Control de un equipo de alta fidelidad usando frases habladas de manera natural

Javier Ferreiros López, José Colás Pasamontes, Javier Macías Guarasa,
Ricardo de Córdoba Herralde y José Manuel Pardo Muñoz

Grupo de Tecnología del Habla, IEL, ETSI Telecomunicación, UPM

Resumen

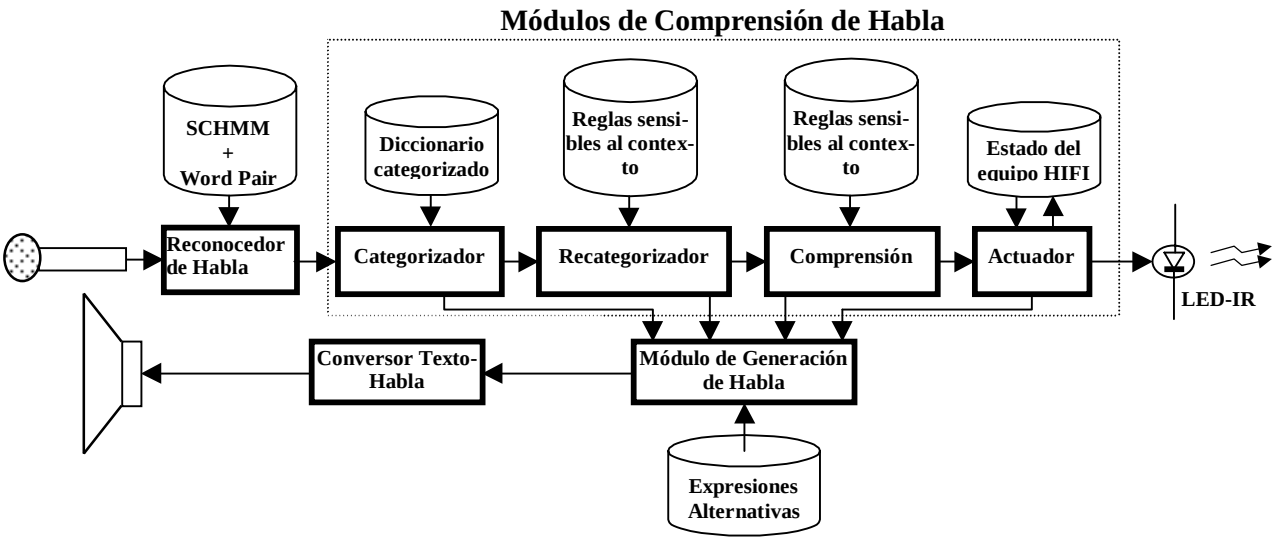
En el Grupo de Tecnología del Habla del Departamento de Ingeniería Electrónica de la Universidad Politécnica de Madrid hemos recogido nuestra experiencia a lo largo de 23 años de investigación en sistemas automáticos de reconocimiento, comprensión, generación y síntesis de habla para utilizarla en la creación de potentes interfaces hombre-máquina utilizando lenguaje hablado. Pensando en la aplicación concreta de esta tecnología a la ayuda de personas con discapacidad y a modo de demostración de las posibilidades y de exploración del comportamiento del usuario frente a esta tecnología, entre otras aplicaciones hemos desarrollado el sistema de control por voz de un equipo de alta fidelidad objeto de esta ponencia.

1. Introducción

El equipo de audio que estamos controlando está compuesto de un triple cargador de discos compactos, una doble pletina de cintas de audio y un receptor de radio. Este equipo comercial tiene un grado muy alto de controlabilidad a través de un mando de infrarrojos y es esta misma vía la que vamos a aprovechar para comandar el equipo desde el ordenador personal que contiene nuestra interfaz de habla. Por lo tanto, la persona que quiere actuar sobre el equipo, habla delante de un micrófono expresando su intención. A diferencia de otros sistemas existentes en el campo de la domótica, el nuestro permite la ejecución de órdenes complejas a partir de un único comando emitido como una frase de habla natural. Además, en la construcción de esta frase el locutor tiene una gran libertad de expresión sin tener que memorizar comandos o sintaxis concretas para conseguir manejar satisfactoriamente el equipo. El sistema interpreta la frase y confirma a través de un sintetizador de voz la acción que se va a realizar. De esta manera, incluso una persona ciega puede saber qué ha hecho exactamente el sistema en cada momento. En futuras versiones es nuestra intención incluir un módulo de diálogo más potente que permita negociar, de una manera más versátil aún, la operación deseada sobre el equipo. Finalmente, el sistema dispone de un pequeño módulo electrónico conectado al puerto paralelo del ordenador y que se encarga de generar los comandos infrarrojos oportunos.

