parámetros). El proceso de descodificación está formado por dos etapas: la primera de ellas consiste en un algoritmo One-pass del que no obtenemos una única secuencia de palabras sino que obtenemos un grafo o lattice de palabras. Sobre este grafo, en una segunda etapa, aplicaremos modelos de lenguaje más potentes y

calcularemos la N mejores secuencias de palabras con bajo coste computacional. El proceso de descodificación utiliza los modelos acústicos (en nuestro caso modelos ocultos de Markov HMM: Hidden Markov Model) obtenidos de una etapa anterior de entrenamiento, un modelo de lenguaje (ML) calculado con las frases de referencia de los ficheros utilizados en el entrenamiento de los modelos acústicos, y el vocabulario (VOC) obtenido del análisis de todos los ficheros. Las principales conclusiones que se pueden extraer del trabajo realizado en el desarrollo de este sistema son las siguientes:

- Por un lado, la utilización de modelos de Markov con 5 estados y transiciones dobles ha permitido disponer de una mayor potencia y flexibilidad de modelado, redundando en una mejor tasa de reconocimiento.
- El entrenamiento selectivo, aunque no nos ha sido útil para aumentar la tasa de reconocimiento, nos ha permitido evaluar la resolución del modelado acústico, poniendo de manifiesto la posibilidad de entrenar modelos más detallados. En este punto, se ha analizado la evolución del número de centroides según dos criterios de selección de gaussianas en modelos semicontinuos: Fuzzy Vector Quantitation (FVQ) y selección de gaussianas con mayor peso. Se ha concluido que la segunda de las soluciones permite hacer mejor uso de los datos de entrenamiento.
- Otro aspecto importante descrito y analizado en este tema, ha sido la simplificación del algoritmo propuesto por Ney para la construcción de un grafo de palabras en ausencia de la técnica de Beam Search. En este capítulo se ha descrito la incorporación del modelo 3-gram en el grafo de palabras, las diferentes formas de postprocesar el grafo, y la manera de obtener un número N de hipótesis de reconocimiento en lugar de una única frase reconocida.
- Durante el desarrollo del sistema se han analizado las diferencias entre el habla leída y el habla espontánea proponiendo la necesidad de utilizar modelos de lenguaje diferentes para cada una de ellas.

3.2 Resultados de reconocimiento finales

En la figura 4 se muestran las tasas mínimas de error obtenidas cuando se consideran varias hipótesis de reconocimiento a la salida del reconocedor utilizando modelos de lenguaje 3-gram independientes para cada tipo de habla: espontánea y leída.

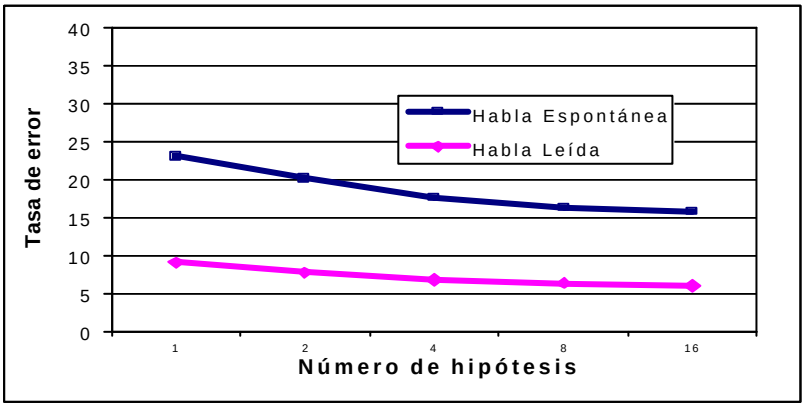

Figura 4: Evolución de la tasa de error según el número de hipótesis.

A medida que vamos aumentando el número de hipótesis consideradas, la tasa mínima de error decrece. Como se observa en la figura 4 esta tasa de error tiende a saturarse entorno al 16% para habla espontánea (partiendo de un 23% con una única hipótesis) y alrededor del 6% para habla leída (partiendo de un 9% para una única hipótesis). Al aumentar el valor de N llegamos a un punto de saturación de la tasa de error a partir del cuál no se reduce de forma importante el error.

4. ANÁLISIS DE MEDIDAS DE CONFIANZA EN SISTEMAS DE RECONOCIMIENTO DE HABLA CONTINUA

Debido a que el reconocimiento automático del habla dista mucho de ser perfecto, se debe analizar la calidad de lo reconocido/comprendido por el sistema con el fin de detectar posibles errores o zonas de gran ambigüedad. Esta necesidad es aún más importante en Servidores Vocales Interactivos donde una mala interpretación de la frase pronunciada puede llevar al sistema a realizar un comportamiento erróneo. Típicamente en los SVIs existen dos módulos anteriores al módulo de gestión de diálogo: reconocimiento y comprensión. Dichos módulos se encargan de extraer la información semántica de la frase pronunciada por el usuario. Esta información es utilizada por el gestor de diálogo para avanzar en su interacción con él. Las medidas de confianza obtenidas en estos dos módulos tienen como objetivo evaluar su comportamiento de forma que el gestor de diálogo pueda medir la calidad de la información recibida y en consecuencia, elegir la acción concreta a realizar: rechazar la frase, preguntar otra vez, o pedir confirmación de alguno de los datos obtenidos. Por otro lado, la evolución del propio diálogo también puede darnos información que ayude a mejorar su gestión.

