SERVICIO DE PÁGINAS AMARILLAS UTILIZANDO RECONOCIMIENTO DISTRIBUIDO DE VOZ

José A. González, Ángel M. Gómez, José L. Carmona y Antonio M. Peinado

Dpto. de Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada

RESUMEN

En este trabajo se describe un prototipo de servicio de acceso a información para dispositivos móviles basado en reconocimiento distribuido de voz (DSR). La aplicación implementa un servicio de páginas amarillas en el que se consulta información sobre establecimientos y lugares de interés en una ubicación determinada. La arquitectura de reconocimiento remoto utilizada hace uso de los estándares DSR de la ETSI, que integran soluciones frente a ruido acústico y errores de transmisión. El sistema se basa en redes de paquetes, de modo que dependiendo de la conectividad del dispositivo se autoconfigura para el uso de distintas redes IP. El sistema de reconocimiento, ubicado en el lado del servidor, se basa en el reconocedor HTK. Los modelos acústicos se han entrenado para el castellano, mientras que el modelo de lenguaje se adecúa al de un servicio de páginas amarillas. Aunque el prototipo sólo puede responder a peticiones y cuestiones en el marco del dominio escogido, el sistema es flexible y permite ser exportado a otros dominios.

1. INTRODUCCIÓN

El desarrollo e implantación de nuevos servicios sobre redes inalámbricas ha encontrado un gran obstáculo en su adaptabilidad a los terminales, ya que estos tienden a disminuir las dimensiones de los interfaces (pantalla, teclado...) para aumentar su portabilidad. Esta tendencia motiva el desarrollo de nuevos interfaces de usuario que provean de una interacción natural, ubicua y multimodal. En este contexto, la naturalidad de la voz hace que las tecnologías del habla jueguen un papel decisivo. Concretamente, el Reconocimiento Automático del Habla (RAH) habilita la aparición de servicios accedidos mediante el habla permitiendo una cómoda interacción hombre-máquina.

Existen dos posibles arquitecturas para la implementación de los servicios accedidos por voz: reconocimiento de voz empotrado (ESR, del inglés *Embedded Speech Recognition*), donde el sistema RAH se encuentra en su totalidad en el dispositivo móvil o portátil, y el reconocimiento distribuido de voz (DSR, del inglés *Distributed Speech Recognition*), basado en una arquitectura cliente-servidor. En este segundo enfoque el cliente sólo parametriza y codifica la señal de voz, mientras que el motor de reconocimiento se integra en el servidor. Frente a ESR, el esquema distribuido presenta como ventaja la reducción de los requerimientos computacionales, así como del consumo de los dispositivos en el lado del cliente, trasladando las partes de mayor complejidad computacional del reconocedor de voz al servidor. Además, permite el fácil mantenimiento y actualización del núcleo del reconocedor en el servidor [1], así como la posibilidad de añadir nuevos servicios sin modificar el cliente, como por ejemplo el reconocimiento de nuevas lenguas.

A su vez, la creación de los nuevos estándares de acceso inalámbrico a Internet abre un amplio abanico de posibilidades de prestación de servicios, debido a la convergencia de las distintas redes inalámbricas de una futura cuarta generación. Prueba de esta convergencia es la nueva aparición de teléfonos móviles celulares que incluyen conexión *bluetooth* (red de área personal), Wi-Fi (red de área local) y UMTS (*Universal Mobile Telecommunications System*). Esto permite al terminal escoger aquella tecnología que le ofrezca mejores condiciones de acceso en cada situación.

En este trabajo se ha implementado un prototipo DSR para el acceso, a través de redes IP, a un servidor de información remoto. El sistema consta de dos partes: el extractor de características, o *front-end*, incluido en un cliente del tipo PDA (*Personal Digital Assistant*) o un teléfono móvil; y el *back-end,* integrado en el servidor, que procesa la información recibida y que lleva a cabo la tarea de reconocimiento.

Como tipo de aplicación se ha escogido un servicio de páginas amarillas accedido mediante voz. En este caso las consultas se realizan en castellano, aunque también se dispone de una versión en inglés [2]. El cliente lleva a cabo consultas sobre distintos tipos de establecimiento en una cierta ubicación. El servidor a su vez se encarga de realizar una transcripción textual de la petición que es utilizada para obtener una salida multimodal, en la que se muestra un mapa de la ubicación seleccionada y un listado de los establecimientos que se adecúan a la consulta realizada. De este modo, un ejemplo de consulta podría ser: *"¿Dónde hay una cafetería en Madrid?"*, a la que el sistema respondería con la información mostrada en la fi-

Este trabajo ha sido subvencionado con fondos del proyecto MEC/-FEDER TEC2007-66600.