2. Componentes físicos del sistema

El sistema se compone de los siguientes elementos: Un micrófono SHURE SM10, un preamplificador reaprovechado de la cabecera analógica de un conversor A/D D/A de DIGITAL DSC200, una tarjeta de LSI con un DSP32C conectada a una de las ranuras ISA de un ordenador personal y que se ocupa de la detección de principio y fin de frases, de la extracción de características y de la cuantificación vectorial suave necesaria para los modelos semicontinuos de Markov con que trabaja el sistema, siendo responsabilidad del procesador central del ordenador personal el resto de tareas de la arquitectura (reconocimiento + comprensión + ejecución



de comandos por el canal de infrarrojos), una tarjeta VISHA con otro DSP32C desarrollada en nuestro Departamento y un altavoz como subsistema de síntesis y un pequeño módulo de recepción / emisión de infrarrojos conectado al puerto paralelo del PC que nos permite copiar comandos de un mando a distancia y posteriormente reproducirlos a voluntad bajo control de nuestra aplicación. Esta compleja arquitectura electrónica se ha arrastrado históricamente desde el comienzo de esta idea cuando era ventajoso el apoyo de estos diferentes módulos. Como se puede observar en el apartado de líneas futuras de trabajo, uno de los objetivos actuales es librarnos de gran parte de esta parafernalia innecesaria hoy en día gracias a las tarjetas de entrada / salida de sonido y los potentes ordenadores disponibles.

3. Arquitectura del sistema

A continuación, y haciendo referencia a la ilustración con el diagrama de módulos del sistema, vamos a relatar la tecnología implicada. Se pueden encontrar más detalles en [2][3][4].

3.1. Captura de comandos

El habla es recogida a través de un micrófono y muestreada usando una tarjeta de sonido. Utilizamos un micrófono SHURE SM10 que el usuario se coloca en la cabeza. Este micrófono recoge el habla desde unos 2 cm. de la comisura de los labios y el fabricante asegura además una cierta capacidad de eliminación del ruido ambiente. También hemos incluido algunas técnicas de detección de habla en el sistema de captura: Por una parte hemos incorporado un detector basado en un autómata que examina una secuencia válida de niveles de energía para determinar los puntos de comienzo y fin de un comando hablado. Este sistema utiliza un conjunto de umbrales de energía que determinan la transición entre los estados del autómata que se adaptan al ruido ambiente presente en el entorno de utilización del sistema. En las primeras versiones, esta adaptación se realizaba explícitamente tras pedírselo al sistema con el comando "Prueba de nivel" y con la recomendación a los usuarios de que al menos al comenzar una sesión se solicitara una adaptación. La prueba de nivel pide al usuario que se mantenga en silencio durante unos dos segundos y posteriormente que hable durante otros dos. En las versiones actuales esta adaptación se hace automáticamente gracias a un algoritmo que se adapta a las condiciones acústicas en todo momento, lo cual ha simplificado la utilización del sistema al liberar al usuario de esta preocupación. Por otra parte, el sistema puede encontrarse en dos estados concretos: uno, el estado de reposo, en el que no realiza ninguna interpretación de lo pronunciado por el locutor quedando a la espera de un comando concreto (actualmente, la frase "Atiende ahora") para comenzar a procesar comandos. El otro estado, el activo, permite la ejecución de varias órdenes sobre el equipo HIFI hasta que se pronuncie otro comando especial (actualmente, la frase "Descansa ahora") con el que se le ordena pasar de nuevo al estado de reposo. De esta manera el usuario puede hablar tranquilamente durante el estado de reposo sin que el sistema intente interpretar todo lo que dice. Al ser comandos vocales los que controlan también el estado del sistema, podemos decir que el sistema es totalmente manos libres, característica fundamental para personas con discapacidades motrices severas.