4.1 Parámetros para la estima de confianza

Los parámetros considerados para obtener estas medidas de confianza se pueden clasificar según su origen en:

- ***Parámetros del Descodificador***: son medidas obtenidas de la evolución del reconocedor a lo largo de la etapa de descodificación de la señal de habla. Estas medidas tratan de detectar zonas de voz en las que existe un desacople importante entre los modelos acústicos y la voz pronunciada, o zonas en las que aparecen varias alternativas con gran confusión acústica entre ellas.
- ***Parámetros exclusivos del Modelo de Lenguaje****:* estos parámetros tratan de validar que la secuencia de palabras obtenida, corresponde con patrones gramaticales característicos, observados en las frases de los usuarios a lo largo de sus interacciones con el sistema.
- ***Parámetros de Comprensión****:* son medidas obtenidas del analizador semántico y tratan de reflejar la fiabilidad con la que han sido obtenidos los conceptos a partir de la secuencia de palabras reconocidas.
- ***Parámetros del Gestor de diálogo****:* el punto del diálogo en el que estemos, también puede ayudarnos a evaluar la calidad de los datos obtenidos: por ejemplo si el sistema está solicitando del usuario el nombre de la ciudad origen de un viaje, el hecho de obtener un nombre de un hotel nos debe alertar sobre un cambio de intención por parte del usuario o de la existencia de problemas en el reconocimiento. Por otro lado, la información obtenida durante la propia evolución del diálogo, como el número de interacciones necesarias para conseguir un determinado objetivo o el número de confirmaciones negativas y/o positivas durante la consulta, nos puede dar una idea de cómo se está desarrollando dicha interacción.

4.2 Niveles de estima de confianza

Según la resolución de las medidas de confianza, podemos clasificarlas en 4 niveles diferentes:

- ***Nivel de palabra**:* en este caso el objetivo es detectar palabras mal reconocidas. Estos errores de reconocimiento se pueden haber producido por problemas del descodificador, o porque la palabra no está incluida en el vocabulario de reconocimiento (OOV: out of vocabulary). En el caso de un sistema de habla continua se pretende detectar las inserciones y substituciones de palabras.
- ***Nivel de concepto**:* en este caso se pretende detectar conceptos erróneos dentro de una frase determinada. Las medidas de confianza en este caso son muy importantes para la gestión de diálogo puesto que es la información semántica, la que se utiliza para realizar esta labor de gestión y decidir cuales van a ser las acciones del sistema en su interacción con el usuario.
- ***Nivel de frase**:* en este nivel, el objetivo es detectar por un lado frases fuera del dominio de la aplicación y por otro, frases del dominio con problemas en el reconocimiento que no tienen ninguna información semántica o concepto correcto. Se pretende por tanto, detectar frases que no van a ser correctamente reconocidas y comprendidas por nuestro sistema, evitando que se detecte algún concepto erróneamente que le haga al gestor realizar una mala interpretación.
- ***Nivel de interacción**:* el principal objetivo a este nivel es medir la calidad de la interacción. Con estas medidas se pretende detectar situaciones problemáticas como las siguientes: que el diálogo emprenda un camino que diverge de las necesidades del usuario debido a un error de compresión, que la tasa de reconocimiento del sistema esté siendo muy baja por problemas de gran ruido ambiente, o situaciones en las que la respuesta del usuario no se ajusta a las preguntas del sistema por desconocimiento de la funcionalidad y/o las limitaciones del servicio. En estos casos es necesario dotar de mecanismos de corrección ágiles que permitan volver a puntos anteriores del diálogo, disponer de variedad de sistemas de reconocimiento que permitan mayor robustez en ambientes ruidosos aun a costa de perder flexibilidad, y también son necesarias estrategias de modelado de usuario que permitan adaptar las preguntas, informaciones o ayudas del sistema, a la destreza del usuario.

4.3 Principales resultados obtenidos

En el presente apartado se describen los resultados obtenidos en el análisis de medidas de confianza para el sistema CU Communicator, desarrollado en The Center for Spoken Language Research (CSLR) de la Universidad de Colorado, y en los sistemas de reconocimiento de habla continua realizados en la presente tesis: reconocimiento de nombres deletreados y reconocimiento de fechas y horas.

Sobre el sistema CU Communicator se ha trabajado sobre en los niveles de palabra, concepto semántico y frase. En este estudio, hemos utilizado parámetros provenientes del descodificador, del modelo de lenguaje y del módulo de comprensión. Por otro lado, en el reconocedor de fechas y horas trabajaremos principalmente al nivel de palabra, utilizando parámetros tanto del descodificador como del modelo de lenguaje. Las mismas fuentes de parámetros se han utilizado para obtener medidas de confianza al nivel de frase (secuencia de letras correspondientes a un nombre) en el caso del reconocedor de nombres deletreados. En este último sistema, se analizará también el problema de la detección de nombres fuera del vocabulario de reconocimiento (OOV: Out Of Vocabulary). En todos los casos, para la combinación de los diferentes parámetros considerados y la obtención de un único valor de confianza, utilizaremos una Red Neuronal sencilla, un Perceptrón Multi-Capa.

En cuanto a los experimentos realizados sobre el sistema CU Communicator cabe resaltar las siguientes conclusiones:

- Se ha realizado un estudio importante de diferentes parámetros con el fin de proporcionar medidas de confianza tanto para el sistema de reconocimiento como el sistema de comprensión. De los resultados presentados podemos resumir que considerando como punto de trabajo un Rechazo Incorrecto (RI) del 5%, hemos conseguido rechazar más del 50% de palabras erróneas y conceptos incorrectos, y más del 76% de frases mal interpretadas semánticamente por el sistema.
- Al nivel de palabra (figura 5), los parámetros obtenidos del modelo de lenguaje funcionan mejor para tasas de RI bajas. Combinando los parámetros del proceso de descodificación y del modelo de lenguaje se consiguen resultados bastante mejores que utilizando cada grupo de parámetros de forma independiente, lo que pone de manifiesto la complementariedad de ambas fuentes de información.