Figura 1. Ejemplo de salida multimodal a la consulta *"¿Dónde hay una cafetería en Madrid?"*.

gura 1. Además, en caso de que el cliente disponga de un sistema GPS (*Global Positioning System*) se utiliza la información de posicionamiento para precisar la respuesta.

Esta comunicación se organiza del siguiente modo: en la sección 2 se muestra la arquitectura general del sistema; en la sección 3 se describen las características del sistema DSR, los modelos acústicos empleados, así como el vocabulario y la gramática utilizados; el sistema de información se describe en la sección 4. Finalmente, en la sección 5 se presentan las conclusiones.

2. ARQUITECTURA DEL SISTEMA

Como se ha comentado anteriormente, en este trabajo se describe un sistema de información de páginas amarillas accedido por voz. Para implementar el sistema se ha seguido un enfoque distribuido como el que aparece en la figura 2. La aplicación consta de tres elementos principales: 1) un cliente, encargado de parametrizar la voz del usuario y realizar las consultas al sistema de información; 2) un servidor RAH, que reconoce la voz del usuario a partir de los parámetros transmitidos por el cliente y 3) un servidor que retorna información sobre la consulta efectuada.

Bajo este esquema, el terminal cliente (PDA o teléfono móvil) parametriza la señal de voz del usuario utilizando un *front-end*. Los parámetros de la voz calculados se envían al servidor de reconocimiento a través de una red de paquetes (en nuestro caso TCP/IP). Dado que se pueden producir pérdidas en la red, el servidor de reconocimiento incluye un *back-end* que aplica técnicas de mitigación para aliviar el efecto de estos errores. Además, el *back-end* procesa los parámetros recibidos añadiendo características que modelan la evolución temporal de la voz (parámetros dinámicos). Finalmente, el motor de reconocimiento decodifica el mensaje del usuario a partir de estos parámetros y envía el texto reconocido de vuelta al cliente.

En la aplicación que describimos, el usuario puede realizar consultas sobre un tipo de establecimiento en una determinada área. Para flexibilizar la interacción con el usuario, la posición en la que se efectúa la búsqueda de establecimientos se puede indicar de forma explícita mediante voz (por ejemplo diciendo el nombre de una ciudad) o de forma implícita, utilizando las coordenadas geográficas del usuario. La elección entre uno u otro método la realiza de forma automática el cliente analizando el texto reconocido. En concreto, si el usuario no ha especificado ningún lugar de búsqueda en su consulta, se utilizan las coordenadas geográficas proporcionadas por un GPS instalado en el terminal (si éste lo incluye). En el caso que no se disponga de GPS ni se haya especificado ninguna localización, la búsqueda se realiza en el lugar establecido por defecto durante la instalación del sistema.

Las consultas aceptadas por el servidor de información responden al siguiente formato: ⟨ *establecimiento, posición* ⟩. En esta consulta, *establecimiento* es el servicio buscado (por ejemplo restaurante, banco, estación de autobuses, ...) y *posición* es la localización alrededor de la cual se efectúa la búsqueda (el nombre de una ciudad o un par de coordenadas GPS). En este trabajo el servidor de información utilizado es Google Maps [3]. La información devuelta por éste consiste en un mapa con la posición de los establecimientos buscados, así como una lista con información sobre éstos.

3. SISTEMA DSR

En esta sección se describe el sistema de reconocimiento distribuido, los modelos acústicos y el modelo de lenguaje empleado.

3.1. Arquitectura Distribuida

Tal y como ilustra la figura 3, el sistema DSR está constituido por dos módulos principales: el extractor de características, o *front-end,* y el motor de reconocimiento. En nuestro caso, hemos utilizado el *front-end* avanzado (AFE, *Advanced Front-End* en inglés) propuesto por el organismo ETSI [4] y el reconocedor de voz HTK [5]. El AFE se encarga de la extracción de vectores de características adecuados para el reconocimiento de voz y de la detección de actividad de voz (VAD, del inglés *Voice Activity Detection*).

Uno de los principales problemas del RAH viene dado por el ruido acústico del entorno. En un contexto móvil, como el definido por la arquitectura de reconocimiento

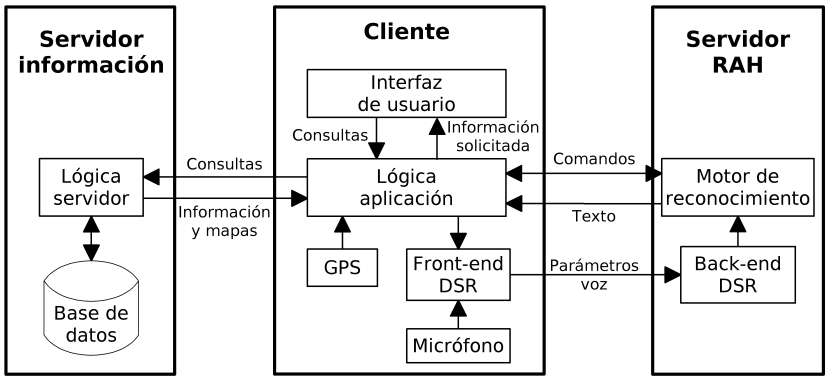

Figura 2. Esquema general de la aplicación desarrollada.

