1.1 Servidores Vocales Interactivos

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 automátización 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.

El trabajo que se presenta en esta tesis pretende el análisis y mejora de varios de los aspectos a tener en cuenta en el diseño y desarrollo de un SVI. Es mucha la variedad de servicios que se pueden ofrecer a través de los SVIs: servicios de banca telefónica, reserva de billetes de transporte, reserva de entradas para espectáculos o sistemas de información de tráfico, información meteorológica o de números de teléfonos. 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 el apéndice A, se puede consultar una descripción más detallada de estos servicios.

En la figura 1-1, se muestra un esquema modular genérico de un SVI.

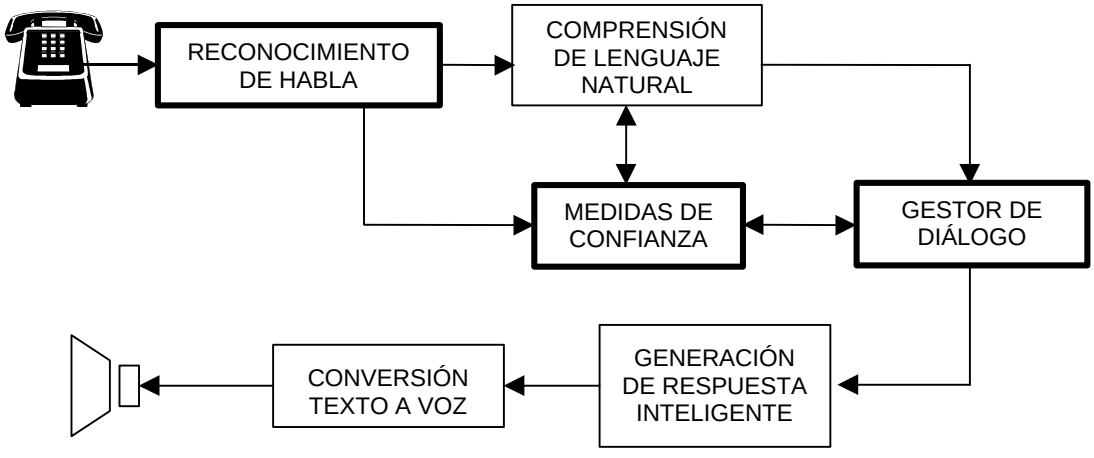

Figura 1-1: Diagrama de bloques de un Servidor Vocal Interactivo.

- El **módulo de reconocimiento** realiza la función de transcribir o descodificar la señal de habla, obteniendo a su salida una secuencia de palabras pertenecientes a un vocabulario de reconocimiento. En un servicio por teléfono, la tarea de reconocimiento debe ser independiente del locutor, de forma que se pueda atender a cualquier usuario de una misma lengua. Además, este módulo debe hacer frente a los ruidos e interferencias que pueda introducir la línea telefónica, así como a la reducción en el ancho de banda que sufre la señal. Otro factor importante que se debe tener en cuenta es el alto grado de espontaneidad que puede llegar a ofrecer un usuario en su interacción con el sistema proveedor del servicio. El habla espontánea frente al habla leída supone una dificultad adicional en su reconocimiento: aparecen una mayor cantidad de ruidos, imprecisiones al hablar, cambios bruscos de velocidad de locución, relajación en las construcciones gramaticales y efectos de co-articulación más pronunciados.

  Habitualmente un mismo SVI puede disponer de varios módulos de reconocimiento con características diferentes; en un mismo servicio se podría disponer de un sistema de reconocimiento de habla aislada y gran vocabulario para nombres propios, un sistema para el reconocimiento de nombres deletreados con el fin de diferenciar homófonos (ej: Jiménez y Giménez) o un sistema de reconocimiento de habla continua y vocabulario medio para dominios restringidos, como por ejemplo para el caso de fechas y horas. La utilización de uno u otro sistema vendrá determinada por el gestor de diálogo quien debe decidir, según la dificultad de la tarea, el sistema a utilizar para obtener la mayor tasa de acierto.

- La **comprensión de lenguaje natural** pretende la interpretación semántica de la secuencia de palabras obtenidas del reconocimiento (Ward, 1994; Colás, 1999). En esta interpretación, el objetivo es detectar los conceptos relevantes de la frase que hacen referencia al dominio de la aplicación o servicio ofrecido. De estos conceptos se pueden extraer datos importantes como la ciudad origen o destino de un viaje, en el caso de un servicio de reserva de billetes de viaje, o pueden hacer referencia a intenciones o partes del servicio deseadas en concreto, como por ejemplo si se quiere información de horarios de tren o si también se desea realizar la reserva del billete.

- El **gestor de diálogo** es el encargado de definir el flujo de la aplicación. Considerando las respuestas del usuario, la historia del diálogo y el estado actual del mismo, se deben definir tanto la acción a realizar: pedir un dato, confirmar un dato, dar información,... como su contenido: el dato a pedir o a confirmar y la información a ofrecer. El gestor del diálogo utilizará principalmente los conceptos extraídos del etiquetador semántico así como las medidas de confianza sobre la certeza de dichos conceptos.