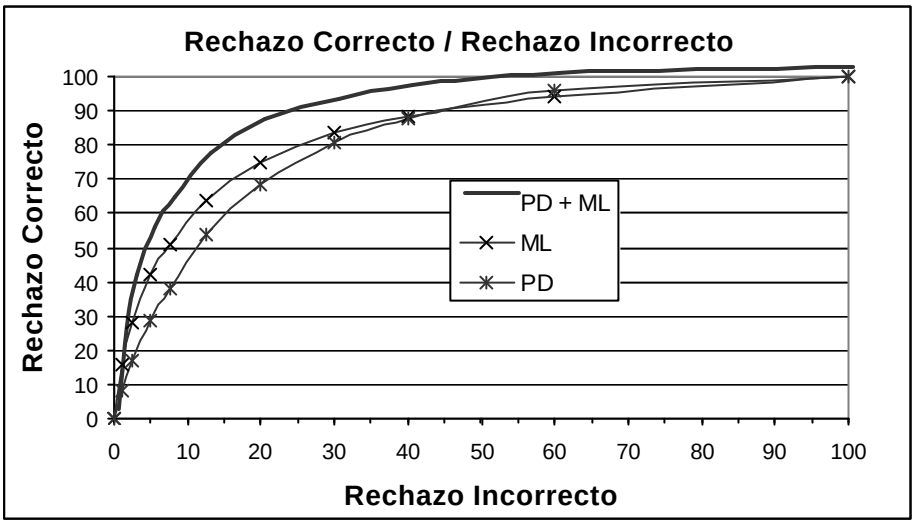

Figura 5: Rechazo Correcto (RC) vs Rechazo Incorrecto (RI) para los parámetros del proceso de descodificación (PD) del modelo de lenguaje (ML) o todos (PD + ML) (NIVEL DE PALABRA).

- En los niveles de concepto y frase, cabe comentar que las medidas obtenidas al nivel de palabra y de concepto respectivamente, son muy útiles para predecir la confianza en niveles superiores.
- En esta tesis también se propone la utilización de las medidas de confianza para combinar varias hipótesis de reconocimiento de uno o varios descodificadores. Para realizar esta combinación se han propuesto dos métodos diferentes: FLCR y el WGCR, consiguiendo reducciones relativas del error superiores al 15%. Esta reducción se consigue cuando se combinan hipótesis de varios reconocedores, y no cuando combinamos exclusivamente hipótesis del mismo reconocedor.

Sobre el reconocedor de nombres deletreados las conclusiones son las siguientes:

- En este reconocedor se proponen parámetros provenientes de los diferentes pasos que forman el proceso de descodificación, consiguiendo rechazos correctos del 58% de errores de reconocimiento y del 68% de nombres deletreados no pertenecientes al diccionario, para RI del 5%.
- Otro aspecto que conviene resaltar es el gran poder de discriminación ofrecido por el parámetro “Diferencia de Verosimilitudes entre Módulos (DVM-3)” para la detección de nombres fuera del diccionario de reconocimiento.

- A medida que se van utilizando parámetros de etapas más avanzadas, se consigue un mejor poder de discriminación. En este caso, cada etapa más avanzada utiliza fuentes de información más potentes, y además, las decisiones tomadas en estos módulos, tienen una influencia directa sobre la tasa final de reconocimiento.
- La discriminación entre errores y palabras fuera del diccionario de reconocimiento es una tarea muy complicada y los resultados dependen fuertemente del número de casos de ejemplo disponibles para entrenar la Red Neuronal utilizada como clasificador.

En relación con el reconocedor de fechas y horas se han realizando experimentos para habla leída y espontánea. Muchas de las conclusiones obtenidas son similares a las mostradas sobre el sistema CU Communicator al nivel de palabra, por esta razón, comentaremos únicamente las conclusiones adicionales:

- Los resultados obtenidos en este caso son peores que los obtenidos para el sistema CU Communicator al nivel de palabra, lo que pone de manifiesto que un reconocedor con mejores modelos acústicos y/o lingüísticos permite obtener mejores parámetros para la obtención de medidas de confianza.
- Para este reconocedor también obtenemos mejores resultados cuando se utiliza la medida de confianza obtenida al nivel de palabra como heurístico para la combinación de hipótesis de reconocimiento, aunque en este caso las diferencias no sean estadísticamente significativas por no disponer de suficientes datos de evaluación.

5. DISEÑO DE GESTORES DE DIÁLOGO

El gestor de diálogo es el módulo más importante de un Servidor Vocal Interactivo. Este módulo es el encargado de gestionar los recursos ofrecidos por el resto de módulos para dirigir la interacción del sistema con el usuario. La importancia de este módulo se debe a que el diálogo es la ventana a través de la cual los usuarios perciben el comportamiento y la aparente “inteligencia” del sistema. En esta tesis se propone una metodología para el diseño de gestores de diálogo.

Esta metodología se presenta mediante su aplicación al diseño de un gestor de diálogo para un servicio de información y reserva de billetes de tren. La metodología propuesta está formada por 5 etapas. La primera etapa es el Análisis de la Base de Datos en la que se describe, mediante un diagrama Entidad-Relación (E-R), el contenido de la base de datos utilizada para dar el servicio. En la etapa de diseño por Intuición se realiza un “brain-storming” sobre el diagrama E-R para proponer diferentes alternativas de diálogo. El tercer paso es el diseño por observación, donde se evalúan las transcripciones de diálogos usuario–operador. El siguiente paso consiste en simular el sistema mediante un Mago de Oz con el fin de aprender las características específicas de las interacciones entre un sistema automático y el usuario. El quinto y último paso es el diseño por Mejora Iterativa en el que el sistema entra en un proceso de prueba (por usuarios reales) y mejora constante

5.1. Análisis de la base de datos

Lo primero que debemos remarcar es que la base de datos a analizar no se refiere a una base de datos de voz ni de diálogos etiquetados. Se refiere a la base de datos que contiene la información que queremos proveer en nuestro servicio. Por ejemplo, en el caso del servicio de información y reserva de billetes de tren, la base de datos contiene la información de todos los posibles viajes en tren a lo largo de la geografía española; tanto la información de horarios y precios como de las reservas de los

usuarios. Este análisis consiste en una descripción a alto nivel de la base de datos. La mejor forma de describir una base de datos es representando su diagrama Entidad-Relación (E-R). Este diagrama no tiene porqué representar fielmente el contenido completo de la base de datos. El objetivo de esta representación es mostrar las relaciones existentes entre los datos *que serán útiles para el desarrollo del servicio*.

El diagrama E-R no tiene una solución única y el resultado final depende del desarrollador y del servicio final que queramos diseñar. A la hora de desarrollar este diagrama E-R debemos tener en cuenta los siguientes puntos:

- Hay que prestar especial interés a la definición de los Conjuntos Entidad y Conjuntos Asociación porque la información contenida en ellos será la que defina los posibles objetivos (partes del servicio: información de horarios, información de precios...) que se pueden ofrecer en nuestro servicio.