distribuida, el ruido acústico se traduce en una de las principales fuentes de degradación del rendimiento del sistema. Para combatir este problema el AFE, a diferencia de su predecesor [6], incluye un bloque de procesado de señal robusto ante ruido acústico basado en un doble filtro de Wiener [7]. Éste es el principal motivo que justifica la selección de este estándar en el diseño de nuestro prototipo.

Como resultado, el AFE obtiene un conjunto de parámetros extraídos a partir de segmentos de señal de 25 ms, distanciados 10 ms. Para cada una de estas tramas, se calcula un vector de características formado por 13 coeficientes MFCC (*Mel-Frequency Cepstral Components*) y un parámetro de energía en escala logarítmica. Finalmente, los vectores de características son codificados y encapsulados en paquetes (2 tramas por paquete).

El envío de estos parámetros se realiza utilizando el protocolo TCP. A pesar de que el retardo en el envío de información no está acotado, este protocolo permite una implementación más sencilla y soslaya la pérdida de paquetes al precio de introducir un mayor retardo en la comunicación.

En el lado servidor los paquetes recibidos son procesados por el *back-end* correspondiente al AFE. Este módulo, en primer lugar, se encarga de detectar y mitigar los posibles errores de transmisión. Posteriormente, a partir de la secuencia de vectores de características reconstruidos son calculadas las correspondientes componentes dinámicas, conformando vectores de características de 39 componentes (13 estáticas, 13 de velocidad y 13 de aceleración). Adicionalmente, se aplica CMN (*Cepstral Mean Normalization*) [7] como técnica básica de compensación de características. Finalmente, la secuencia compensada de parámetros es introducida como entrada al motor de reconocimiento HTK. El reconocedor presenta su resultado tras el final de la locución, marcada por el usuario. Este resultado es devuelto al cliente.

Tanto en el cliente como en el servidor se incluyen dos módulos de control que se encargan de dar soporte lógico al intercambio de comandos. Entre este tipo de instrucciones se encuentran el inicio y final de reconocimiento, así como distintos tipos de comandos para la correcta configuración del sistema.

El cliente DSR ha sido evaluado sobre una PDA HP IPAQ hw6915, que dispone de un microprocesador Intel PXA 270 416 MHz y 64 MB de memoria RAM. Este dispositivo permite el muestreo de la señal de voz a 8 kHz y es capaz de realizar el procesado de voz del AFE en tiempo real.

3.2. Modelo acústico de la voz

El modelado acústico de la voz se realiza a nivel de subpalabra con dependencia del contexto. En concreto, se dispone de un conjunto de modelos de trifonemas (sobre una base inicial de 32 alófonos) dependientes del contexto tanto a nivel interno de palabra, como entre palabras. Cada trifonema se modela mediante un modelo oculto de Markov (HMM, *Hidden Markov Model* en inglés) [7] continuo con una topología de izquierda a derecha, 3 estados y 5 gausianas por estado. Durante el entrenamiento de los HMMs, se lleva a cabo una agrupación de los estados similares mediante árboles de decisión binarios. Esto ayuda a reducir el número de parámetros que necesitan entrenarse además de mejorar este proceso cuando se dispone de una cantidad de datos limitada.

Para entrenar los modelos acústicos se ha utilizado la base de datos de voz ATLAS Spanish Microphone Database (MICROAES) [8]. Esta base de datos consta de grabaciones de 300 locutores realizadas con micrófonos de diferentes calidades, colocados a distintas distancias del locutor. Además, se hacen consideraciones respecto a dialectos del español (se recogen 5 dialectos diferentes), así como diferencias en cuanto a género y edad. En total la base de datos contiene 30 horas de voz.

3.3. Modelo del lenguaje

Por simplicidad, en la versión que describimos el modelo del lenguaje del sistema de reconocimiento viene dado por una gramática regular. Ésta modela las consultas aceptadas por el sistema de información mediante de reglas de producción de la forma,