- Las **medidas de confianza** juegan un papel importante en el desarrollo de un SVI robusto. Dado que estamos muy lejos de los sistemas perfectos, no sólo es necesario intentar reconocer o comprender lo que dice el usuario sino que además debemos saber la fiabilidad o calidad de lo reconocido/comprendido. El objetivo

de las medidas de confianza es intentar detectar posibles errores en la secuencia de palabras o conceptos que hagan que el sistema lleve la interacción hombre-máquina por caminos que divergen de los objetivos del usuario, y como consecuencia, le hagan sentirse frustrado e insatisfecho con el sistema.

- El módulo de **generación de respuesta** realiza la labor de construir las frases que debe producir el sistema. Partiendo de la acción a realizar, decidida por el gestor de diálogo, y del contenido de dicha acción, el generador de respuesta debe diseñar la frase con la que el sistema pedirá o confirmará un dato al usuario, o bien le ofrecerá la información solicitada.

- El **conversor de texto a voz** es el módulo encargado de convertir la respuesta del sistema a una señal de habla. Esta conversión se puede hacer reproduciendo ficheros concatenados, mediante síntesis de voz (Pardo et al,1995) o combinando ambas técnicas.

En esta tesis se ha realizado la mayor cantidad de trabajo en los módulos resaltados en la figura 1-1: 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. Este sistema pretende ser un apoyo a un sistema de reconocimiento de nombres con gran vocabulario. 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.

1.2 Objetivos de la Tesis

Los objetivos más importantes planteados en esta tesis se enmarcan en los módulos comentados anteriormente: reconocimiento del habla, obtención de medidas de confianza y diseño de la gestión del diálogo.

1.2.1 Módulo de reconocimiento del habla

Los objetivos perseguidos en relación con este módulo son los siguientes:

- **Desarrollo e implementación de un sistema de reconocimiento de nombres deletreados en castellano con tasas de acierto comparables a otros sistemas desarrollados para otros idiomas como inglés o francés.** Este objetivo lleva consigo la obtención de los siguientes subobjetivos:

1. Análisis de la tarea del deletreo para el castellano. Se analizarán los hábitos y comportamientos más frecuentes de los usuarios de lengua castellana a la hora de deletrear: estudio de posibles errores de elocución, diversidad de pronunciaciones para hacer referencia a las letras de nuestro alfabeto y análisis de la confusión entre los nombres de las letras que formarán nuestro diccionario de reconocimiento.

2. Validación de las técnicas de reconocimiento utilizadas en el caso del inglés para su aplicación en castellano y desarrollo de nuevas técnicas para hacer frente a la gran variedad de ruidos presentes en la señal de habla y a la variabilidad de pausas entre las letras.

3. Adaptación del mecanismo de generación de grafos de palabras propuesto por Ney (Ney, 1994), (Ney, 1999) al caso del deletreo donde las palabras a reconocer son las letras y donde no se aplicarán técnicas de Beam Search puesto que el tamaño del espacio de búsqueda en reconocimiento es reducido y de gran confusión.

- **Desarrollo e implementación de un sistema de reconocimiento de habla continua para dominios restringidos. El dominio elegido es el de fechas y horas.** Las acciones realizadas para conseguir este objetivo son las siguientes:

1. Análisis del dominio de fechas y horas con el fin de obtener tanto el vocabulario de reconocimiento como un análisis de la estructura lingüística que nos permita definir modelos de lenguaje. Estos modelos serán muy útiles para guiar el proceso de decodificación.

2. Estudio de la configuración de los HMMs más apropiados para el modelado acústico. Selección que se deberá realizar atendiendo a un compromiso entre potencia de modelado y datos disponibles para entrenar dichos modelos.

3. Adaptación y simplificación del algoritmo de generación de un grafo de palabras propuesto por Ney como segunda etapa de reconocimiento, para la aplicación de modelos de lenguaje potentes y la generación de las N-mejores secuencias de palabras con coste computacional reducido.

4. Análisis de la repercusión del habla espontánea frente al habla leída en la tasa de reconocimiento.

1.2.2 Medidas de confianza

Los objetivos referidos a las medidas de confianza son los siguientes:

- **Estudio de medidas de confianza tanto para el módulo de reconocimiento de habla continua como para el módulo de comprensión en el sistema CU Communicator.** Este sistema ofrece un servicio de información de viajes de avión, reserva de hotel y coches de alquiler desarrollado en la Universidad de

Colorado, Boulder (USA) (Ward y Pellom, 1999). Más detalles sobre este sistema se pueden consultar en el apéndice A.

- **Análisis de medidas de confianza para el sistema de reconocimiento de nombres deletreados por teléfono.** En este caso se analizarán diferentes parámetros para la detección tanto de errores en el reconocimiento, como de palabras fuera del vocabulario de reconocimiento.

