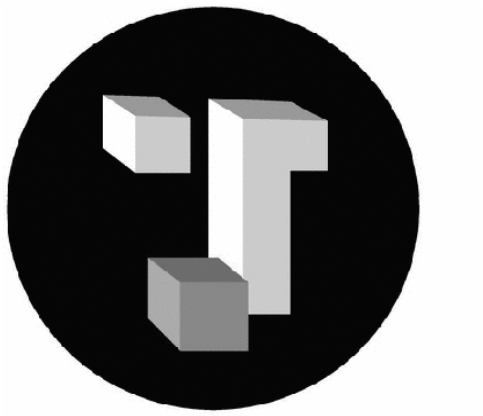
UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

RESUMEN DEL PROYECTO FIN DE CARRERA

Controlling a High Fidelity system from speech



AUTOR: Rafael García Sánchez
TUTOR: Fernando Fernández Martínez

2006

Índice

Capítulo 1

Descripción del proyecto

1.1. Motivación

En el ámbito domótico podemos definir el diálogo como un proceso de comunicación orientado a la consecución de determinados objetivos que responden a las necesidades de control sobre diversos aparatos electrónicos domésticos. El origen del proyecto documentado en [1], fue dotar al Departamento de Tecnología del Habla de un sistema de demostración que permitiera mostrar nuevas soluciones de gestión natural del diálogo mediante un sistema vocal interactivo independiente de locutor. Para llevar a cabo este proyecto, se partió del sistema detallado en [2] basado en reglas de traducción que carecía de diálogo hombre-máquina y entrenado para un único usuario.

1.2. Objetivos

A diferencia de los típicos sistemas de control basados en comandos simples pronunciados de forma aislada, el interfaz vocal que hemos querido desarrollar es un interfaz conversacional que permite a los usuarios controlar el sistema HiFi mediante frases habladas de manera natural. Los usuarios deben tener libertad para formular varias órdenes complejas a partir de una única frase. Por otra parte, no necesitan memorizar una lista de posibles comandos o una fraseología específica con las que poder controlar el sistema de manera satisfactoria. Hemos querido poder controlar un sistema HiFi comercial compuesto de un reproductor de CD con cargador para tres discos, un receptor de radio y un reproductor/grabador de casetes dotado de doble

pletina. Normalmente el sistema se controla mediante el uso de un control remoto infrarrojo (IR). En su lugar, los usuarios controlan el sistema mediante órdenes vocales a través de un micrófono. Las frases pronunciadas contienen la intención del usuario y son traducidas al conjunto de comandos IR necesarios para llevar a cabo una cierta acción sobre el sistema. Esta traducción se realiza de tal forma que el conjunto apropiado de comandos IR se envían modificando el estado del equipo según la intención del usuario. La interfaz se encarga también de mantener en memoria el estado actual del equipo. De esta forma se consigue un control casi total sobre el equipo. Únicamente no se tiene control de aquellas funciones que conllevan acciones físicas, como la carga de CD's, por la imposibilidad de ser controladas. En cualquier caso, el correcto control del equipo está supeditado a que sea la interfaz vocal el único que lo controla, ya que en cualquier otro caso, y al no existir comunicación desde el equipo a la interfaz, podría perderse el sincronismo y desconocerse su estado.

Con estos supuestos y utilizando los recursos disponibles en el Departamento se ha conseguido realizar este sistema que presentamos, un interfaz vocal para el control de un equipo de alta fidelidad usando órdenes o frases habladas de manera natural.

1.3. Desarrollo

El desarrollo del proyecto se realizó por módulos, al igual que la arquitectura de un sistema vocal interactivo, como puede verse en la figura 1.1.