3.2. Preproceso y reconocimiento

A continuación en la arquitectura, el ordenador extrae un conjunto de características del habla que son actualizadas cada 10 milisegundos. El reconocedor de habla asume un modelo dinámico estocástico del habla conocido en la literatura como Modelo Oculto de Markov usando como unidades básicas de reconocimiento las palabras, construidas a partir de los modelos acústicos básicos que son los alófonos. La versión básica maneja un vocabulario de 175 palabras aunque el inventario concreto es fácilmente ampliable o adaptable a las costumbres del usuario como destacaremos también más adelante. Concretamente estamos utilizando en esta aplicación modelos semicontinuos con 256 gausianas en cada *codebook* y trabajamos con dos *codebooks*: uno para 10 parámetros MFCC y la energía y otro para la primera derivada temporal de los anteriores. El algoritmo de reconocimiento busca la secuencia más probable de palabras (la frase) que ha pronunciado el locutor. Se trata de un *one-pass* [1] guiado por la transcripción alofónica del vocabulario de la aplicación y una gramática de pares de palabras entrenada con frases típicas de control y posteriormente suavizada a mano para incrementar su cobertura. Una de las características importantes de un sistema como el que aquí hemos desarrollado se encuentra en su capacidad de responder en el menor tiempo posible. Por ello se ha trabajado en el aprovechamiento de la capacidad de proceso disponible paralelizando al máximo las tareas. En la arquitectura que hemos dado en llamar "histórica", el DSP32C de la tarjeta LSI realiza las tareas de muestreo, extracción de características, detección de principio y fin de comandos y cuantificación suave de los modelos semicontinuos de Markov. Tan pronto como se verifica un principio de una frase, el procesador central del ordenador comienza el proceso de reconocimiento con sincronía en trama. Es decir, estamos reconociendo ya al mismo tiempo que el locutor sigue pronunciando su comando. Esto permite que con un muy pequeño retardo tras el fin de la pronunciación de la frase (y dependiendo obviamente de su longitud), obtengamos ya la secuencia de palabras reconocida. Los módulos de comprensión y síntesis ocupan un tiempo despreciable de proceso posterior y la sensación del usuario es una respuesta prácticamente inmediata a su petición (el retardo medio desde el fin de frase está en el entorno de un segundo).

3.3. Comprensión de habla

Detrás del módulo de reconocimiento entra en juego el sistema de comprensión, que empieza por etiquetar cada una de las palabras reconocidas con una categoría sintáctico-semántica que tiene ya bastante que ver con el dominio concreto de la aplicación. Esta tarea la realiza el módulo que hemos llamado Categorizador que simplemente asigna una o varias categorías a cada palabra en función de un diccionario categorizado de que dispone el sistema. Una característica importante del sistema es que este diccionario categorizado se obtiene de la recopilación de las palabras que pertenecen a cada categoría, estando estas listas reflejadas en ficheros separados para cada una. Este pequeño detalle de implementación permite la fácil ampliación del vocabulario manejado por el sistema. Si, por ejemplo, un usuario utiliza frecuentemente la expresión "arranca la cinta" para reproducir una cinta, las modificaciones a hacer en el sistema para que soporten la nueva palabra "arranca" son las siguientes: transcribir a alófonos la palabra arranca y consignarlo en el diccionario de reconocimiento utilizando una herramienta automática de transcripción grafema-alófono e incluir el grafema "arranca" en todas los ficheros de listas de palabras pertenecientes a categorías concretas que tengan esa misma función sintáctico-semántica. Como guía para esta tarea se puede buscar dónde está consignada una palabra sinónimo (la palabra "reproduce" en nuestro ejemplo). El conjunto de reglas de los módulos de comprensión no tienen que ser alterados y la nueva palabra será admitida como sinónimo automáticamente. Estas primeras categorías son refinadas por el módulo Recategorizador haciendo uso de un conjunto de reglas sensibles al contexto que empiezan a integrar y a buscar coherencia entre la semántica que aportan las distintas palabras componentes de la frase reconocida. Finalmente, en el módulo de comprensión, otro conjunto de reglas sensibles al contexto disparan una interpretación de la frase rellenando a su vez un conjunto de campos en uno o varios marcos semánticos que son enviados al intérprete para confirmar la acción al usuario y realizarla de forma efectiva. Nuestros marcos semánticos son muy simples: de manera general se componen de tres ranuras: un aparato sujeto de la acción, un parámetro de ese aparato a controlar y un valor que deseamos adopte el parámetro indicado. Existen acciones que no necesitan siquiera que todas las ranuras del marco estén completas.

