7.1 Conclusiones

El trabajo realizado en esta tesis doctoral se ha centrado en tres de los aspectos más importantes para la implementación de Servidores Vocales Interactivos: reconocimiento automático del habla, análisis de medidas de confianza para los módulos de reconocimiento y compresión de lenguaje natural, y el diseño de la gestión del diálogo.

Una primera aportación del trabajo ha sido el estudio bibliográfico realizado. En relación con el reconocimiento de nombres deletreados se han analizado y comparado las principales arquitecturas de reconocimiento propuestas al nivel internacional. En medidas de confianza hemos propuesto una clasificación de los parámetros a considerar según su origen, y una clasificación de las medidas de confianza según su nivel de aplicación: palabra, concepto, frase e interacción. En relación con la gestión de diálogo se han introducido las principales estrategias de diseño, y en el apartado 6.2, se ha realizado una descripción detallada de los modelos de diálogo utilizados actualmente al nivel internacional. Esta descripción permite entender mejor los conceptos más importantes en la gestión de diálogo que se manejan a lo largo del capítulo 6. Por otro lado en el apéndice F, se ha hecho un esfuerzo de síntesis para recoger las principales recomendaciones a tener en cuenta en el diseño de gestores de diálogo.

En cuanto al módulo de reconocimiento, por un lado, se ha implementado el primer reconocedor de nombres deletreados en castellano con tasas de acierto comparables a los realizados en otros idiomas, y por otro lado, se ha desarrollado un reconocedor de habla continua para frases que expresan fechas y horas. Ambos sistemas han sido diseñados y entrenados para voz por línea telefónica e independiente del locutor.

En relación con el análisis de medidas de confianza, se ha trabajado fundamentalmente sobre el sistema DARPA Communicator desarrollado en el Centro de Investigación de Lenguaje Hablado (CSLR: The Center for Spoken Language Research) de la Universidad de Colorado (Boulder) en Estados Unidos (consultar apéndice A, para ver una descripción más detallada). Además, se han estudiado medidas de confianza para los sistemas de reconocimiento desarrollados en la presente tesis.

Sobre el módulo de gestión de diálogo se ha definido una metodología de diseño que se presenta mediante su aplicación en el desarrollo de un sistema de información y reserva de billetes de tren por teléfono.

A continuación se presentan las principales conclusiones obtenidas de cada uno de los aspectos analizados.

7.1.1 Reconocedor de nombres deletreados

El primer aspecto tratado en este capítulo ha sido el análisis de la tarea de deletreo en castellano. A tenor de los estudios realizados, podemos concluir que los castellano–hablantes no estamos acostumbrados a deletrear debido a que hay una correspondencia directa entre la escritura y la pronunciación de una palabra. Dicha correspondencia hace

que el proceso de deletreo no sea necesario para desambiguar posibles escrituras de una misma pronunciación. Esta falta de hábito se manifiesta en el habla en tres aspectos importantes: realización de pausas largas entre las pronunciaciones de las letras y de gran variabilidad, incorporación de gran cantidad de ruidos cometidos por el locutor (pausas rellenas, falsos comienzos, respiraciones fuertes,...), e introducción de errores en la secuencia de letras pronunciada.

Tras un análisis de diferentes arquitecturas de reconocimiento se ha propuesto una estrategia basada en dos etapas: hipótesis y verificación, como estructura con mejor compromiso entre tasa de reconocimiento y tiempo de procesado. En una primera fase (hipótesis), se realiza una selección de los nombres del diccionario que más se parecen a la secuencia pronunciada por el locutor, para después, realizar con estos nombres candidato, un reconocimiento más potente que ofrezca el resultado definitivo.

En el análisis realizado sobre la etapa de hipótesis podemos resaltar las siguientes conclusiones:

- Para hacer frente a la gran cantidad de pausas entre letras y la variabilidad de su longitud se ha propuesto una nueva topología de HMM con modelos de silencio contextuales a los modelos de letra. Con esta solución se ha conseguido una mejora absoluta de 6,6 puntos (21,9% relativa) en la Tasa de Error de letra en comparación con utilizar un único modelo de silencio.