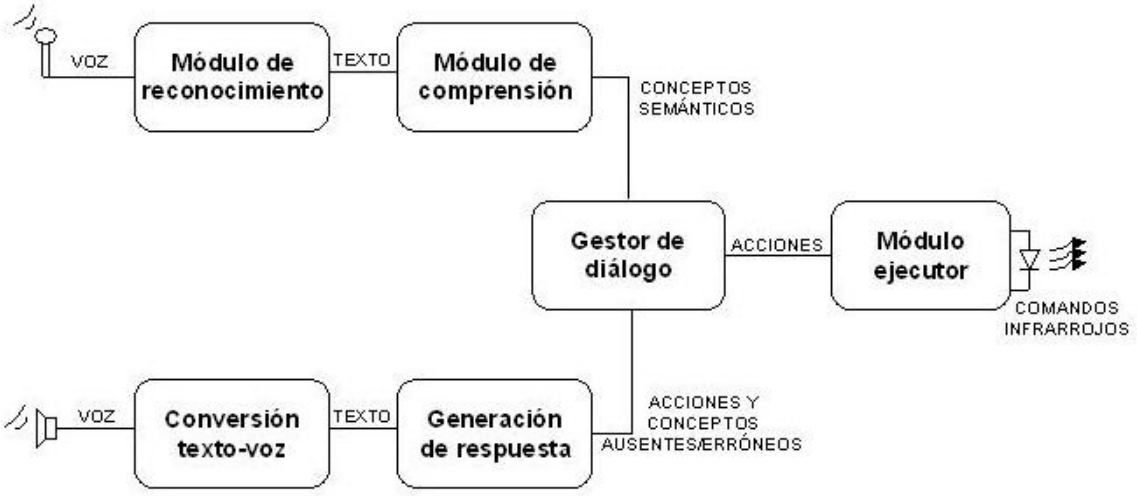


Figura 1.1 Arquitectura del sistema

1.3.1 Módulo de reconocimiento

Hubo un trabajo previo de estudio sobre sistemas de diálogo, posibles vocabularios empleado en el control de un equipo HiFi y diferentes tipos de reconocedores de voz. Con esta base, decidimos comenzar por integrar el módulo de reconocimiento de voz. Este módulo es el responsable de transcribir la señal acústica del micrófono a la secuencia de palabras reconocidas (la unidad mínima de reconocimiento es la palabra). Para llevar a cabo esta funcionalidad se necesitaba: el vocabulario de la aplicación (la lista de palabras que se podrían reconocer), la lista de conversión grafemas[1]-fonemas[2] (que permite conocer los fonemas que componen cada palabra de nuestro vocabulario), los tri-fonemas permitidos (el sonido de un fonema depende del anterior y del posterior, el fonema /a/ precedido del /k/ no es igual al precedido del /f/), los modelos acústicos de Markov (que representan cada unidad por 3 estados codificando la diferente duración de cada unidad mediante la auto-transición o la transición al siguiente estado), y por último una gramática que introduce las restricciones sintácticas a nuestro reconocedor, que modela el conocimiento gramatical que incorporamos en el proceso de reconocimiento. Usamos una gramática estocástica, más concretamente de tipo *N-gram*, donde la probabilidad de que se produzca una palabra depende de las palabras precedentes. Para simplificar los cálculos, optamos por una gramática del tipo bigrama (*2-gram*), donde codificamos las probabilidades de dos palabras consecutivas.

Tras el análisis de diferentes situaciones de control del equipo y con la ayuda de un lingüista experto, definimos los vocabularios y diccionarios de nuestra aplicación. En base a estos diccionarios, se generaron 400 frases de entrenamiento, que sirvieron para obtener las probabilidades que definirían nuestra gramática del tipo bigrama. De esta manera, ha sido posible realizar un reconocedor independiente de locutor, esto quiere decir, que es capaz de transcribir la señal acústica producida por cualquier usuario a un texto reconocido.

[1] Grafema: unidad mínima e indivisible de la escritura de una lengua.

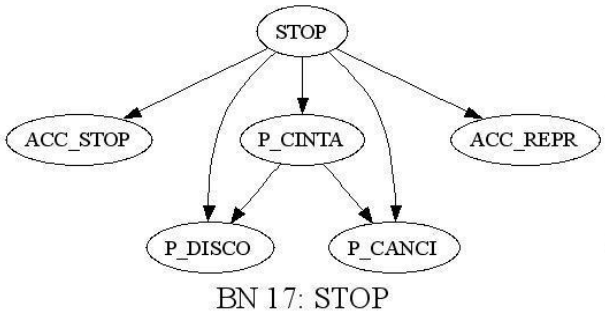
[2] Fonema: cada uno de los sonidos diferenciables de una lengua.

