PROTOTIPO DE SISTEMA DE AUTOMATIZACIÓN DE ENCUESTAS TELEFÓNICAS

Alejandro H. Toselli[1], *Elsa Cubel Barea*[1], *Alberto Sanchis Navarro*[2]

[1]Institut Tecnològic d'Informàtica
[2]Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{ahector,ecubel,asanchis}@iti.upv.es

RESUMEN

En este trabajo se presenta un prototipo de sistema de automatización de encuestas telefónicas (ADET). Se describe las características relevantes del mismo, la complejidad de la tarea que implica una encuesta real y se dan las prestaciones del mismo para una encuesta dada. El prototipo es completamente funcional y accesible al público desde internet: **http://prhltdemos.iti.es/∼adet**.

1. INTRODUCCIÓN

En este trabajo se presenta un prototipo de automatización de encuesta telefónicas (ADET), desarrollado por el *Instituto Tecnológico de Informática* (ITI - www.iti.es) en colaboración con la empresa ODEC (empresa de servicios y aplicaciones informáticas de captura y proceso de datos - www.odec.es).

El objetivo concreto del prototipo es evaluar la viabilidad para el desarrollo de futuros sistemas basados en tecnología de reconocimientos espontáneo del habla, donde una de sus potenciales aplicaciones serían los sistemas de automatización de encuestas telefónicas.

En líneas generales, el prototipo ADET se ha desarrollado bajo el esquema cliente-servidor (ver figura 1), interactuando ambas partes a través de un canal de internet y de una conexión telefónica. La parte del cliente involucra un ordenador donde se ejecuta una interfaz de visualización de preguntas y resultados reconocidos (respuestas) y un teléfono. El servidor incorpora un motor de reconocimiento del habla y un módem con el cual establece la comunicación telefónica. El prototipo es completamente operativo y accesible desde **http://prhltdemos.iti.es/∼adet**.

Desde el punto de vista funcional, el prototipo se especializa, por un lado, en el reconocimiento propio de las respuestas de voz telefónica (para cada pregunta que se va realizando), obteniendo y almacenando de cada una de ellas la transcripción ortográfica de la información relevante. Por otro lado (el prototipo) se encarga también de realizar las preguntas de la encuesta, mediante la reproducción de ficheros de audio previamente grabados. No se descarta que en un futuro, esta funcionalidad esté implementada en cambio por un sintetizador de voz, que reproducirá las preguntas basándose en información textual.

El desarrollo del prototipo ha estado supeditado a tres directivas a saber:

a) La puesta a punto del prototipo para la realización de una encuesta específica, se debe realizar de forma automática con una mínima supervisión. Esta característica permite que el sistema pueda adaptarse fácilmente a cualquier tipo de encuesta.

b) El encuestado puede emitir la respuesta de un modo natural y el prototipo identifica cuál es la palabra o palabras relevantes que indican la opción elegida entre las posibles respuestas.

c) El prototipo debe incorporar la capacidad de detectar errores de reconocimiento. De este modo, cuando el sistema identifica mal la respuesta escogida, se repite la pregunta.

Sin embargo, por la gran complejidad que puede entrañar una encuesta real, solo se consideran en este caso las que involucran respuestas del tipo *multiple-choise*.

El presente trabajo se ha estructurado en cinco secciones principales. La sección 2 incluye un análisis de la complejidad que supone la automatización de encuestas telefónicas, basándonos en ejemplos de encuestas reales[1]. En la sección 3 se incluye una descripción detallada del prototipo desarrollado que permite la automatización de las encuestas telefónicas. La sección 4 muestra los resultados experimentales obtenidos para validación del prototipo. Por último, en la sección 5 se destacan las conclusiones principales de este informe.

2. ANÁLISIS DE LA COMPLEJIDAD DE LA TAREA

Una encuesta es llevada a cabo mediante la formulación, por parte del encuestador, de una serie de preguntas, y la contestación correspondiente a cada una de ellas por

Este trabajo ha sido financiado parcialmente por *Conselleria d'Empresa, Universitat i Ciència - General itat Valenciana* ID:GV06/252.

[1]Muestras proporcionadas por ODEC, S.A. (`www.odec.es`)

el encuestado. En la gran mayoría de los casos, al encuestado se le ofrece un conjunto de posibles respuestas de las que debe elegir sólo una.

Con el fin de estudiar la complejidad que supone la automatización de encuestas telefónicas, la compañía ODEC proporcionó un total de 75 ficheros de audio. Cada uno de estos ficheros contiene la grabación de una encuesta telefónica. El guión de la encuesta grabada en todas las muestras es el mismo, variando la persona encuestada y, aunque en menor frecuencia, los encuestadores/as.