3.4. Actuación sobre el equipo

El resultado final se traduce en una secuencia adecuada de comandos infrarrojos que produce el actuador para llevar a cabo la acción deseada y que hemos copiado previamente en el diseño del sistema con un algoritmo automático del mando a distancia original del equipo HIFI. Este subsistema de copia de comandos originales nos permitirá en posteriores proyectos manejar cualquier otro equipo que disponga de mando a distancia por infrarrojos. Un problema a destacar de este tipo de control del equipo es que la información es en el único sentido mando – equipo de manera que no podemos recuperar ninguna información del estado del equipo, por ejemplo. Esto obliga a que el módulo actuador conserve una base de datos con el estado completo de todos los parámetros del equipo HIFI y que tengamos que prohibir absolutamente cualquier actuación manual sobre el equipo desconocida por nuestro sistema. El sistema parte al conectarse por primera vez de una configuración estándar que hay que asegurar en el equipo HIFI y a partir de ese momento guarda en disco el estado entre consecutivos apagados y encendidos del sistema de control para poder actuar siempre sin problemas. Otra característica del módulo actuador es que también posee cierta inteligencia propia. Como ejemplo trivial, si el sistema de comprensión pide al actuador que encienda el equipo y el equipo estaba ya encendido, el actuador no repite el comando y envía un mensaje al usuario informándole de que el equipo ya estaba encendido. Esta es una característica de robustez ante frases perfectamente comprendidas pero que plantean problemas prácticos de ejecución con peligrosas pérdidas de control si no se tratan.

3.5. Respuesta vocal

El sistema se completa con un subsistema de respuesta vocal que hace uso de un sintetizador de voz para confirmar al usuario en todo momento las interpretaciones y acciones que realiza el sistema. Este subsistema se compone en primer lugar de un generador de habla que compone mensajes recogiendo información de todos los módulos de comprensión. Este generador de habla trabaja en un primer nivel con un relleno de plantillas adecuadas a la información que se desea dar y que es alimentado directamente por las diversas reglas que actúan en la interpretación de los mensajes hablados. Las plantillas especifican por una parte los conceptos que componen las frases y por otro lado cadenas de caracteres literales que contienen datos específicos a presentar en el mensaje concreto. Trabajar de esta manera, utilizando la mezcla de identificadores de conceptos con literales nos permite en un segundo nivel utilizar una sustitución aleatoria de cada concepto por una expresión concreta, dentro de un conjunto de diversas alternativas, que lleven la misma carga semántica. El resultado final es que el texto a sintetizar es variable en cuanto a su expresión, pero no en cuanto a su contenido, cada vez que se utiliza un mismo tipo de mensaje (una misma plantilla). Esto produce el efecto de una comunicación variada que da una sensación de naturalidad importante al usuario del sistema y evita el tedio de escuchar siempre la misma expresión para distintas ocasiones del mismo caso.

4. Explicación del razonamiento como técnica de depuración del sistema

Cuando el sistema comenzó a funcionar con las primeras reglas implementadas, tuvimos que recurrir inmediatamente a la necesaria fase de depuración. En particular, el diseño de los diversos conjuntos de reglas de los módulos de comprensión requirió la inversión de mucho esfuerzo manual de ajuste. La inexistencia de procedimientos automáticos de extracción de estas reglas llevan a la necesidad de la inversión de este esfuerzo manual. En un principio este trabajo se realizaba con ficheros adicionales de salida de depuración del sistema donde se imprimían la serie de reglas y los resultados parciales que se obtenían de su aplicación al conjunto de frases que teníamos de referencia y a otras que fuimos añadiendo durante el desarrollo. En cierto punto de este tedioso trabajo se nos ocurrió añadir a cada regla del sistema la generación, utilizando la metodología en dos niveles anteriormente descrita, del mensaje que explicaba qué estaba haciendo esa regla. En conclusión, era el mismo sistema el que nos explicaba los "razonamientos" que le conducían a una interpretación concreta de las frases. Escuchar estos razonamientos a través del sintetizador de voz fue a la par divertido y útil para abstraer y comprender más fácilmente qué reglas estaban causando problemas de interpretación y por qué.