1.3.2 Módulo de comprensión

Este módulo tiene la misión de extraer los conceptos semánticos relevantes a partir de la frase reconocida. El conocimiento experto en el dominio de la aplicación permitió definir un total de 70 categorías semánticas que pueden clasificarse como: acciones (e.g. reproduce), aparatos (e.g. reproductor de cd), parámetros (e.g. volumen), y valores (e.g. cinco). Con objeto de refinar el etiquetado de cada palabra se generaron un conjunto de reglas dependientes de contexto que eliminan la ambigüedad relativa a su significado específico teniendo en cuenta el contexto en que aparecen en la frase. Así, obtenemos los conceptos semánticos más relevantes presentes en cada frase.

1.3.3 Módulo de gestión de diálogo

Como objetivos de diálogo definimos cada una de las posibles acciones (20 en total) que pueden realizarse sobre el sistema. Mediante las *belief networks* (BN) usando el procedimiento de inferencia directa se identifican los objetivos de diálogo presentes en la frase conforme a la intención del usuario a partir de los conceptos semánticos extraídos.



P(STOP)

on	off
0.5	0.5

P(ACC_STOP|STOP)

STOP	on	off
on	0.9	0.1
off	0.1	0.9

P(P_CINTA|STOP)

STOP	on	off
on	0.5	0.5
off	0.4	0.6

P(P_DISCO|STOP,P_CINTA)

STOP	P_CINTA	on	off
on	on	0.5	0.5
on	off	0.9	0.1
off	on	0.25	0.75
off	off	0.9	0.1

P(P_CANCI|STOP,P_CINTA)

STOP	P_CINTA	on	off
on	on	0.5	0.5
on	off	0.9	0.1
off	on	0.25	0.75
off	off	0.9	0.1

P(ACC_REPR|STOP)

STOP	on	off
on	0.5	0.5
off	0.9	0.1

Figura 1.2 Ejemplo de topología y probabilidades que codifican una BN

Como ejemplo de BN tenemos la figura 1.2, donde se muestran la relación entre los conceptos semánticos y el objetivo STOP, y la probabilidad de que el objetivo STOP esté activo si en nuestra frase aparecen o no cada uno de esos conceptos semánticos. Además, para cada objetivo inferido es posible detectar, mediante el procedimiento de inferencia inversa (calculando la probabilidad de que aparezca en la frase cada concepto semántico si el objetivo STOP está activo), qué conceptos han sido omitidos para solicitarlos al usuario, y cuáles son erróneos u opcionales para tratar de resolverlos. Por ejemplo, si el usuario dice “párala” se activa el objetivo de STOP, y nuestro sistema será capaz de detectar que le falta el parámetro que defina la acción, es decir qué es lo

que se quiere parar, la cinta, la canción o el disco. Para resolverlo, el sistema hará uso de la memoria de diálogo que hemos implementado. Ésta consta del estado del equipo (si el equipo está reproduciendo la cinta, se entenderá que se quiere parar la cinta), y de los conceptos semánticos dichos recientemente (si el usuario ha dicho recientemente algo referido a la cinta, o al CD, es lógico pensar que se quiere parar uno de estos elementos). Si por el contrario, nuestra memoria de diálogo es incapaz de resolver esta ambigüedad, el gestor de diálogo será responsable de realizar una pregunta al usuario para obtener los conceptos omitidos por el usuario.

A continuación se completa la interpretación de la frase rellenando un conjunto de marcos semánticos, uno por cada acción, que son enviados al módulo de ejecución. Nuestros marcos semánticos son muy simples, de manera general se componen de tres ranuras: un aparato sujeto de la acción (e.g. cd), un parámetro de ese aparato a controlar (e.g. pista), y un valor que deseamos adopte el parámetro indicado (e.g. uno). Siguiendo esta filosofía, hemos definido y ajustado las BN que mejor codificaban la comunicación de un usuario y el sistema.

