Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 451-452 recibido 19-05-2005; aceptado 01-06-2005

Demostración de una interfaz vocal para el control de un sistema de alta fidelidad

F. Fernández, J. Ferreiros, V. Sama, J. M. Montero, R. García
Grupo de Tecnología del Habla, Universidad Politécnica de Madrid
Ciudad Universitaria s/n, Madrid, Spain, 28040
{efhes, jfl, vsama, juancho, rgarcia}@die.upm.es

Resumen: 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. Presentamos una interfaz vocal para el control de un equipo de alta fidelidad usando órdenes o frases habladas de manera natural. Se describen los principales módulos que componen la interfaz destacando el de gestión de diálogo para el que se ha adoptado una estrategia basada en Redes Bayesianas.
Palabras clave: Redes de Creencia Bayesiana, gestión de diálogo, domótica, interfaces vocales.

Abstract: In domotics we can define the dialogue as a communication process aimed at the achievement of some goals. Such goals could be the execution of some control commands on diverse domestic electronic devices. We present a speech interface for the control of a high fidelity system using spoken natural language. The main modules of the interface are described. A Bayesian Networks (BN) approach has been adopted for the dialogue management module.
Keywords: Bayesian Belief Networks, dialog modelling, domotic, speech interfaces.

1 Introducción

A diferencia de los típicos sistemas de control basados en comandos simples pronunciados de forma aislada, esta interfaz conversacional permite a los usuarios controlar el sistema Hifi mediante frases habladas de manera natural. Los usuarios tienen 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.

Queremos 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 controlarán el sistema mediante órdenes vocales a través de un micrófono. Las frases pronunciadas contienen la intención del usuario y serán traducidas al conjunto de comandos IR necesarios para llevar a cabo una cierta acción sobre el sistema. Esta traducción se realizará de tal forma que el conjunto apropiado de comandos IR se envíen 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 tendrá control de aquellas funciones que conlleven funciones físicas, como la carga de cd's, o que no estén programadas en el mando a distancia, como la memorización de sintonías de radio, y que por lo tanto no tengan comandos IR asociados para poder realizarlas. En cualquier caso, el correcto control del equipo está supeditado a que sea la interfaz vocal el único que lo controle, 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.

2 Arquitectura del sistema

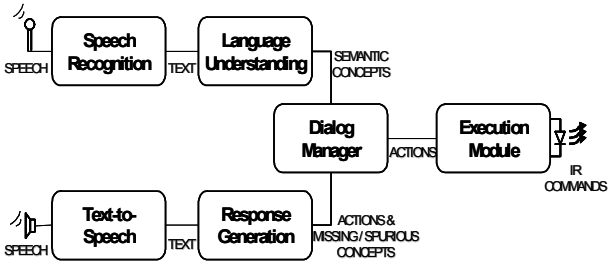

Ilustración 1. Arquitectura del sistema

ISSN: 1135-5948

2.1 Módulo de reconocimiento

Esta constituido por un reconocedor de habla continua que busca la secuencia de palabras (para un vocabulario de 250 palabras) más probable que ha pronunciado el usuario.

2.2 Módulo de comprensión

Este módulo tiene la misión de extraer los conceptos semánticos relevantes a partir de la frase reconocida. Se han definido 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 aplica 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.

2.3 Módulo de gestión de diálogo

Como objetivos de diálogo hemos definido cada una de las posibles acciones (20 en total) que pueden realizarse sobre el sistema. Mediante las BN se identifican los objetivos de diálogo presentes en la frase conforme a la intención del usuario a partir de los conceptos semánticos extraídos. Además, para cada objetivo inferido es posible detectar qué conceptos han sido omitidos para solicitarlos al usuario, y cuáles son erróneos u opcionales para tratar de resolverlos. 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).

2.4 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.

2.5 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. También define las preguntas que deben ser formuladas con el fin de poder recoger información del usuario.

2.6 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.

3 Descripción de la demo

3.1 HW necesario

Se contará con un micrófono inalámbrico, un equipo Hifi, un pc portátil con la interfaz de control vocal instalado y equipado con tarjeta de sonido y un dispositivo hardware vía USB de envío/recepción de comandos IR.

3.2 Desarrollo de la demo

La duración estimada de la demo será de aproximadamente 15 minutos. Esta consistirá básicamente en la interacción con el equipo Hifi haciendo uso de la interfaz vocal con objeto de llevar a cabo un conjunto de acciones. Así mismo se presentará la aplicación gráfica que ha sido desarrollada con fines de evaluación y depuración del sistema enumerándose las diferentes posibilidades que esta ofrece e.g. presentación de información relevante para los diferentes módulos que componen el sistema.

3.3 Objetivos de la demo

El objetivo principal será poder observar como gracias al modelado de diálogo basado en BN el sistema es capaz de resolver situaciones en las que el usuario proporciona una información incompleta o inexacta en relación a los objetivos de diálogo identificados como presentes en las frases procesadas y que reflejan la intención del usuario. Para ello el sistema establece un proceso de negociación con el usuario mediante el que se solicitan los elementos de información omitidos y se trata de resolver los erróneos. Otro de los objetivos de la demo será poner de relieve la utilidad de la estrategia de diálogo implementada que permite mediante el uso de de la historia del diálogo y del estado del sistema la recuperación automática de cierta información omitida dando lugar, en definitiva, a una interacción mucho más flexible y natural.