Un Sistema de Diálogo para la Consulta de Correo Electrónico en Lenguaje Natural *

Leandro Rodríguez Liñares[1], Carmen García Mateo[2], Santiago Pardo Ríos[2], Víctor Darriba Bilbao[1]
[1]E.T.S.I. Informática - [2]E.T.S.E. Telecomunicación
Universidade de Vigo

Resumen: En los últimos años ha surgido un nuevo tipo de interfaces hombre-máquina que combina varias tecnologías de habla con el fin de permitir a las personas conversar con los ordenadores mediante la voz. En este artículo, presentamos uno de estos sistemas, diseñado para facilitar el acceso a un servidor de correo electrónico usando lenguaje natural.
Palabras clave: Interfaces conversacionales, reconocimiento de voz, conversión texto-voz, sistemas de diálogo, procesamiento del lenguaje natural

Abstract: Last years have witnessed the emergence of a new class of human-computer interfaces that combine several human language technologies to enable humans to converse with computers using speech. In this paper, we describe one of these systems, devised for accessing an e-mail server using natural language.
Keywords: Conversational interfaces, speech understanding, text-to-speech conversion, dialogue systems, natural language processing

1. Introducción

Cada vez es más habitual encontrar sistemas que proporcionan diversos tipos de información utilizando el habla (Zue y Glass, 2000; Bernsen, Dybkjaer, y Dybkjaer, 1998; Hacioglu y Ward, 2001; Pellom, Ward, y Pradhan, 2002). En la construcción de interfaces de comunicación de este tipo están involucradas tecnologías como pueden ser el reconocimiento de habla, la conversión texto-voz y el control de diálogo.

Muchos interfaces de comunicación basados en habla pueden ser considerados conversacionales en el sentido de que tiene lugar una conversación real entre el sistema y el usuario. Estos sistemas pueden ser clasificados en función del grado de actividad del sistema en la conversación. En un extremo tenemos las interacciones en las cuales el diálogo es dirigido o *de iniciativa en el sistema*. En este caso el diálogo es rígido y el usuario debe completar una serie de preguntas formuladas por el sistema en un orden determinado. En el otro extremo están los sistemas *de iniciativa en el usuario*. En este caso el sistema es pasivo y como mucho realiza preguntas cuando necesita información adicional. El típico ejemplo es el sistema que le pregunta al usuario algo del tipo "¿qué desea?". En principio, en este tipo de sistemas las frases del usuario no presentan ninguna estructura predefinida, por lo que la comprensión del lenguaje natural es un aspecto clave. Entre estos dos extremos se sitúan los denominados sistemas *de iniciativa mixta*, en los cuales ambas partes participan activamente para llevar a buen término la interacción.

En esta comunicación se presenta un sistema de diálogo de iniciativa en el usuario para la consulta de correo electrónico. El sistema fue desarrollado adaptando un Conversor Texto-Voz y un Reconocedor de Habla disponibles previamente y funciona en ordenadores PC sobre sistema operativo Linux. En su fase actual de desarrollo, el sistema funciona utilizando como entrada la tarjeta de sonido del ordenador, aunque sería perfectamente factible su adaptación a una tarjeta de gestión telefónica. El sistema de diálogo es un prototipo en constante revisión, por lo que son posibles numerosas mejoras del mismo.

El resto de este artículo está estructurado como sigue: en primer lugar se hace una descripción general del sistema desde el punto de vista de su funcionamiento global. A continuación se describen los bloques funcionales, pasando en la sección 4 a comentar el objetivo principal de este artículo: el procesado de lenguaje natural. En las últimas secciones se presentan conclusiones y posibles mejoras del sistema.