1.3.4 Módulo de ejecución

A partir de los marcos de ejecución, este módulo interpreta las diferentes acciones determinando el conjunto de comandos IR que deben ser enviados secuencialmente al sistema Hifi para llevar a cabo la acción deseada. Para ello hemos integrado un dispositivo hardware vía USB de envío/recepción de comandos IR que sustituye la funcionalidad del mando a distancia del equipo.

1.3.5 Módulo de generación de respuesta

Este módulo junto con el de conversión texto-voz permite comunicar al usuario en todo momento las interpretaciones y acciones que realiza el sistema. Hemos optado por una solución basada en plantillas conceptuales, que permite la substitución aleatoria de cada concepto por una expresión correcta dentro de un conjunto de alternativas con la misma carga semántica. De esta manera obtenemos una mayor naturalidad del diálogo.

1.3.6 Módulo de conversión texto-voz

Es el encargado de sintetizar los mensajes de texto generados por el módulo de respuesta con objeto de facilitar una realimentación de información útil para los usuarios. Para ello, hemos integrado el sintetizador BORIS [3], desarrollado por el "Grupo de Tecnología del Habla" de la ETSIT (UPM).

1.4. Conclusiones

El trabajo de este proyecto final de carrera ha permitido crear un interfaz vocal para el control de un equipo HiFi en el que se ha ampliado notablemente el vocabulario que se puede reconocer, que junto con un módulo de reconocimiento independiente de locutor, permite ser utilizado por cualquier persona sin necesidad de ser previamente entrenado. La creación de nuevas reglas dependientes de contexto para la comprensión y la categorización semántica más exacta de ciertas palabras, permite que nuestro sistema extraiga la información relevante más relacionada con la intención del usuario. De esta manera, y junto con la estrategia de diálogo, se amplía el número de giros lingüísticos que nuestro sistema interpreta correctamente. Esa estrategia de diálogo basada en el uso de las BN, permite que nuestro sistema y el usuario tengan un diálogo más natural, siendo capaz la aplicación de extraer toda la información necesaria para actuar sobre el equipo HiFi. La integración del sistema de ejecución acorde con los comandos estándar del mando a distancia y el uso de un dispositivo hardware USB, permite que este sistema pueda ser utilizado para controlar otros equipos HiFi realizando pocos cambios, e incluso siguiendo la filosofía del proyecto, con un pequeño trabajo extra se permitiría controlar cualquier equipo que disponga de un interfaz infrarrojo.

En resumen, se ha desarrollado un sistema que cualquier persona desde un inicio puede controlar el equipo mediante la voz de la forma más natural posible, con las consiguientes aplicaciones prácticas para personas discapacitadas y para entornos domóticos.

Capítulo 2

Originalidad del tema

Después de cinco años de asignaturas teóricas y prácticas a lo largo de la carrera de ingeniería de telecomunicaciones, el alumno necesita retos nuevos. Este proyecto crea un interfaz vocal que permite la comunicación hombre-máquina basada en el diálogo. El habla es el modo de comunicación más natural para los seres humanos, pero actualmente es necesario aprender complejos códigos y mecanismos de interactividad cuando se quiere utilizar cualquier sistema. Sin embargo, este interfaz compatibiliza el diálogo natural de un ser humano con la secuencia de comandos de control remoto de un equipo de música, siendo el interfaz capaz de pedir al usuario que especifique su información si es necesario, surgiendo así un diálogo eficaz y flexible entre el interfaz y el usuario.

Aunque actualmente existen sistemas de reconocimiento de voz para ayudar al usuario en la utilización de diferentes entornos domóticos, este proyecto final de carrera pretende ir un poco más lejos. Ya que nuestro sistema se basa en un diálogo natural con el usuario, en lugar de un interfaz de comandos vocales que caracterizan la gran mayoría de los interfaces actuales. De esta manera el usuario no necesitará de ningún aprendizaje previo a la hora de comunicar sus intenciones sobre el equipo, con su consiguiente facilidad de uso.