- Por otro lado, para modelar los posibles ruidos existentes en la señal de voz (tanto producidos por el locutor como por el canal de comunicaciones o el entorno), se han entrenado modelos acústicos específicos para estos ruidos consiguiendo reducir el error de letra otros 4,5 puntos (de 23,8% a 19,3%), obteniendo una Tasa de Error de 8,0% al nivel de nombre. A tenor de las grandes mejoras conseguidas podemos concluir que las soluciones propuestas en estos dos puntos se ajustan muy bien a los problemas detectados en el proceso de deletreo, y además, son técnicas complementarias entre sí.

- Por último, para hacer frente a los errores de deletreo cometidos por el locutor se han incorporado modelos de lenguaje de tipo N-gram (2-gram y 3-gram) con el fin de guiar el proceso de decodificación de la señal de habla. Estos modelos se han obtenido utilizando los nombres del diccionario considerado. Otro aspecto importante, ha sido la generación de las N mejores cadenas de letras en lugar de una única secuencia, para hacer más robusta la selección de nombres del diccionario. En este punto ha tenido una relevancia especial la generación de un grafo de letras que nos ha permitido incorporar estas técnicas (modelos de lenguaje 3-gram y generación de las N-mejores cadenas de letras), con un coste computacional reducido, permitiendo el funcionamiento del sistema en tiempo real.

Como configuración final de esta etapa de hipótesis se ha considerado la incorporación del modelo 3-gram y la generación de las 2 mejores cadenas de letras. En este caso, se ha obtenido un error de letra de 11,9% y un error de nombre de 5,8% para un diccionario de 1.000 nombres.

En la etapa de verificación se utiliza una gramática muy restrictiva generada con los M mejores candidatos propuestos por la etapa de hipótesis, y se vuelve a realizar un proceso de reconocimiento sobre dicha estructura para conseguir el nombre finalmente reconocido. En esta etapa se ha presentado el análisis realizado para calcular M (número de candidatos a considerar) con un buen compromiso entre tasa de reconocimiento y tiempo de procesado. Para la evaluación final del sistema completo se han utilizado diccionarios de 1.000, 5.000 y 10.000 nombres, consiguiendo tasas del 96,3%, 92,8% y 90,3% respectivamente. Estos resultados son comparables a los obtenidos para sistemas similares en otros idiomas como en inglés (Junqua, 1997).

Finalmente, se ha demostrado la utilidad del sistema desarrollado incorporándole en un servicio de información telefónica. En la evaluación de campo realizada se comprobó que gracias a este sistema se pudo aumentar un 49% la tasa de llamadas atendidas automáticamente.

7.1.2 Reconocedor de habla continua para fechas y horas

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 utilizado, 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 disponibles.

- 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 y las diferentes formas de postprocesar el grafo. Este mismo algoritmo ha sido el utilizado para el reconocedor de nombre deletreados, aunque el análisis en detalle se ha realizado sobre el reconocedor de fechas y horas.

- 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. Para el caso del habla espontánea se ha conseguido un error de 23,0% al nivel de palabra. Este error es menor al obtenido con el reconocedor Sphinx en el sistema CU Communicator (27,2%), aunque los

resultados no son directamente comparables. En primer lugar, utilizamos un diccionario de reconocimiento mucho más pequeño 400 palabras frente a las 1.800 utilizadas en CU Communicator. En segundo lugar, disponemos de menor cantidad de datos para entrenar los modelos tanto acústicos (3 horas frente a las más de 100 horas de voz utilizadas en Sphinx) como lingüísticos (1000 frases frente a las más de 25.000 utilizadas en CU Communicator).

7.1.3 Análisis de medidas de confianza

En relación con el análisis de medidas de confianza hemos trabajado en primer lugar sobre el sistema CU Communicator, en el que se han presentados estudios independientes para los niveles de palabra, concepto semántico y frase completa. Por otro lado, también se han realizado análisis para los reconocedores desarrollados en la presente tesis, centrándonos en los niveles de frase para el sistema de nombres deletreados, y en el nivel de palabra para el reconocedor desarrollado en el dominio de fechas y horas. 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, los parámetros obtenidos del modelo de lenguaje funcionan mejor para tasas de RI bajas. Combinando los parámetros del proceso de decodificación 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.