* Este trabajo ha sido financiado parcialmente mediante los Proyectos CICYT TIC2000-1104-C02-01 y TIC2000-0370-C02-01

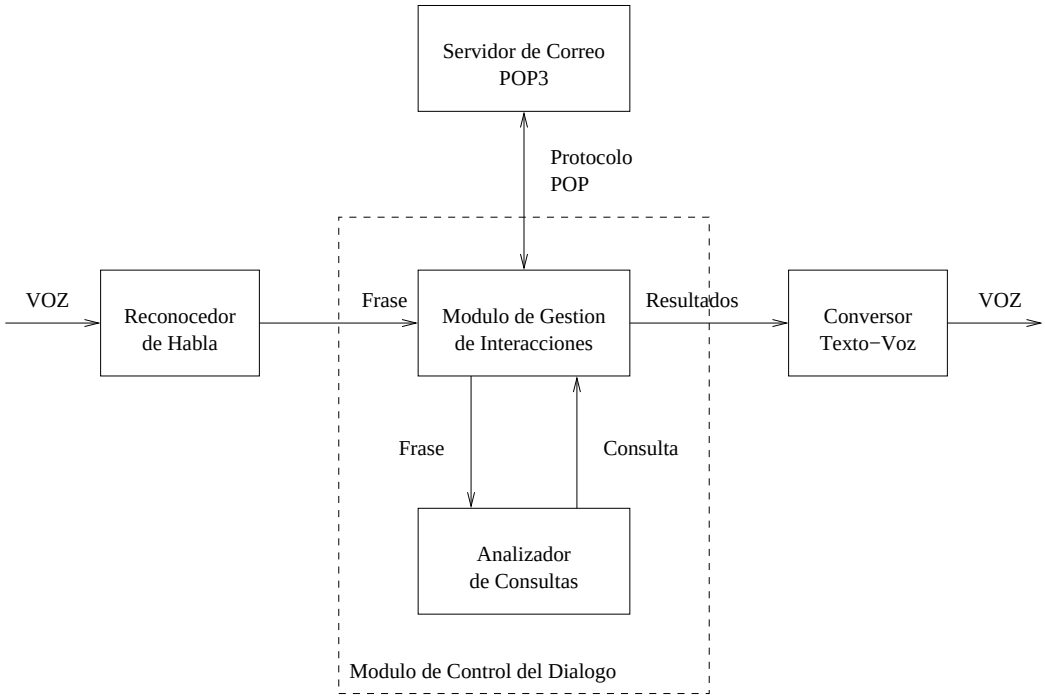

Figura 1: Diagrama de bloques del sistema

2. Descripción general

Tal como se ha comentado en la introducción, el objetivo del sistema presentado en este artículo es el de permitir la consulta de los usuarios de correo electrónico por medio de habla natural. El diagrama del bloques de la aplicación se puede ver en la figura 1. La estructura del sistema es similar a la presentada en (Pérez-Piñar y García-Mateo, 2002). La entrada al sistema será por tanto una frase en lenguaje natural obtenida mediante un *Reconocedor de Habla*, mientras que la salida devolverá la información solicitada mediante un *Conversor Texto-Voz*. De estas dos herramientas (el Reconocedor de Habla y el Conversor Texto-Voz) se disponía previamente, y se realizó su adaptación a la tarea que nos ocupa. Dichas herramientas están descritas de forma breve en las secciones 3.2 y 3.3, respectivamente.

El núcleo de la aplicación presentada es el *Módulo de Control de Diálogo*, que consta de un *Submódulo Analizador de Consultas* y de un *Submódulo de Gestión de Interacciones*.

El objetivo es poder acceder a un servidor de correo entrante para poder llevar a cabo alguna de las siguientes acciones:

- Consultar si se ha recibido algún mensaje y contar los mismos.
- Consultar características de los mensajes tales como la fecha, el remite o el asunto.
- Leer alguno de los mensajes existentes.
- Borrar mensajes del buzón.

En cada caso habrá varias maneras de seleccionar el mensaje o mensajes que se desean consultar, es decir, varios criterios de filtrado de los mismos:

- Indicando su posición dentro de la lista de mensajes.
- Pidiendo los N primeros o los N últimos.
- Indicando el remitente.
- Especificando la fecha de envío. Se podrán solicitar los mensajes enviados en una fecha concreta o que sean anteriores o posteriores a la misma.
- Concretando si son nuevos o ya han sido leídos.

Estos criterios podrán combinarse, por lo que podremos, por ejemplo, solicitar el último mensaje de un remitente concreto, o el primer mensaje recibido en una fecha determinada.

3. Estructura del sistema

En esta sección se presentan de forma breve las funciones y los detalles principales de los bloques del sistema que se pueden observar en la figura 1.

3.1. Módulo de Control del Diálogo

Este módulo es el núcleo principal de la aplicación. Sus funciones incluyen la interacción con las herramientas de habla (el Conversor Texto-Voz y el Reconocedor de Habla) y con el servidor de correo electrónico.