El interfaz vocal es una herramienta que permite simplificar el funcionamiento del equipo mediante la interpretación del lenguaje natural de los seres humanos. Con este principio, su desarrollo requiere de la comprensión de muchos elementos que no son intuitivos y en cambio novedosos para el estudiante, ya que debe adaptar los

conceptos de un sistema de comunicación a este caso particular. Además hay que destacar el aprendizaje y definición de una metodología de diseño fácilmente extensible a otros dominios o ámbitos de control. Por otra parte, el diseño del sistema basado en la arquitectura modular y la correcta definición de la información intercambiada entre estos módulos, hace muy fácil la mejora o sustitución de un módulo concreto.

El proyecto demuestra la aplicación práctica de este tipo de sistemas para el uso concreto del HiFi, pero gracias a la metodología diseñada cabe la posibilidad de extender el sistema al control de nuevos aparatos dentro del entorno domótico. Si pensamos que el sistema de comunicación entre el interfaz y el equipo HiFi está basado en un sistema de comandos infrarrojos mediante puerto USB, con una pequeña modificación seríamos capaces de controlar otros aparatos que tienen mando a distancia.

Capítulo 3

Resultados obtenidos

Después de varios meses de trabajo es importante resaltar los resultados obtenidos con este proyecto.

El primer gran resultado es que se ha desarrollado un interfaz vocal con capacidad de entablar un diálogo con el usuario con el fin de obtener la información necesaria para interaccionar sobre un equipo de alta fidelidad. De esta manera cualquier persona sin conocimientos previos del sistema será capaz de controlar un equipo HiFi. De esta manera se demuestra la posibilidad de realizar este tipo de sistemas dentro de un entorno domótico.

Los resultados obtenidos en cada uno de los módulos ya se ha comentado en apartados anteriores, pero destacamos como más importantes los siguientes:

- Se ha ampliado notablemente el vocabulario que se puede reconocer, que junto con la utilización de un módulo de reconocimiento independiente de locutor, permite ser utilizado por cualquier persona sin realizar ningún tipo de ajuste previo.

- La creación de nuevas reglas dependientes de contexto para la comprensión, ampliamente probadas y diseñadas por expertos conocedores del dominio de la aplicación, y la categorización semántica más exacta de ciertas palabras, permite que nuestro sistema extraiga la información relevante más

relacionada con la intención del usuario, permitiendo expresiones comunes y simples hasta más complejas.

- El módulo responsable de la gestión del diálogo ha sido diseñado mediante BN para permitir usar técnicas de aprendizaje automático con la posibilidad de incorporar el diseño manual usando el conocimiento experto de las topologías, permitiendo así en un futuro desarrollar este tipo de aplicaciones sin la necesidad de destinar grandes recursos a su realización.

- La integración del sistema de ejecución acorde con los comandos estándar del mando a distancia y el uso de un dispositivo hardware USB, permite que este sistema pueda ser utilizado para controlar otros equipos HiFi realizando pocos cambios. Además, siguiendo la metodología de este proyecto, se podría controlar cualquier equipo que disponga de un interfaz infrarrojo con un mínimo esfuerzo de modificación.

En todas las demostraciones realizadas del sistema, se ha comprobado su versatilidad a la hora de interactuar diferentes usuarios sobre el equipo de alta fidelidad. Durante la última parte del proyecto se realizaron pruebas por parte de agentes externos para comprobar la fiabilidad del sistema, comprobando su buen funcionamiento en situaciones no previstas con anterioridad.

Por lo tanto, el sistema, como ejemplo y prueba de todo lo que se puede realizar en el entorno de los reconocedores de voz y de gestión de diálogo, superó con creces las previsiones iniciales. Cabe destacar que de este proyecto final de carrera se realizaron dos publicaciones, una en dentro de la*"9th Conference on Speech Communication and Technology (INTERSPEECH 2005)"* en Lisboa [4] y otra en las jornadas del XXI Congreso SEPLN [5].

Capítulo 4

Aplicabilidad práctica

Es fácil imaginar la importancia de este proyecto como demostración de todos los sistemas que se pueden crear a partir de esta idea.