- **Estudio de medidas de confianza para el sistema de reconocimiento de fechas y horas.** En este caso se analizará la potencia de las medidas de confianza para la combinación de hipótesis de reconocimiento obtenidas de diferentes decodificadores (ej: modelos acústicos independientes del sexo, modelos acústicos adaptados para hombres y otro con modelos adaptados para mujeres).

1.2.3 Gestión de diálogo

Las investigaciones en relación con la gestión del diálogo se realizaron sobre el entorno de desarrollo de aplicaciones telefónicas TADE (ver apéndice A, apartado A.2), desarrollado íntegramente en el Grupo de Tecnología del Habla. La aplicación concreta será un servicio de información sobre horarios y precios de los trenes para viajar entre dos ciudades españolas, así como la posible reserva del viaje. El objetivo perseguido en este caso es:

- **Propuesta de una metodología para el diseño de sistemas de gestión de diálogo**. En esta metodología se pretende, además, definir mecanismos para:

    1. Incorporar las medidas de confianza en la gestión de diálogo con el fin de diseñar las estrategias de confirmación de los datos.

    2. Definir un modelo de usuario a través del cuál, las preguntas e informaciones proporcionadas, se adecuen mejor a la destreza demostrada por el usuario en su interacción con el sistema.

1.3 Contenido de la Tesis

La presente tesis está estructurada en siete capítulos:

- **Capítulo 1. Introducción**. En este capítulo se comentan los módulos que constituyen un Servidor Vocal Interactivo, remarcando aquellos en los que se centran los trabajos de investigación realizados en la presente tesis así como los objetivos perseguidos con dichos trabajos.

- **Capítulo 2. Encuadre Científico-Tecnológico**. Este capítulo describe el marco tecnológico en el que se sitúa la tesis, analizando las técnicas más importantes que se están utilizando tanto en el reconocimiento de habla continua, la obtención de las medidas de confianza como en la gestión del diálogo en los Servidores

Vocales Interactivos. Esta descripción nos permitirá entender mejor las decisiones tomadas en la realización de este trabajo así como el mayor o menor alcance de las conclusiones obtenidas de él.

- **Capítulo 3. Sistema de reconocimiento de nombres deletreados**. Este capítulo presenta las características de la tarea de deletreo en castellano y describe la implementación de un sistema de reconocimiento de nombres deletreados con tasas comparables a otros idiomas. En esta descripción se detallan los trabajos de investigación realizados tanto para la adaptación y ajuste de técnicas empleadas en otros idiomas, como para la incorporación de nuevas ideas. Este sistema se evaluará en pruebas de laboratorio y en un servicio real de páginas blancas, en el que se utilizará como apoyo el reconocedor de habla aislada y gran vocabulario necesario para nombres y apellidos.

- **Capítulo 4. Sistema de reconocimiento en dominios restringidos: fechas y horas**. En este capítulo se describen los trabajos realizados para el desarrollo de un sistema de reconocimiento de fechas y horas. En este desarrollo se propone la utilización de la técnica de entrenamiento selectivo como mecanismo para evaluar la resolución de los modelos utilizados y su adecuación a la cantidad de datos de entrenamiento disponibles. Por otro lado, se propondrá una simplificación del algoritmo de Ney para la generación de grafos de palabras y se estudiará la repercusión del tipo de habla, leída o espontánea, sobre la tasa de reconocimiento.

- **Capítulo 5. Análisis de medidas de confianza en SVIs**. Los análisis de medidas de confianza realizados se centrarán en el servicio CU Communicator (ver apéndice A). En este análisis se utilizarán parámetros acústicos, gramaticales (modelo de lenguaje) y semánticos (sistema de comprensión) con el fin de definir medidas de confianza al nivel de palabra, concepto semántico, y frase que permitan una mejor gestión del diálogo. En este capítulo, también se extenderá este estudio a los reconocedores desarrollados a lo largo de esta tesis doctoral.

- **Capítulo 6. Metodología para el diseño del gestor de diálogo**. En este capítulo se detallan las fases llevadas a cabo para el diseño del gestor de diálogo en el desarrollo de un sistema de información y reserva de billetes de tren. Estas fases constituyen la metodología de diseño que se presenta y se propone en este trabajo de investigación. En cada una de las fases de diseño se describen los objetivos a alcanzar así como las medidas de evaluación propuestas para poder comparar las diferentes estrategias de diálogo. En este capítulo, también se describe la incorporación de las medidas de confianza para la gestión de las confirmaciones de los datos y se comenta el modelado de usuario implementado.

- **Capítulo 7. Conclusiones y líneas futuras**. Se concretan las conclusiones más importantes extraídas de este trabajo de investigación así como las líneas futuras que se pueden apoyar en esta tesis.

Además de los capítulos comentados, se ha introducido una sección adicional de apéndices complementarios donde se puede consultar información más detallada de algunas partes de la tesis para permitir una mejor compresión de las mismas.