- Por otro lado, los atributos que forman la Clave de un Conjunto Entidad o Asociación serán los datos obligatorios que el sistema debe preguntar al usuario para poder ofrecer información de dicho Conjunto Entidad. El resto de atributos serán datos opcionales.

5.2. Diseño por Intuición

En este paso se pretende definir, sobre el diagrama E-R, posibles propuestas sobre el tipo de servicio o conjunto de objetivos que deseamos ofrecer así como posibles flujos de diálogo: tanto la secuencia de objetivos dentro del propio servicio, como las secuencias de preguntas al usuario que nos permiten obtener los datos necesarios para satisfacer dichos objetivos. Este paso es como un “brain-storming” realizado sobre el diagrama E-R obtenido en la fase anterior. El principal objetivo de esta fase es proponer alternativas sobre los siguientes aspectos del diálogo:

- **Objetivos que forman el servicio:** en este caso la idea es proponer varios objetivos que podrían ser resueltos por el sistema a tenor de la información representada en el diagrama E-R. Debemos prestar especial atención sobre los conjuntos entidad y asociación puesto que estos conjuntos representan los conceptos principales de la aplicación y su contenido es la base del servicio y de los objetivos de diálogo que lo forman.
- **Datos del usuario para cada objetivo:** debemos seleccionar los datos que el sistema necesita para satisfacer un objetivo, es decir, la información que se necesita del usuario para ofrecerle el servicio correctamente. En este punto debemos considerar los atributos que forman la clave de cada uno de los conjuntos entidad o asociación. Estos atributos son los datos obligatorios que se deben especificar para hacer referencia a una unidad dentro del conjunto. Además es necesario plantearse las posibilidades que utilizan los usuarios para especificar cada uno de estos atributos teniendo en mente la tecnología del habla disponible.
- **Secuencia por defecto para preguntar los datos:** se deben plantear varias alternativas para la definición del orden en el que se preguntan los datos al usuario.

5.3. Diseño por Observación

El diseño por observación está basado en el análisis de conversaciones reales entre los usuarios y los operadores humanos en un servicio análogo al que se quiere automatizar. En este análisis, se debe apuntar las veces que aparecen las alternativas propuestas en la fase anterior, con el fin de evaluar el impacto de cada una de ellas.

Para ello, podemos utilizar la hoja o tabla obtenida en el diseño por intuición. Puede ocurrir que aparezcan nuevas opciones no consideradas anteriormente, lo que obligaría a incluirlas en nuestro análisis.

Veamos los aspectos que debemos anotar y evaluar en esta fase:

- **En relación con los objetivos del servicio.** Debemos detectar los *objetivos que más frecuentemente se solicitan* por parte del usuario y el orden en el que se especifican. Otro aspecto importante es la *información que el operador humano ofrece* al usuario para satisfacer un determinado objetivo: tanto la información concreta como el orden o la manera de ofrecerla.
- **En relación con los datos a preguntar al usuario.** Debemos anotar *qué datos son los necesarios para completar un objetivo* y la secuencia en la que el operador los pregunta. Clasificación de cada dato como *obligatorio u opcional.* Clasificación de cada dato como *simple o complejo* (consideraremos un dato como complejo cuando se puede dividir en datos simples). Otro aspecto importante es el análisis de las *diferentes formas* que utiliza un usuario *para especificar el valor de un dato* concreto. El último aspecto a analizar relacionado con los datos es la secuencia en la que el operador va preguntando al usuario los datos necesarios. Para ello debemos anotar la posición de cada dato en la secuencia. Definida la secuencia, se debe hacer una agrupación de datos en subobjetivos o pasos del diálogo. Esta agrupación se suele realizar por afinidad semántica entre los datos (ej: estaciones de tren de origen y destino) o por proximidad en la secuencia de petición al usuario. Estos pasos nos permiten definir un ritmo del diálogo.
- **En relación con la negociación entre sistema y usuario.** Consideraremos que hay negociación cuando el usuario debe elegir o rechazar alguna de las alternativas que el sistema le ofrece, pudiendo cambiar alguna de las restricciones. Los aspectos que debemos analizar son: criterios que más frecuentemente se utilizan para elegir una opción de entre un conjunto de opciones: precio, duración, hora de salida,... Estos criterios serán los que nos permitan diseñar el tipo de información que debemos ofrecer para cada una de las opciones (ofrecer primero la información más útil para elegir).
  El proceso de negociación se puede realizar de dos maneras: presentar una única opción y permitir al usuario que solicite la opción anterior/posterior (navegación), o presentar varias alternativas y que el usuario elija una de ellas Para los casos en los que se opte por la segunda estrategia de negociación, debemos analizar el número de opciones que el usuario puede retener simultáneamente.

La principal limitación de esta etapa es que estamos analizando conversaciones entre personas (usuario–operador) y en estos casos el comportamiento del usuario al hablar es diferente que cuando interacciona con un sistema automático. En esta fase por tanto, se pueden aprender características generales del diálogo como las que se han descrito, pero no detalles de comportamiento de los usuarios frente a un sistema automático.

5.4. Diseño por Simulación

En esta fase se pretende simular el sistema mediante una herramienta de Mago de Oz. Con esta herramienta podemos implementar rápidamente las propuestas sugeridas en los pasos anteriores y evaluar su comportamiento. Además, nos permite experimentar con usuarios reales del servicio original sin que la calidad del servicio se vea afectada de manera importante. En esta etapa, la estrategia de diseño se basa en probar varias alternativas del diálogo y evaluar el funcionamiento de cada una de ellas.