$CONSULTA : [En $LUGAR] $TC [un | una] $EST |

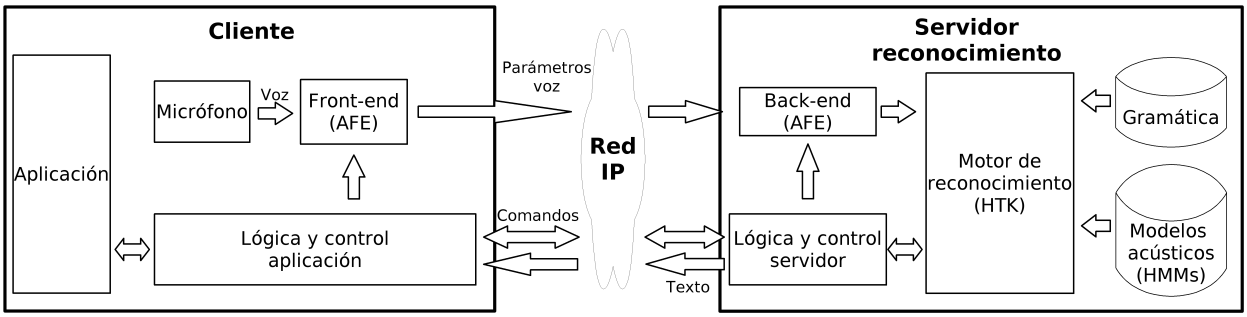

Figura 3. Arquitectura DSR.

```
$TC [un|una] $EST [en $LUGAR]
```

donde los corchetes indican contenido opcional, la barra vertical indica distintas alternativas, $LUGAR es el nombre de una población o ciudad (por ejemplo Bilbao, Granada, etc), $EST es el establecimiento buscado (por ejemplo caja de ahorros, gasolinera, etc) y $TC es una locución del tipo *dónde hay*, *busca*, *enséñame*, etc. El vocabulario resultante contiene unas 160 palabras, entre las cuales se incluyen los nombres de todas las capitales de provincia de España y gran parte de los establecimientos más utilizados en el día a día. La perplejidad de la gramática obtenida es 26,63.

4. SISTEMA DE INFORMACIÓN

El servidor de información se encarga de dar respuesta a las consultas formuladas por el usuario en el dominio de uso del sistema. En la fase de diseño del prototipo descrito, se planteó la necesidad de incluir distintos tipos de información en la respuesta a la consulta realizada. En particular, se consideró oportuno el ofrecer información gráfica en forma de mapas sobre la posición de los establecimientos buscados, así como otros datos relevantes del tipo nombre del establecimiento, dirección, teléfono, etc. Dada la dificultad del diseño y mantenimiento de un servidor de información con estas características, por simplicidad en este trabajo se utiliza como sistema de información externo el proporcionado por Google mediante la herramienta Google Maps [3].

Google Maps es un servicio de aplicaciones de mapas vía Web ofrecido por Google. De entre las características más relevantes de éste encontramos la visualización de mapas e imágenes de satélite del mundo entero, el cálculo de rutas entre distintas ubicaciones, la búsqueda de información sobre lugares de interés (e información sobre negocios) y la posibilidad de incorporar información sobre GPS a las consultas realizadas. Para los propósitos de la aplicación descrita, hemos utilizado las capacidades de este servicio para proporcionar información sobre empresas y lugares de interés. En concreto, la aplicación cliente efectúa consultas a Google Maps utilizando peticiones con el método GET del protocolo HTTP [9]. Cada consulta, por tanto, se traduce en una URL donde se modifica el valor de dos parámetros: el establecimiento buscado y el lugar donde se busca.

5. CONCLUSIONES

En este trabajo hemos presentado un servicio remoto de páginas amarillas accedido mediante voz. El sistema hace uso de la tecnología DSR, adoptando una arquitectura distribuida, en la que el reconocedor de voz y el sistema de información se encuentran ubicados en diferentes servidores. El prototipo desarrollado permite llevar a cabo consultas desde una PDA, obteniendo una respuesta multimodal con la información solicitada. El desarrollo de este prototipo demuestra la madurez de las tecnologías empleadas, así como su utilización para el desarrollo de aplicaciones comerciales.

6. BIBLIOGRAFÍA

[1] Z.-H. Tan, P. Dalsgaard, y B. Lindberg, "Automatic speech recognition over error-prone wireless networks," *Speech Communications*, vol. 47, pp. 220–242, 2005.

[2] *AVIOS Student Contest 2008, http://avios.org/contest2008/individual.htm*.

[3] *Google Maps, http://maps.google.es*.

[4] *ETSI Standard ES 202 212. Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm*, November 2003.

[5] S. Young, G. Evermann, y T. Hain, *The HTK book*, Cambridge Universtiy Engineering Department, 2007.

[6] *ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, 2000.

[7] Antonio M. Peinado y Jose C. Segura, *Speech Recognition Over Digital Channels: Robustness and Standards*, Wiley, 2006.

[8] European Language Resources Association (ELRA), *The ATLAS Spanish Microphone Database (MICROAES), http://www.elda.org/catalogue/en/speech/S0165.html*.

[9] The W3C consortium, *HTTP - Hypertext Transfer Protocol, http://www.w3.org/Protocols*.