En las grabaciones se observan una serie de comportamientos que aumentan enormemente la complejidad de la automatización de este tipo de encuestas. El hecho de que la encuesta sea realizada por una persona condiciona totalmente el comportamiento de la persona encuestada, siendo éste muy diferente a cómo se comportaría si la encuesta fuese realizada por una máquina. Concretamente se observan comportamientos como:

- Conversaciones fuera del ámbito de la encuesta entre la persona encuestada y encuestador.
- Formulación de opiniones personales de las personas encuestadas ante ciertas preguntas sin responder entre las opciones debidas. En estos casos el encuestador intenta reconducir a la persona encuestada para que se limite a contestar entre las posibles opciones.
- En muchas ocasiones la persona encuestada responde a la pregunta haciendo alusión a respuestas anteriores. Lógicamente el nivel de comprensión de un humano no es comparable al de una máquina.
- El encuestado no sólo emite la respuesta concreta, sino que puede incluir antes y después de la opción elegida cualquier tipo de frase.

Esta serie de dificultades, propiciadas por el marco en el que se desarrolla la encuesta, impiden la aplicación satisfactoria de un sistema de reconocimiento automático del habla que permita la transcripción automática de las respuestas de la encuesta. La automatización de encuestas telefónicas requiere pues que el procedimiento de realización de la encuesta sea distinto.

La transcripción automática de las respuestas de la encuesta puede lograrse, con un alto grado de acierto (ver sección 4), siempre y cuando se cumplan una serie de requisitos. Hay que tener en cuenta que el prototipo desarrollado (ver sección 3) sea capaz de realizar la encuesta, simplemente utilizando la información de cuáles son las posibles respuestas para cada pregunta. Éste es un aspecto interesante, ya que el prototipo no necesita de supervisión externa, aunque limita sus prestaciones o, más bien, el modo en que se debe interactuar con él.

Para que la transcripción automática de las respuestas sea posible es necesario que la encuesta se limite a ser una sucesión de preguntas-respuestas. El sistema hace la pregunta (ofreciendo las posibles respuestas) y el encuestado se limita a contestar una de las posibles opciones. La respuesta emitida no tiene por qué seguir exactamente la sintaxis de las posibles contestaciones, sino que existe un margen para la omisión o inserción de palabras. El sistema es capaz de identificar cuales son las palabras relevantes y, en consecuencia, transcribir la opción escogida. Éste sería el comportamiento lógico si la encuesta fuese realizada por una máquina, ya que se elimina la posibilidad de que el encuestado incurra en los comportamientos indicados anteriormente.

Para hacer frente a lo anterior, los modelos de lenguaje y de léxico que se han utilizado con el motor de reconocimiento del habla se generan en forma automática a partir de las posibles respuestas de cada pregunta, sin ningún tipo de supervisión. Los modelos léxico (palabras) son directamente representados por autómatas de estados finito que modelan la concatenación secuencial de sus caracteres constituyentes. Para los modelos de lenguaje se probaron diferentes tipos y configuraciones, incluyendo aquellos modelos (gramáticas) que solo aceptan la frase misma de la respuesta, pasando por modelos que buscan palabras claves dentro de la frase reconocida (word spotting), hasta llegar a los modelos n-gramas [5, 6]. Todos ellos son representados también por autómatas de estados finitos (ver figura 3).

3. EL PROTOTIPO ADET

El esquema cliente-servidor del prototipo ADET es mostrado esquemáticamente en la figura 1. El cliente se encuentra conformado por una estación de trabajo y un teléfono, que interactúan a través de una conexión de internet y una conexión telefónica (respectivamente), con un servidor donde se lleva a cabo el proceso de reconocimiento de las repuestas enviadas por teléfono. En la estación de trabajo del cliente se ejecuta una interfaz, desarrollada completamente en JAVA (ver figura 2), donde el operario, entre otras cosas, puede:

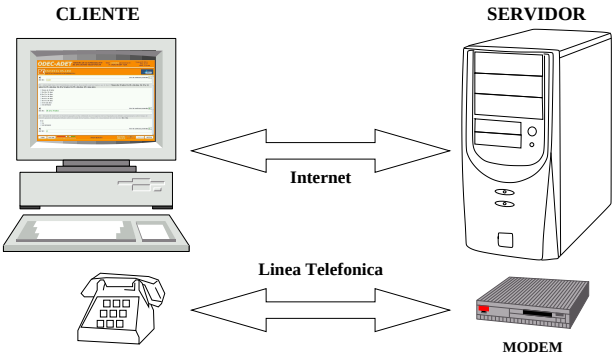

Figura 1. Esquema cliente-servidor del prototipo ADET.