Aspectos a analizar:

- Objetivos del servicio. Hay que ver si la cobertura del servicio es la adecuada o si los usuarios se esperaban más funcionalidad, y si la organización de los objetivos dentro del diálogo es la esperada.
- Datos:
  - Analizar la inteligibilidad de las preguntas, y ver si las respuestas de los usuarios pueden ser incluidas en el vocabulario del reconocedor.
  - Si la secuencia de preguntas es la apropiada o si los usuarios tienen problemas al contestar a unas preguntas antes que otras.
  - El caso en el que tengamos que plantear varias preguntas para recoger un dato complejo, debemos decidir si la secuencia de preguntas ofrece una mayor tasa de éxito requiriendo un menor tiempo de consulta.
- Negociación. En nuestro caso hemos decidido utilizar la estrategia de presentar varias opciones a la vez por analogía con el servicio actual. En este caso se debe analizar la información a incluir en cada opción y el número de opciones a presentar simultáneamente.

Medidas de evaluación:

- Evaluación de la cobertura y estructura de los objetivos: el número de veces que un determinado objetivo es solicitado por el usuario, el tiempo requerido y el número de turnos de diálogo (pregunta–respuesta) necesarios para satisfacerlo. En un cuestionario se puede preguntar al usuario por nuevas funcionalidades.
- En relación con el diseño de las preguntas: grabación de las respuestas de los usuarios, nº de veces que el usuario se queda callado ante una pregunta, tasa de reconocimiento (con una primera versión del reconocedor).
- Para analizar la mejor secuencia de preguntas que forman el diálogo. Se implementan varias soluciones posibles y elegiremos una u otra en cada llamada de forma aleatoria. Como medidas de evaluación consideraremos la grabación de las respuestas de los usuarios y la tasa de reconocimiento de la secuencia obtenida como multiplicación de las tasas de cada uno de los datos.
- Definición de varias posibilidades en cuanto al número de opciones a presentar y el patrón en el que está estructurada la información para cada una de las opciones. Estas posibilidades cambian de forma aleatoria de unas llamadas a otras (pero no dentro de una misma llamada) de forma que los usuarios puedan probar varias de las opciones y mediremos el nº de preguntas necesarias en la negociación y el tiempo consumido.

5.5. Diseño por Mejora Iterativa

Este último paso de la metodología consiste en un proceso iterativo de prueba y mejora del sistema hasta llegar a una situación o versión estable. En esta etapa nos hemos centrado en dos aspectos importantes: diseño de los mecanismos de confirmación y modelado del usuario.

Para conseguir una gestión eficiente de estos mecanismos de confirmación, es necesario utilizar las medidas de confianza obtenidas en la fase de reconocimiento. Dependiendo de la confianza obtenida en reconocimiento debemos adoptar una u otra estrategia de confirmación.

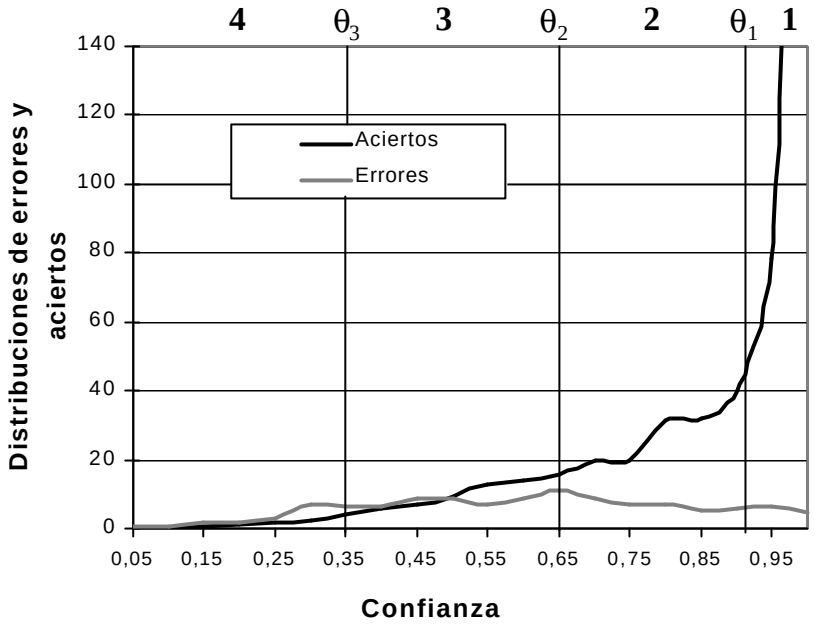

Figura 6**:** Distribución de aciertos y errores con los umbrales para el diseño de los mecanismos de confirmación.

Para utilizar las medidas de confianza en el diseño de los mecanismos de confirmación, debemos representar las distribuciones de aciertos y errores de reconocimiento en función del valor de confianza obtenido. Sobre esta distribución, se definen varios umbrales que dividen el gráfico en varias zonas (figura 6). Para cada una de estas zonas definimos estrategias de confirmación diferentes. En las zonas de mayor confianza se adoptarán estrategias de confirmación más arriesgadas y que consuman poco tiempo, mientras que en las zonas de baja confianza se adoptarán estrategias más seguras independientemente del tiempo necesario.

La técnica de modelado del usuario propuesta en la presente tesis está basada en la definición de niveles de destreza para cada uno de los aspectos a modificar de nuestro diálogo, y en la consideración de ciertos eventos que, ocurridos en el transcurso de la interacción, nos hacen cambiar de uno a otro nivel de destreza. Dependiendo del nivel de destreza se definen diferentes preguntas, ayudas, mecanismos de confirmación,... que pretenden adaptarse mejor a las características del usuario. En la figura 6, se muestra un ejemplo con 4 niveles de destreza.

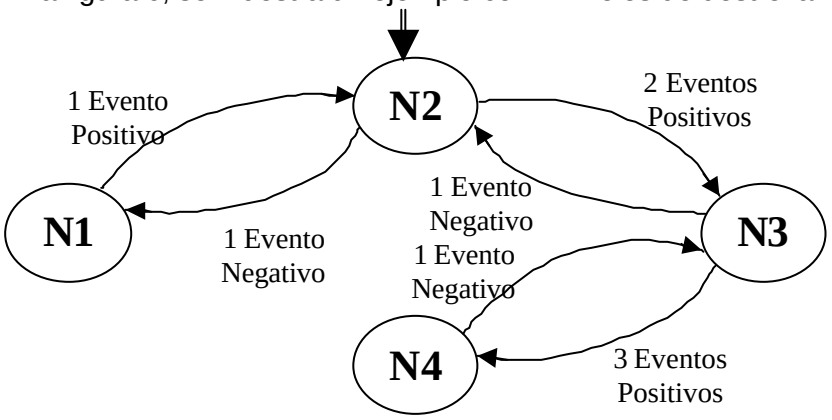

Figura 7**:** Diagrama de niveles de destreza en la técnica de modelado de usuario.