5. Líneas futuras de trabajo

La continuación de este trabajo sigue dos líneas importantes: La primera es un transporte del sistema desde su corsé actual delimitado por la utilización del sistema operativo D.O.S. y de un hardware especial de procesamiento de la señal de habla (Tarjetas de LSI y VISHA que contienen un DSP32C) que lo convierte en un sistema único de demostración prácticamente irreproducible y que fue condicionado todo ello por el estado de la tecnología en los momentos iniciales de la idea, a una plataforma más dinámica y versátil como la que se nos ofrece hoy en día con el uso de tarjetas convencionales de entrada / salida de audio de que disponen prácticamente todos los ordenadores actuales, el transporte al sistema operativo Windows ampliamente diseminado y el aprovechamiento de la capacidad de proceso de los procesadores Pentium de que disponen los ordenadores actuales que hacen innecesario el apoyo que antiguamente nos daban las tarjetas con DSP.

La segunda es un objetivo de mejora de las capacidades del sistema añadiéndole más tecnología. Nos referimos concretamente a añadir una capacidad más elaborada de diálogo con el usuario que permita la negociación de la acción deseada. Esta será una característica relevante del sistema en cuanto a que aumentará la robustez del mismo. Actualmente, hay frases que llegan al sistema de comprensión sin todos los datos necesarios. Por ejemplo, podemos tener la frase: "reproduce la cinta" y ya que disponemos de dos cintas, nos quedaría la duda de cuál es la cinta que el usuario desea escuchar. Tal como está el sistema, hay acciones de este estilo que simplemente no se ejecutan y se informa al usuario de que no se le ha comprendido o en el mejor de los casos se utilizan decisiones por defecto (en el ejemplo, podemos decidir reproducir la cinta que se haya estado utilizando más recientemente). Evidentemente, la solución correcta es informar al usuario de que se ha comprendido que desea reproducir una cinta pero que necesitamos que nos indique qué cinta concreta desea escuchar. Por lo tanto podremos a continuación lanzar un reconocimiento restringido para intentar captar la opción concreta que necesitamos para completar la interpretación o cancelar esa interpretación parcial que se ha hecho si el usuario nos indica que es incorrecta. Las razones por las cuales una frase puede llegar incompleta son varias. Por una parte el reconocedor de voz puede cometer errores y entregarnos una frase equivocada (entendemos frase equivocada como aquella en la que una palabra con relevancia semántica ha sido omitida, insertada o substituida por otra). Por otra parte, el usuario puede no haber especificado completamente el comando y esto en mayor medida en este sistema donde hemos querido dotar al usuario de naturalidad de expresión y no tener que memorizar comandos concretos para interactuar con el equipo HIFI. El incrementar la robustez ante estas dos situaciones es el objetivo que nos lleva al deseo de incluir esta capacidad de diálogo aumentada en futuras versiones.

6. Agradecimientos

Deseamos agradecer a todo el Grupo de Tecnología del Habla y en particular a los alumnos de Proyecto Fin de Carrera Natalia París Navajas, Yolanda López Moreno, Alejandro Ruiz Sánchez y Oscar Benjamín García Toledo por su colaboración y esfuerzo en la consecución de este objetivo.

Referencias

[1] Ney H. "The use of one-stage dynamic programming algorithm for connected word recognition", IEEE trans. On Acoustic, Speech and Signal Processing, Vol 32 (2) pp. 263-271, April 1984

[2] J. Ferreiros, J. Colás, J. Macías-Guarasa, A. Ruiz, J. M. Pardo "Controlling a HIFI with a continuous speech understanding system", ICSLP 98, The 5th International Conference on Spoken Language Processing, Sydney, Australia, noviembre 1998, ISBN 1-876346-17-5

[3] J. Colás, J. Ferreiros, J. M. Montero, J. Pastor, A. Gallardo, J. M. Pardo "On the Limitations of Stochastic Conceptual Finite-State Language Models for Speech Understanding", ICSLP 98, The 5th International Conference on Spoken Language Processing, Sydney, Australia, noviembre 1998, ISBN 1-876346-17-5

[4] J. Colás, J. M. Montero, J. Ferreiros, J. M. Pardo "An alternative and flexible approach in robust information retrieval system", EUROSPEECH' 97, pp 2683-2686, ISSN 1018-4074, Rhodes, Greece