1. Visualizar la secuencia de preguntas a medida que se van formulando (guión de la encuesta) y que se

pueden oír simultáneamente a través del auricular del teléfono.

2. Visualizar las posibles clases de respuestas esperadas, de estilo *multiple-choise*, y de las cuales se espera una contestación aproximada a algunas de ellas.

3. Visualizar de la respuesta reconocida (enviada por el servidor) conjuntamente con una medida de confianza [1, 2] sobre la certeza de ese reconocimiento.

4. Interactuar con el sistema por medio de una serie de controles, como por ejemplo cancelar, repetir, o finalizar la encuesta, fijar niveles de confianza del reconocimiento, etc.

Figura 2. Imagen de la interfaz gráfica del prototipo ADET.

La parte del servidor se compone principalmente de un PC y de un módem externo con el cual se establece la conexión telefónica con el cliente y por donde se reciben sus respuestas de voz. En la PC del servidor, principalmente se ejecuta un motor de reconocimiento del habla continuo operativo, denominado "ATROS" [3], desarrollado completamente en el *Instituto Tecnológico de Informática* (ITI - www.iti.es). Entre las características más sobresalientes del mismo se encuentran:

a - su flexibilidad de configuración mediante parámetros, que permiten adaptarlo a los requerimientos de memoria y tiempos de respuesta en la adquisición y reconocimiento de la señal acústica telefónica.

b - su capacidad para cargar en memoria diferentes modelos de lenguaje y poder cambiar entre ellos en función de las respuestas diferentes que deban reconocerse para cada pregunta de la encuesta.

c - compatibilidad con el formato de los modelos acústicos de Markov (HMM) utilizado por la herramienta de reconocimiento del habla HTK [4]. El entrenamiento de los **modelos acústicos** HMM para el reconocimiento de voz telefónica a 8KHZ fue realizado con este software (HTK).

Además del motor de reconocimiento del habla, en el servidor se ejecutan programas encargados de:

1 - Facilitar la interacción entre módem y motor de reconocimiento del habla, mediante un protocolo de comunicación módem-motor-de-reconocimiento y códigos-mensajes de error (previamente definidos) para evaluación del estado respectivo de ambos.

2 - La gestión y ejecución del flujo de una encuesta dada, así como la consulta del estado de evolución de la misma.

Para el intercambio de información/datos entre los procesos y programas que se ejecutan en el cliente y el servidor, se ha desarrollado e implementado un protocolo de comunicación entre ambos a través del canal de internet (TCP/IP, utilizando sockets). Así también se implementaron procesos destinados a mantener la sincronización de la transferencia de información del canal de internet y la del canal telefónico.

Como se ha mencionado anteriormente, una de las características importantes del prototipo ADET es la incorporación de las llamadas *medidas de confianza*. En cualquier problema de reconocimiento de formas, ya sea reconocimiento automático del habla, traducción automática, o reconocimiento de texto manuscrito continuo, los sistemas automáticos no son infalibles y cometen errores. Añadir a estos sistemas la capacidad de detectar cuándo se ha cometido un error es de gran interés, ya que permite al propio sistema reaccionar ante sus propios errores. Dentro de nuestro contexto, las mismas darían la opción al sistema de poder rechazar la respuesta si esta no supera un determinado umbral de confianza pre-establecido, y volver a realizar la pregunta. Este es el objetivo que persigue la estimación de medidas de confianza. La estimación de medidas de confianza se ha aplicado extensamente en el área de reconocimiento automático del habla [1].

4. EVALUACIÓN

Para evaluar el prototipo se ha utilizado una encuesta facilitada por ODEC, la cual consiste básicamente en una serie de preguntas y posibles respuestas asociadas (*multiple-choise*) sobre la situación política en España. Las preguntas fueron clasificadas según el posible conjunto de respuestas. Es decir, todas aquellas preguntas cuyo posible conjunto de respuestas coincide, pertenecen a la misma clase o tipo. Para la encuesta dada, los tipos identificados, son mostrados en la tabla 1. El número total de preguntas evaluadas es de 66.