Al comienzo de la interacción el sistema ofrece una ayuda inicial y se sitúa en el segundo nivel por defecto. Cuando ocurre un evento negativo desciende al nivel 1, mientras que si ocurren dos eventos positivos el sistema aumenta hasta el nivel 3. Como podemos observar de la figura 7, a medida que aumentamos el nivel de

destreza exigimos un mayor número de eventos positivos para seguir aumentando de nivel.

5.6. Resultados finales

Medida	Valor
Duración media de la consulta (segundos)	195
Nº medio de preguntas realizadas por el sistema	18,83
Nivel medio de destreza del usuario	1,88
Varianza del nivel de destreza del usuario	0,32
% de confirmaciones implícitas	71,0%
% de confirmaciones explícitas	29,0%
Nº volver a empezar	0,26
Nº de veces que el usuario solicita la corrección de un dato	0,32
Duración de la negociación de la opción del tren (segundos)	58
Nº medio de veces que el usuario solicita repetición	0,18

Tabla 3: Medidas obtenidas de las anotaciones realizadas por el sistema.

En este apartado vamos a presentar los resultados de la evaluación de campo realizada sobre el sistema de información y reserva de billetes de tren. Esta evaluación ha consistido en diversas llamadas al sistema por parte de 105 usuarios para completar 4 escenarios de viaje. En total se obtuvieron 355 consultas puesto que no todos los usuarios completaron los 4 escenarios. Sólo en un 19,8% de las llamadas el usuario colgó sin recibir ningún tipo de información, luego el porcentaje de llamadas en las que se ofreció información sobre los trenes fue de un 80,2%.

En cuanto al cuestionario, se han analizado aspectos subjetivos pidiendo al usuario que evalúe cada uno de ellos de 1 a 5. Los resultados medios obtenidos se presentan en la tabla 4.

Medida subjetiva	Valor
El sistema comprende lo que le dices.	3,1
Las respuestas del sistema son claras y concisas.	3,4
Entiendo lo que el sistema me dice.	3,5
Se accede a la información de trenes rápidamente.	2,9
El sistema es fácil de usar.	3,5
Es fácil de aprender su funcionamiento.	3,7
El sistema me ayuda durante la conversación.	3,1
En caso de error la corrección fue fácil.	2,9
El sistema me hace las preguntas en un orden lógico.	3,6
En general, es un buen sistema.	3,0

Tabla 4: Medidas subjetivas recogidas de los cuestionarios.

En los resultados presentados en la tabla 4, podemos ver cómo muchos de los aspectos obtienen una puntuación superior a la media (3), lo que indica que el sistema

está funcionando razonablemente bien. Los aspectos peor valorados son la agilidad o rapidez del sistema y las opciones de corrección. Las mejores puntuación se obtuvieron para la inteligibilidad del sistema y para la estructura del diálogo.

PUBLICACIONES RELACIONADAS CON LA TESIS DOCTORAL

Las publicaciones derivadas de la presente Tesis Doctoral se distribuirán en cuatro apartados: revistas internacionales, congresos internacionales, revistas nacionales, congresos nacionales e informes técnicos. La referencias se han ordenado cronológicamente en cada uno de los apartados.

REVISTAS INTERNACIONALES DE INVESTIGACIÓN

Spanish recognizer of continuously spelled names over the telephone

R. San-Segundo, J. Colás, R. Córdoba, J.M. Pardo.
Speech Communication 38 pp.287-303 Octubre de 2002. (ISSN 0167-6393)
Factor de Impacto 0.44

En proceso de revisión:

Different Levels for Confidence Annotation in Spoken Dialogue Systems

R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, J.M. Pardo
Speech Communication. (ISSN 0167-6393)

Knowledge Combining Methodology for Dialogue Design in Spoken Language Systems

R. San-Segundo, J.M. Montero, J. Ferreiros, J.M. Pardo
International Journal of Speech Technology. (ISBN 1-55798-241-4)

Adapting a search algorithm to the Spanish Railway Network

R. San-Segundo, J. Macías-Guarasa, J. Ferreiros, J.M. Montero, J.M. Pardo.
Transportation Research Part A: Policy and Practice (ISSN 0965-8564)

CONGRESOS INTERNACIONALES

Methodology for Dialogue Design in telephone-based spoken dialogue systems: a Spanish train information system

R. San-Segundo, J. M. Montero, J. Colás, J. Gutiérrez, J.M. Ramos, J.M. Pardo.
EUROSPEECH'2001. Sep 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

Detection of Recognition Errors and Out of the spelling dictionary names in a spelled name recognizer for Spanish

R. San-Segundo, J. Macías-Guarasa, J. Ferreiros, P. Martín, J.M. Pardo.
EUROSPEECH'2001. Sep 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

An Interactive Directory Assistance Service for Spanish with Large-Vocabulary Recognition

R. Córdoba, R. San-Segundo, J.M. Montero, J. Colás, J. Ferreiros, J. Macías, J.M. Pardo.
EUROSPEECH'2001. Sep 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

Confidence measures for Spoken Dialogue Systems

R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, J.M. Pardo.
ICASSP'2001, Mayo 5-11, Salt Lake City, Utah, USA. (ISBN: 0-7803-7043-0)

A Telephone-Based Railway Information System for Spanish: Development of a Methodology for Spoken Dialogue Design

R. San-Segundo, J. M. Montero, J. Gutiérrez, A. Gallardo, J.D. Romeral, J.M. Pardo.
SIGDIAL 2001 WORKSHOP. Septiembre 1-2, 2001. Aalborg (Dinamarca).

Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish

R. San-Segundo, J. M. Montero, J. Ferreiros, R. Córdoba, J.M. Pardo.
SIGDIAL 2001 WORKSHOP. Septiembre 1-2, 2001. Aalborg (Dinamarca).