Pero en concreto, este proyecto puede ser entendido como ayuda a personas con ciertas discapacidades en las que su única manera de comunicación puede ser la voz, a esto le debemos sumar las nuevas tendencias en entornos domóticos, donde la voz cobra gran importancia. La facilidad de uso del sistema demostrada en las jornadas del XXI Congreso SEPLN [5] le dan al sistema potencialmente un amplio abanico de personas destinatarias, con su consiguiente aplicabilidad práctica.

La demostración de este sistema, con el amplio público destinatario, indica la posibilidad que tienen los sistemas vocales interactivos en un futuro para controlar de una manera eficiente y lo más natural posible todos los equipos del hogar, evitando sus complejos modos de funcionamiento y los diferentes sistemas de control de cada uno. De la misma manera que hemos sido capaces de desarrollar este sistema, con los conocimientos adquiridos y los modos automáticos de generación de bn's, es fácil expandir la funcionalidad del sistema al control de otros equipos sin la necesidad de grandes recursos.

También es importante señalar la aplicación de este sistema como demostración de conceptos teóricos estudiados en el "Grupo de Tecnología del Habla" de la UPM. Como ejemplo de la aplicación de conceptos consolidados y banco de pruebas para el estudio de nuevas ideas es una herramienta útil y versátil. Al igual que en este proyecto

se han estudiado ciertas soluciones para diferentes problemas como es el diálogo, en un futuro modificando cada uno de los módulos que compone el sistema, se pueden realizar pruebas de nuevas soluciones.

Pero hay que señalar que una vez vista una demostración del sistema, es más fácil hacerse una idea del amplio abanico de posibles aplicaciones prácticas que el trabajo de este proyecto puede dar lugar.

Bibliografía

[1] R. García Sánchez, *"Controlling a High Fidelity system from speech"*. Proyecto final de carrera de la ETSIT (UPM), Madrid 2005.

[2] O. García Toledo, *"Optimización de un reconocedor de voz y de un sistema de comprensión del habla mediante un ordenador personal"*. Proyecto final de carrera de la ETSIT (UPM), Madrid 2002.

[3] Ricardo de Córdoba Herralde, José Manuel Pardo Muñoz, Juan Manuel Montero Martínez y Juana María Gutiérrez Arriola, *"Boris. Conversor texto voz del GTH"*. Registro de la Propiedad Intelectual (Madrid): M-001714/2003. Grupo de Tecnología del Habla (UPM), 1998.

[4] F. Fernández, J. Ferreiros, V. Sama, J.M. Montero, R. San Segundo, J. Macías, R. García, *"Speech interface for controlling an Hi-fi Audio system based on a bayesian belief networks approach for dialog modelling"*. 9th Conference on Speech Communication and Technology (INTERSPEECH 2005), pp.3421-3424, Lisboa (Portugal), ISSN: 1018-4074, Sep. 2005.

[5] F. Fernández, J. Ferreiros, V. Sama, J.M. Montero, R. García, *"Demostración de una interfaz vocal para el control de un sistema de alta fidelidad"*. Procesamiento del lenguaje natural Nº 35, pp. 451-452, ISSN 1135-5948, Sep. 2005.

Apéndice

A.1. Financiación obtenida

Como información relevante que demuestra la valía de este proyecto, hay que destacar que los últimos meses del proyecto final de carrera fueron financiados por el "Departamento de Tecnología del Habla" al alumno firmante como recompensa del gran trabajo desarrollado.

A.2. Trabajos publicados

Además, de este proyecto final de carrera se realizaron dos publicaciones, una de ellas internacional y la otra nacional:

- La primera de ellas llamada *"Speech interface for controlling an Hi-fi Audio system based on a bayesian belief networks approach for dialog modelling"* [4] se enmarca en la "9th European Conference on Speech Communication and Technology", conocida como INTERSPEECH 2005-EUROSPEECH. INTERSPEECH es uno de los congresos más importantes en cuanto a Tecnología del Habla a nivel mundial, estas conferencias son de carácter bianual.

- La segunda publicación derivada del proyecto *"Demostración de una interfaz vocal para el control de un sistema de alta fidelidad"* [5] estaba dentro del "XXI Congreso anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)" (2005). La publicación llevó asociada la demostración del sistema en la que cualquier persona que acudió al congreso pudo probar el sistema desarrollado.