A continuación se describen los submódulos que lo integran.

3.1.1. Submódulo de Gestión de Interacciones

Este submódulo recibe como entrada la frase reconocida en formato texto y se la envía al Submódulo Analizador de Consultas. A partir de la información que éste le devuelve llevará a cabo la interacción con el servidor POP3 y obtendrá de éste la información requerida en la consulta. Esta información será enviada en formato texto al Conversor Texto-Voz.

En el caso de que la consulta no siga las reglas de la gramática especificada, la información enviada el Conversor Texto-Voz consistirá en una frase que comunicará dicho hecho al usuario.

Este submódulo ha sido desarrollado en lenguaje C e incorpora una interfaz de comunicación con un intérprete Prolog (Vilares, Alonso, y Valderruten, 1998; Convington, 1994), lenguaje en que ha sido desarrollado el Submódulo Analizador de Consultas, que se presenta a continuación.

3.1.2. Submódulo Analizador de Consultas

Este submódulo recibe la frase a analizar del Submódulo de Gestión de Interacciones como una lista Prolog. La comunicación se lleva a cabo a través de un predicado `consulta/6`, que tiene un término de entrada y cinco de salida:

```
consulta(+A,-Accion,-Indicador,
  -Numero, -Remite, -Fecha)
```

En `A` se introduce la frase a analizar, y el resto de las variables son instanciadas en el submódulo según el resultado del análisis:

- `Accion:` indica la acción a realizar (*contar, leer, borrar, consultar_remite, consultar_fecha, consultar_tema*).
- `Indicador:` sirve para establecer el criterio de selección de los mensajes de la consulta. Algunos de sus posibles valores son *todos, numero, leidos, remite, fecha_desde, fecha_antes....* Además, estos criterios pueden ser combinados mediante el signo *+*. Por ejemplo, si la consulta se refiere a los mensajes de un determinado remite y en una fecha determinada el valor de `Indicador` sería *remite+fecha*.
- `Numero:` contendrá el número de mensajes a leer o la posición de un mensaje determinado.
- `Remite:` el remitente cuyos mensajes se deseen seleccionar.
- `Fecha:` en el caso de que se deseen seleccionar los mensajes por su fecha. En esta primera versión, se han limitado las capacidades del sistema de tal modo que esta variable solamente podrá contener los valores *hoy, ayer, anteayer, lunes...domingo*. De este modo, solamente se podrán especificar por fecha mensajes de antigüedad igual o inferior a una semana.

La aplicación soporta la combinación de más de un criterio de filtrado de mensajes. En este caso, se combinarían los parámetros correspondientes a cada uno de los criterios. En la tabla 1 se muestran algunos ejemplos de combinaciones de este tipo.

3.2. Reconocedor de Habla

El Reconocedor de Habla utilizado en este sistema se encuentra descrito en (Cardenal López, 2001; Cardenal, Diéguez, y García-Mateo, 2002). Está basado en un motor de reconocimiento que utiliza 25 modelos preentrenados de fonemas. Dichos modelos consisten en HMM's (Modelos Ocultos de Markov) (Rabiner, 1989) que fueron entrenados utilizando una base de datos de voz telefónica, aunque posteriormente fueron adaptados para ser usados para reconocer voz obtenida directamente de la tarjeta de sonido de un PC. Los modelos son de los denominados de izquierda a derecha con 3 estados y 16 mezclas por estado.

La voz obtenida es muestreada a 8 KHz. A partir de la misma, se obtienen la energía y 12 coeficientes mel-cepstra utilizando una ventana de análisis de longitud 20 mseg. con un desplazamiento de 10 mseg. Se aplica un liftering con factor 22 y se añaden las primeras y segundas derivadas de los parámetros,

Consulta	Accion	Indicador	Numero	Remite	Fecha
"Borra los dos primeros mensajes de Manolo"	*borrar*	*remite+primeros*	*2*	*Manolo*	*null*
"Lee los mensajes de Manolo recibidos el viernes"	*leer*	*remite+fecha*	*null*	*Manolo*	*viernes*
"Dime el asunto del primer mensaje nuevo"	*consultar_tema*	*posicion+noleidos*	*1*	*null*	*null*