Sistema de información ferroviaria por teléfono: propuesta de una metodología de diseño de gestores de diálogo

R. San-Segundo, J. M. Montero, J.M. Pardo.
SLPLT-2. Septiembre 14-15, 2001. Jaén. España.

Spanish Recogniser of continuously spelled names over the telephone

R. San-Segundo, J.Colás, J.Ferreiros, J.Macías-Guarasa and J.M. Pardo.
ICSLP'2000, Oct 7-11, Pekín, China. pp 863-866, Vol II. (ISBN 7-80150-114-4/G.18)

Stress assignment in Spanish Proper Names

R. San-Segundo, J.M. Montero, R. Córdoba, J.M. Gutiérrez.
ICSLP'2000, Oct 7-11, Pekín, China. pp 346-349, Vol III. (ISBN 7-80150-114-4/G.18)

Confidence measures for dialogue management in the CU COMMUNICATOR system

R. San-Segundo, B. Pellom, W. Ward, J.M. Pardo.
ICASSP'2000, Junio 5-9, Estambul, Turquía.

Acoustical and Lexical Based Confidence Measures for a Very Large Vocabulary Telephone Speech Hypothesis-Verification System

J. Macías-Guarasa, J. Ferreiros, R. San-Segundo, J.M. Montero and J.M. Pardo.
ICSLP'2000, Pekín, China. (ISBN 7-80150-114-4/G.18)

IDAS : Interactive Directory Assistance Service

Lehtinen, G., S. Safra, …, J.M. Pardo, R. Córdoba, R. San-Segundo, et al.
VOTS-2000 Workshop, Belgium.

Efficient vector quantization using an N-path Tree Search Algorithm

R. San-Segundo, R. Córdoba, J. Ferreiros, A. Gallardo, J. Colás, J. Pastor, Y. López.

EUROSPEECH'99, Sep 5-10,. Budapest (Hungría). pp. I 93-96. (ISSN: 1018-4074)

A variable preselection list length estimation using neural networks in a telephone speech hypothesis-verification system

J. Macías, J. Ferreiros, A. Gallardo, R. San-Segundo, J.M. Pardo and L. Villarrubia.
EUROSPEECH'1999, Septiembre 5-10, 1999. Budapest (Hungría). (ISSN: 1018-4074)

An asymmetric stochastic language model based on multi-tagged words

J. Pastor, J. Colás, R. San-Segundo, J.M. Pardo.
ICSLP'1998 , Diciembre 1-4, 1998 Sydney (Australia). (ISBN 1-876346-17-5)

REVISTAS NACIONALES

Tecnología del Habla para aplicaciones Multilingüe, Multiservicio y Multiplataforma

L. Villarrubia, M.A. Rodriguez, J. Relaño, F.J. Garijo, J. Bernat, L. Hernández,
R. San-Segundo, D. Tapias, L.A. María.
COMUNICACIONES DE TELEFÓNICA I+D, nº 30 Marzo, 2003. (ISSN 1130-4693)

CONGRESOS NACIONALES

Tecnología del Habla para aplicaciones Multilingües, Multiservicio y Multiplataforma

Luis Villarrubia, R. San-Segundo, Luis Hernández, Gregorio Escalada,
II JORNADAS DE TECNOLOGÍA DEL HABLA 16-18, Diciembre, 2002.

Entorno para el desarrollo de aplicaciones multimedia con síntesis y reconocimiento de voz

R. San-Segundo, J.M. Montero, J.Colás, R. Córdoba, J. Ferreiros, A. Gallardo, J. Macías-Guarasa, J.M. Gutiérrez, J. Pastor, J.M. Pardo.

X JORNADAS TELECOM I+D Barcelona-Madrid, Noviembre, 2000.

(ISBN: 84-607-1397-0)

Optimización de un servicio automático de paginas blancas por teléfono: proyecto IDAS

R. Córdoba, R. San-Segundo, J.Colás, J.M. Montero, J. Ferreiros, J. Macías-Guarasa, A. Gallardo, J.M. Gutiérrez, J. Pastor, J.M. Pardo.

X JORNADAS TELECOM I+D Barcelona-Madrid, Noviembre, 2000.

(ISBN: 84-607-1397-0)

Servidores vocales interactivos: desarrollo de un servicio de páginas blancas por teléfono con reconocimiento de voz. Proyecto IDAS (Interactive telephone-based Directory Assistance Service)

R. San-Segundo, J.Colás, J.M. Montero, R. Córdoba, J. Ferreiros, J. Macías-Guarasa, A. Gallardo, J.M. Gutiérrez, J. Pastor, J.M. Pardo.

IX JORNADAS TELECOM I+D Barcelona-Madrid, Noviembre, 1999. (ISBN: 84-7653-730-1)

INFORMES TÉCNICOS

Different Levels of Confidence Annotation in Spoken Dialogue Systems

R. San Segundo, B. Pellom, K. Hacioglu, W. Ward y J.M. Pardo.
(GTH-DIE-ETSIT-UPM / 1-01)

Methodology for developing dialogue managers in spoken dialogue systems

R. San Segundo y J.M. Montero
(GTH-DIE-ETSIT-UPM / 2-01).

Diseño semiautomático de diálogos de iniciativa mixta con independencia de modalidad e independencia de idioma

J.M. Montero y R. San Segundo
(GTH-DIE-ETSIT-UPM / 1-02).

OTROS MÉRITOS ATRIBUIDOS A LA REALIZACIÓN DE LA TESIS DOCTORAL

Los principales méritos atribuidos a la realización de la Tesis se resumen en tres aspectos diferentes. En primer lugar las publicaciones en las que se referencia la Tesis Doctoral o los artículos publicados con los resultados de dicha Tesis. En segundo lugar se describirá la aplicación industrial de la investigación realizada y por último, se comentará la transferencia de tecnología desde Estados Unidos con motivo de la realización de varias estancias en este país durante la realización de la Tesis.