- 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 decodificadores. 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 decodificación, consiguiendo rechazos correctos del 58% de errores de reconocimiento y del 68% de nombres deletreados no pertenecientes al diccionario, para Rechazos Incorrectos 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 habla 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.

7.1.4 Metodología de diseño de gestores de diálogo

La principal aportación de esta tesis en la gestión del diálogo ha sido la propuesta de una metodología de diseño en la que se combina información de diferentes fuentes: análisis de base de datos, observación de conversaciones reales, simulación del servicio y funcionamiento con usuarios reales. La presentación de esta metodología se ha realizado sobre su aplicación al caso de un servicio de información y reserva de billetes de tren.

Esta metodología está formada por 5 fases. En la primera fase se realiza un análisis de la base de datos con la información disponible para ofrecer el servicio, mediante la descripción del diagrama Entidad-Relación que ofrece una representación semántica de los datos. En la segunda etapa “diseño por intuición”, proponemos la técnica de “braim-storming”, realizada sobre el diagrama Entidad-Relación, con la idea de plantear diferentes opciones de diseño en todos los aspectos relacionados con la gestión del diálogo.

En la fase de diseño por observación, es necesario grabar conversaciones entre los usuarios y operadores humanos. Con estas grabaciones se evalúan las alternativas sugeridas en la etapa anterior sin haber realizado aún ninguna implementación del sistema. El problema que surge de este análisis es que las interacciones usuario–operador son diferentes de las interacciones usuario–sistema, luego podremos hacer estudios sobre aspectos generales pero no sobre detalles muy concretos de la interacción.

En la fase de simulación utilizamos la herramienta de Mago de Oz para simular una interacción usuario–sistema. En la presente tesis se describe la utilización de esta técnica y los procedimientos utilizados para probar y evaluar diferentes estrategias de diálogo con usuarios reales. En esta simulación, el aspecto más importante es dotar al Mago de herramientas para que su tiempo de respuesta sea similar al de un sistema automático.

En la etapa de mejora iterativa se describe la utilización de medidas de confianza para la gestión y el diseño de los mecanismos de confirmación. Además, se presentan recomendaciones para el diseño de las frases de confirmación y se proponen medidas de evaluación para los mecanismos definidos. En esta etapa, también se describe una técnica de modelado del usuario, basada en niveles de destreza. A través de esta técnica se consigue adaptar el funcionamiento del sistema a la habilidad demostrada por el usuario a lo largo de su interacción.

Considerando los resultados de evaluación presentados en el apartado 6.4.6, podemos concluir que la metodología propuesta nos ha permitido implementar un sistema con buena aceptabilidad por parte de los usuarios. Los usuarios dieron un valor medio de 3,0 en una escala de 1 a 5.

7.2 Líneas futuras de trabajo

En este apartado se describirán las principales líneas futuras de trabajo en cada uno de los aspectos analizados en esta tesis doctoral. En primer lugar describiremos las líneas propuestas para el sistema de reconocimiento de nombres deletreados.

- Una primera línea de trabajo sería realizar un estudio más exhaustivo del modelado acústico utilizado: análisis de diferentes criterios para fijar el número de mezclas por estado ó la consideración de un entrenamiento discriminativo para reforzar las diferencias entre los modelos de letras con gran parecido acústico.

- Otro aspecto susceptible de un mayor análisis es el algoritmo de comparación de las secuencias de letras con los nombres del diccionario. En este punto se podrían evaluar nuevos criterios o mecanismos para definir las penalizaciones de letras y se podría evaluar el impacto de considerar penalizaciones con información contextual para aprender mejor los errores generados cuando se incorporan modelos de lenguaje en la obtención de la secuencia de letras. En este caso es necesario disponer de mayor cantidad de datos para que estas penalizaciones queden correctamente entrenadas.

- En la etapa de verificación se puede probar a utilizar modelos acústicos diferentes a los utilizados en la etapa de hipótesis: modelos con mayor resolución ó modelos entrenados con criterios discriminativos que puedan hacer frente a la gran similitud que con frecuencia aparece entre los nombres candidato de la etapa de hipótesis.