Cuadro 1: Ejemplos de análisis de consultas

`frase` ⇒ `<frase_verbal>`
`frase_verbal` ⇒ `<verbo>` + `<complemento_directo>` ("quiero contar los mensajes...")
`frase_verbal` ⇒ "me gustaría" + `<complemento_directo>` ("me gustaría leer el correo...")
`complemento_directo` ⇒ `<artículo>` + `<posición>` + `<nombre>` + `<condición>` ("léeme el tercer correo de Manolo")
`condición` ⇒ "enviados por" + `<remite>` ("... enviados por Manolo")

Cuadro 2: Ejemplos de reglas de análisis

obteniendo así vectores de 39 coeficientes por trama.

Para lograr cierta ecualización, a los vectores de parámetros se les suprime su media, por lo que, cuando se reconoce en tiempo real, se necesitan un mínimo número de vectores de parámetros antes de poder realizar una estimación adecuada. Por tanto, el reconocedor no dará resultados correctos hasta después de un cierto tiempo de funcionamiento (que puede ser alrededor de un segundo).

El funcionamiento es el que sigue: el reconocedor está obteniendo de forma constante muestras de la tarjeta de sonido y aplicando un VAD (detector de actividad) basado en umbrales de energía. Mediante el VAD es detectado el momento en el cual el usuario comienza a hablar. A partir de este instante, se almacenan vectores de parámetros hasta que se detecta que el usuario ha dejado de hablar, momento en el cual se se lleva a cabo el proceso de reconocimiento y la frase reconocida se transmite al Submódulo de Gestión de Interacción. El reconocedor permanecerá inactivo hasta que el Submódulo de Gestión de Interacción le transmita la orden de que debe volver a escuchar de nuevo.

3.3. Conversor Texto-Voz Cotovía

El Conversor Texto-Voz utilizado es el denominado Cotovía (Fernández y Rodríguez, 1999; Rodríguez et al., 2002). Cotovía es un Conversor Texto-Voz bilíngüe castellano/gallego basado en concatenación de unidades. Es capaz de generar voz masculina o femenina de banda ancha (8 KHz). La frecuencia fundamental y la velocidad de articulación son parámetros configurables por el usuario.

Cotovía actúa como un módulo completamente independiente del resto del sistema, e integra todas las funciones necesarias para trabajar como aplicación aislada. Entre estas destacan sus capacidades para realizar preprocesado de texto.

4. Consultas en lenguaje natural

En esta sección se presentan más en profundidad dos aspectos básicos de la consulta en lenguaje natural, que son la generación del modelo del lenguaje para el Reconocedor de Habla y el problema de la especificación de los remitentes de los mensajes.

4.1. Generación del modelo de lenguaje

El Reconocedor de Habla utiliza un modelo de lenguaje probabilístico basado en el formato ARPA, similar al que utilizan el conjunto de herramientas de reconocimiento HTK (Young, 2002). Dicho modelo consiste en un léxico y un conjunto de probabilidades de unigramas, bigramas y trigramas a partir de los cuales se construye una red de reconocimiento. Esta red es utilizada por el motor de reconocimiento, que aplica el algoritmo de Viterbi para obtener la cadena reconocida con mayor probabilidad.

Se puede ver, por tanto, que el modelo de lenguaje es un modelo probabilístico, por lo que, para su obtención, es necesario un corpus de texto suficientemente grande a partir del cual se obtienen tanto el léxico como el conjunto de probabilidades de los n-gramas. Lo ideal es utilizar un corpus adaptado a la tarea de reconocimiento, esto es, que sea representativo del posible conjunto de frases que a las que va a tener que enfrentarse el reconocedor. Para obtener este corpus se ha explotado el hecho de que un analizador Prolog puede ser utilizado de forma sencilla como generador de frases correctas de acuerdo a la gramática que define.