PUBLICACIONES CON REFERENCIAS A LOS RESULTADOS DE LA TESIS

Las principales publicaciones con referencias a los resultados de la Tesis y conocidas por el autor son:

Using Word Confidence Measure for OOV Words Detection in a Spontaneous Spoken Dialog System
Hui Sun, Guoliang Zhang, Fang Zheng, Mingxing Xu,
EUROSPEECH'03. pp. 2713-2716. 2003. ISSN: 1018-4074

Development of a stochastic dialog manager driven by semantics

F. Torres, E. Sanchís, E. Segarra.
EUROSPEECH'03. pp 605-608. 2003. ISSN: 1018-4074

Word Level confidence measurement using semantic features
Ruhi Sarikaya, Yuqing Gao and Michael Picheny.
ICASSP 2003.

Recognition confidence scoring and its use in speech understanding systems.
Hazen, T,. Seneff, S., Polifroni, J., 2002.
Computer Speech and Language (2002) 16, 49-67.

Probabilistic Integration of multiple confidence measures and context information for concept verification
Yi-Ching Lin and Huei-Ming Wang.
ICASSP'2002. Mayo de 2002, Orlando, Florida (USA).

A concept graph based confidence measure
Kadri Hacioglu and Wayne Ward.
ICASSP'2002. Mayo de 2002, Orlando, Florida (USA).

Estimating semantic confidence for spoken dialogue systems
Sameer S. Pradhan and Wayne H. Ward.
ICASSP'2002. Mayo de 2002, Orlando, Florida (USA).

Issue-based dialogue management
Staffan Larson.
Tésis Doctoral. Department of linguistics Göteborg University, Sweden, 2002.

Robust semantic confidence scoring

Didier Guillevic, Simona Gandrabur, Yves Normandin.
ICSLP'2002. Septiembre de 2002, Denver, Colorado (USA).

Word Level Confidence Annotation using Combinations of Features
Rong Zhang and Alexander I. Rudnicky.
EUROSPEECH'2001. Sep. 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

Is This Conversation on Track ?
Paul Carpenter, Chun Jin, Daniel Wilson, Rong Zhang, Dan Bohus, Alex Rudnicky.
EUROSPEECH'2001. Sep. 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

Ambiguity Representation and Resolution in Spoken Dialogue Systems
Egbert Ammicht, Alexandros Potamianos, Eric Fosler-Lussier.
EUROSPEECH'2001. Sep. 3-7, 2001. Aalborg (Dinamarca). (ISBN 87-90834-09-7)

APLICACIÓN INDUSTRIAL DE LOS RESULTADOS

Durante los dos últimos años de realización de la Tesis, el autor ha sido contratado por Telefónica I+D en la División de Tecnología del Habla con el fin de aplicar directamente la investigación realizada en la Tesis, sobre los productos de tecnología del habla desarrollados por Telefónica I+D. Las principales aportaciones se pueden resumir en los siguientes puntos:

- **Desarrollo de un reconocedor de voz de palabras deletreadas por línea telefónica.** Una de las principales aportaciones de la Tesis ha sido el estudio de la tarea deletreo en castellano, la evaluación de diferentes estrategias de reconocimiento, y el desarrollo de una arquitectura de reconocimiento con un buen compromiso entre tasa de reconocimiento y tiempo de proceso (capítulo 3). El reconocedor de palabras deletreadas de Telefónica I+D ha sido desarrollado en su totalidad por el autor de esta Tesis, aprovechando así el trabajo de investigación realizado. Para el desarrollo de este reconocedor se ha seguido la arquitectura propuesta en la Tesis, consiguiendo reducir sensiblemente los tiempos de desarrollo. En el caso del reconocedor de palabras deletreadas de Telefónica, se ha conseguido mejorar los resultados de la presente Tesis gracias a las grandes bases de datos de voz propiedad de Telefónica I+D.
- **Desarrollo del módulo de estima de confianza para el reconocedor fonético de Telefónica I+D (habla aislada) y la realización de una primera versión para el reconocedor de Lenguaje Natural de Telefónica I+D (habla continua).** Estas versiones también han sido desarrolladas íntegramente por el autor de la Tesis. Para su realización se ha basado en los experimentos realizados sobre el sistema CU Communicator de la Universidad de Colorado (Estados Unidos) y que incluyen el estudio de tres niveles de confianza: palabra, frase y concepto semántico. Estos experimentos forman el capítulo 5 de la presente Tesis Doctoral.
- **Consultoría para la realización de servicios orientados al gran público: Servicio de Información Telefónica 1003 y Portal de voz de Telefónica Móviles "Emoción Voz".** La metodología de desarrollo de servicios automáticos por teléfono, descrita en el capítulo 6 de la Tesis, ha servido como propuesta inicial de trabajo para el desarrollo de servicios automáticos de gran impacto como son el Servicio de Información Telefónica 1003 y el Portal de voz "Emoción Voz" con varios miles de llamadas diarias. Si bien, es justo poner de relieve que el potencial económico de una empresa como Telefónica ha permitido incorporar nuevas etapas de diseño y evaluación en el desarrollo de estos sistemas, dando lugar a una metodología de diseño más completa.

TRANSFERENCIA DE TECNOLOGÍA

Una característica importante de esta Tesis ha sido que parte de los trabajos de investigación se han realizado en Estados Unidos en **"The Center of Spoken Language Research"** de la Universidad de Colorado. Concretamente han sido los trabajos sobre medidas de confianza de reconocimiento presentados en el capítulo 5. Estos experimentos se han realizado sobre el sistema **CU Communicator** que es un sistema automático que permite realizar reservas de viajes en avión, estancias en hoteles y alquiler de coches. Este sistema, así como el proyecto DARPA en el que estaba inmerso, ha constituido un referente a nivel mundial para los sistemas automáticos con reconocimiento de voz por teléfono.

La investigación se ha realizado durante 2 estancias de verano (3 meses cada una) en dicho centro, así como en los meses anteriores y posteriores a la realización de las estancias, trabajando desde España.

La realización parcial de la Tesis en Estados Unidos ha permitido transferir a España (y posteriormente a empresas españolas) tecnología desarrollada en este país con un mayor potencial investigador.