- Generalmente el reconocimiento de nombres deletreados se utiliza como apoyo a un reconocedor de nombres de gran vocabulario. Una línea futura interesante es proponer estrategias para combinar las salidas de ambos reconocedores con el fin de mejorar la tasa final obtenida. En este punto las medidas de confianza de uno y otro reconocedor se pueden utilizar como heurístico de combinación.

En relación con el reconocedor de habla continua para fechas y horas se proponen los siguientes trabajos futuros.

- Evaluación de mayor número de criterios para realizar el entrenamiento selectivo: relación señal a ruido de los ficheros o la existencia de cierto tipo de ruidos.

- Ampliar la potencia de modelado utilizando modelos continuos en los que se evalúe, en tasa de reconocimiento y tiempo de proceso, diferentes criterios de compartición de parámetros: compartición de estados completos entre diferentes modelos o compartición de gaussianas entre estados que pertenecen al mismo o a distintos modelos.

- Incorporación de la técnica de Búsqueda en Haz (Beam Search) para reducir el tiempo de proceso. Esta técnica es muy importante si se consideran modelos más potentes, como los modelos continuos, que ralentizan el proceso de descodificación.

- Otro aspecto importante es la obtención de mayor cantidad de datos de habla espontánea que permitan obtener modelos acústicos adaptados a este tipo de habla. En este punto será interesante evaluar el funcionamiento de los modelos acústicos entrenados con habla espontánea en comparación con técnicas de adaptación que permitan generar modelos de habla espontánea a partir de modelos entrenados con habla leída.

- En la línea de utilizar más datos, otro punto importante es la generación de modelos de lenguaje con una mayor cantidad de frases que permitan sacar conclusiones más relevantes acerca del tipo de suavizado a utilizar.

- Otro aspecto relacionado con los modelos de lenguaje es el estudio de técnicas para la generación automática de clases a utilizar en el modelo. Un posible criterio es por similitud contextual.

Sobre el análisis de medidas de confianza proponemos las siguientes líneas de investigación futuras:

- Análisis de parámetros obtenidos de la interacción o diálogo usuario–sistema para su utilización en la obtención de medidas de confianza en los tres niveles descritos en esta tesis: palabra, concepto y frase completa, así como la propuesta de un nuevo nivel de análisis: nivel de interacción. Las medidas propuestas a este nivel permitirán rediseñar o adaptar la gestión del diálogo según la calidad que está teniendo dicha interacción.

- Estudio de nuevos parámetros que puedan aportar información complementaria a los propuestos en esta tesis como por ejemplo las duraciones de los fonemas.

- Evaluación de diferentes técnicas para la selección y combinación de parámetros de confianza como el LDA (Linear Discriminant Analysis) o los árboles de decisión y su comparación con las técnicas basadas en Redes Neuronales.

- Utilización de medidas de confianza para combinar diferentes hipótesis de reconocedores que utilicen varias parametrizaciones diferentes para hacer frente a diferentes condiciones acústicas.

- Análisis de los parámetros descritos y propuestas de nuevos, para la detección de palabras fuera del vocabulario de reconocimiento (OOV: Out of Vocabulary) en sistemas de habla continua.

En cuanto al diseño de gestores de diálogo las principales líneas de trabajo futuras son las siguientes:

- Incorporación de sistemas de reconocimiento de habla continua y sistemas de compresión con el fin de permitir cierta iniciativa mixta: que el usuario pudiera cambiar de objetivo del diálogo o aportar los datos en el orden que mejor le parezca. Estos módulos son especialmente importantes para el reconocimiento de la fecha que es donde mayores problemas tienen los usuarios.

- Aplicación y validación de la metodología presentada utilizando modelos de diálogo más potentes (árboles dinámicos de objetivos) para agilizar y dar más flexibilidad a la interacción usuario–sistema.

- Implementación de modelos de usuario más potentes que consideren mayor cantidad de aspectos modificables, mayor análisis de los eventos positivos y negativos a considerar en cada aspecto, y mayor variedad de niveles de destreza.

- Validación de la metodología presentada en otros servicios y dominios de aplicación.

- Análisis de la diferente aceptación por parte de los usuarios, al utilizar mensajes pre-grabados en lugar del conversor texto-voz para la síntesis de los mensajes emitidos por el sistema.