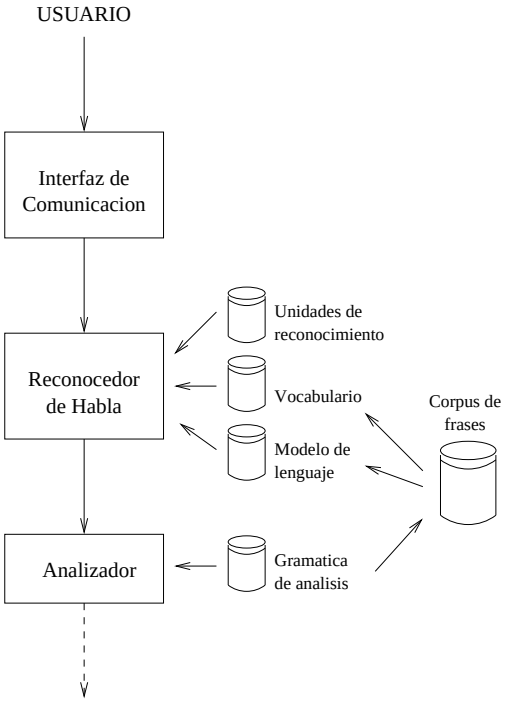

Figura 2: Interfaz de reconocimiento del sistema

En resumen, para obtener el modelo de lenguaje se han llevado a cabo los siguientes pasos (ver figura 2):

1. Definición informal de una gramática para las consultas. Se ha elegido un posible conjunto de estructuras de frase que permitan la obtención de información sobre correo electrónico. Por ejemplo: "léeme los mensajes...", "¿podrías decirme...?", "quiero que me digas el número de..."

2. Implementación de un analizador de la gramática en Prolog. Para ello se ha utilizado un predicado raíz `frase`. A partir de éste, mediante dos reglas, se consideran dos tipos de frases: verbal ("quiero leer...") e interrogativa ("¿podrías decirme...?"). En la tabla 2 se muestran a modo de ejemplo algunas de las reglas utilizadas.

   Se ha tenido en cuenta la concordancia entre género y número para evitar la proliferación de frases gramaticalmente incorrectas.

3. Transformación del analizador en un generador de frases correctas. Generación del corpus de frases correctas.

4. Estimación del modelo de lenguaje. Para ello se ha utilizado el conjunto de herramientas software presentado en (Rosenfeld y Clarkson, 1996).

El vocabulario de la aplicación está compuesto de un total de 144 palabras, y el corpus de frases al que da lugar está compuesto de unos cuantos millones de frases. El número de apariciones por palabra oscila entre las aproximadamente 20.000 veces que aparecen palabras como *nuevo*, *viejo*, etc. y los siete millones que aparece la palabra *el*. La media del número de apariciones para cada palabra es de algo más de medio millón de veces. El modelo de lenguaje estimado consta de 144 palabras, 1637 bigramas y 11801 trigramas. Para la estimación de estas cifras se ha utilizado una agenda de remitentes en la cual había tres personas cuyos nombres constaban de una sola palabra.

4.2. Agenda de remitentes

Aunque son muchos los detalles prácticos a tener en cuenta en la realización y puesta a punto de un sistema de este tipo, uno que merece especial atención es el de los remitentes

de los mensajes. El sistema extraerá el contenido del campo `From` de las cabeceras de los mensajes, campo que puede venir expresado de varias maneras que dependerán de la configuración del mensaje emisor. Lo único seguro es que la dirección de correo electrónico está presente en cualquier caso.

El sistema logrará seleccionar los mensajes de un determinado remitente siempre y cuando logre establecer una relación entre el nombre que se le indique en la consulta por voz (el alias) y el contenido del campo `From` de la cabecera. Para solucionar esto se ha incorporado al sistema una agenda de direcciones que almacenará las correspondencia entre alias y dirección electrónica.

5. *Conclusiones y posibles mejoras*

En esta comunicación se ha desarrollado una aplicación de consulta de correo electrónico con iniciativa en el usuario utilizando el lenguaje Prolog para el análisis de las consultas. La arquitectura presentada ha sido también aplicada a la búsqueda de contenidos en Internet. Esto demuestra que es posible la extensión de esta arquitectura a otras aplicaciones de consulta remota.

Las aplicaciones son ampliables, ya que el diseño del Submódulo Analizador de Consultas permite la extensión de éste, pudiendo tanto añadir palabras al léxico de la aplicación como crear nuevas estructuras sintácticas o nuevas combinaciones de las ya existentes. Sin embargo, el procedimiento para crear los modelos de lenguaje dista de ser cómodo. La generación del modelo de lenguaje a partir de la gramática es un aspecto en el que se está trabajando en el momento en que se escribe este artículo. El analizador Prolog hace uso del formalismo de las DCG's (*Definite Clause Grammars*), a partir del cual es fácil construir algún tipo de script que extraiga tanto el vocabulario como el fichero de gramática en formato HTK (Young, 2002), muy similar al formato de especificación de gramáticas ABNF. La ventaja de utilizar un modelo de lenguaje basado en una gramática de estados frente a un modelo probabilístico es que se evita la necesidad de la construcción de un corpus de frases (ver figura 2), lo que redunda en un aumento de flexibilidad y comodidad en el procedimiento de generación de modelos de lenguaje.

Una limitación introducida en la primera versión del sistema consiste en que los alias de los remitentes están incorporados en el analizador Prolog y en el modelo del lenguaje. Por tanto, añadir un nuevo remitente supone repetir el proceso de generación del modelo del lenguaje desde el principio. La utilización de modelos de lenguaje basados en gramáticas de estados evitará este problema.

Otras modificaciones irán encaminadas al aumento de la inteligencia y la naturalidad del diálogo. Entre estas podemos citar:

- En el estado actual del sistema, éste es sin memoria, es decir, cada consulta se considera completa de forma individual. Una mejora sustancial sería dar al usuario la posibilidad de completar información en varios pasos de interacción.

- Cuando una consulta no es correcta gramaticalmente, el sistema devuelve un mensaje sencillo notificándolo. Otra posible mejora consiste en que el sistema sea capaz de reconocer cuando una consulta es parcialmente correcta e interrogar al usuario por la información que le falta.

Tal como se ha comentado, el sistema presentado está en continua revisión. Por tanto, a falta de una versión estable y definitiva, no se ha realizado una evaluación exhaustiva de las prestaciones del mismo. Las pruebas de funcionamiento se han realizado de una manera más o menos informal, por lo que no han sido incluidas en este artículo.

Bibliografía

Bernsen, N., H. Dybkjaer, y L. Dybkjaer. 1998. *Designing Interactive Speech Systems.* Springer-Verlag.

Cardenal, A., J. Diéguez, y C. García-Mateo. 2002. Fast LM Look-Ahead for Large Vocabulary Continuous Speech Recognition using Perfect Hashing. *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).* Orlando, USA.

Cardenal López, Antonio. 2001. *Realización de un Reconocedor de Voz en Tiempo Real para Habla Continua y Grandes Vocabularios.* Ph.D. tesis, Dpto. de Tecnoloxías das Comunicacións, ETSE Telecomunicación, Universidade de Vigo.

Convington, Michael A. 1994. *Natural Language Processing for Prolog Programmers*. Prentice Hall.

Fernández, X. y E. Rodríguez. 1999. Segmental Duration Modelling in a Text-to-Speech System for the Galician Language. *Proceedings of the EuroSpeech*.

Hacioglu, K. y W. Ward. 2001. Dialog-Context Dependent Language Modeling Using N-Grams and Stochastic Context-Free Grammars. *Proceedings of the ICASSP*. Salt Lake City, USA.

Pellom, B., W. Ward, y S. Pradhan. 2002. The CU Communicator: An Architecture for Dialogue Systems. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Beijing, China.

Pérez-Piñar, D. y C. García-Mateo. 2002. Integración Automática de Fuentes de Conocimiento Lingüístico en el Desarrollo de Sistemas de Diálogo. *Actas del XVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*.

Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications i n Speech Recognition. *Proceedings of the IEEE*, 77:257–285.

Rodríguez, E., F. Campillo, E. Fernández, y F. Méndez. 2002. Sistema de Conversión Texto-Voz en Lengua Gallega Basado en la Selección de Unidades Acústicas y Prosódicas. *Actas del XVIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*.

Rosenfeld, Ronald y Philip Clarkson. 1996. CMU-Cambridge Statistical Language Modelling Toolkit v2.

Vilares, M., M. Alonso, y A. Valderruten. 1998. *Programación lógica*. Ed. Tórculo.

Young, Steve. 2002. Hidden Markov Model Toolkit (HTK). http://htk.eng.cam.ac.uk/.

Zue, Victor W. y James R. Glass. 2000. Conversational Interfaces: Advances and Challenges. *Proceedings of the IEEE*, páginas 1166–1180.