Como muestras de voz para la evaluación se han utilizado dos corpus:

Corpus ODEC: De los 75 ficheros de audio proporcionados por ODEC se seleccionaron 6 encuestas para evaluación. El criterio para seleccionarlas fue comprobar que en ellas se seguía, más o menos, el paradigma pregunta/respuesta. Con las encuestas se-

Tipo	Nº de Preg.	Respuestas
1	15	Sí No No sabe
2	10	Mucho Bastante Poco Nada No sabe No contesta
3	13	Muy bien Bastante bien Bastante mal Muy mal Regular No recuerda No sabe No contesta
4	13	Ha mejorado Ha empeorado Sigue igual No conoce No sabe No contesta
5	13	Con la mayoría de las cosas que dijo Con bastantes de las cosas que dijo Con pocas de las cosas que dijo Con nada o casi nada de lo que dijo No sabe No contesta
6	2	Situación económica / asuntos económicos La política antiterrorista Desarrollo del Estado autonómico / Reforma de los estatutos de autonomía Modelo de financiación autonómica Estabilidad del gobierno Inmigración Temas sociales (sanidad / educación / etc.) Otros Ninguno en especial No sabe No contesta

Tabla 1. Agrupación en 6 clases diferentes de las respuestas más comunes de la encuesta facilitada por ODEC, sobre la temática política en España.

leccionadas se procedió a la extracción manual de los segmentos de voz correspondientes a las contestaciones de las preguntas de evaluación. Para cada muestra de audio extraída fue necesario transcribir cuál era la respuesta escogida.

Corpus ITI: Utilizando el prototipo se procedió a que personal del ITI realizase la encuesta. De todas las encuestas adquiridas se seleccionaron al azar 6, correspondientes a personas distintas. Para cada muestra de audio fue necesario también transcribir cuál era la respuesta escogida.

En la evaluación se experimentaron con distintos modelos del reconocedor de habla. Todos estos modelos fueron generados siempre de forma automática a partir de las posibles respuestas de cada pregunta. Como se mencionó en la sección anterior, se probaron diferentes tipos y configuraciones de modelos de lenguaje, representados por autómatas de estados finitos. Los mismos se caracterizan por el grado de restricción en la aceptación de una respuesta dada, y van desde aceptar las respuestas esperadas tal cual están escritas, hasta especializarse en buscar las palabras claves que caracterizan a cada una de las mismas. La figura 3 muestra un ejemplo de ello. Hay que

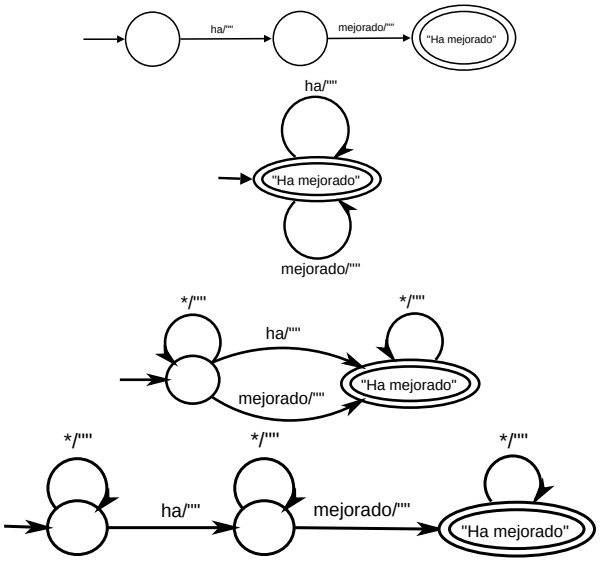

Figura 3. Modelos de lenguaje representados por autómatas de estado finito (los cuales son también transductores) que modelan la frase “ha mejorado”. De arriba hacia abajo: el primero acepta solo la frase “ha mejorado”; el segundo acepta una secuencia de las palabras “ha” y/o “mejorado” en cualquier orden y número de repeticiones; el tercero acepta una secuencia de fonemas cualquiera (representado por el *) + “ha” o “mejorado” + secuencia cualquiera de fonemas; finalmente el último modelo acepta una secuencia de fonemas cualquiera + “ha” + secuencia de fonemas cualquiera + “mejorado” + secuencia de fonemas cualquiera.

destacar que cada uno de estos autómatas que modelan

el conjunto de respuestas a cada pregunta son en realidad transductores, porque cada uno de sus respectivos estados finales emite una de las posibles cadenas de respuesta (ver ejemplo en figura 4).

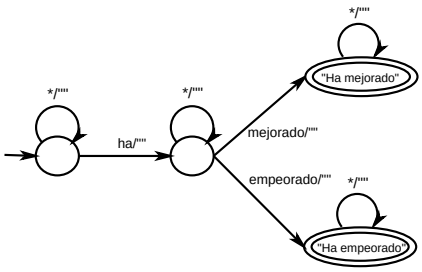

Figura 4. Ejemplo de un transductor que modela las dos posibles respuestas: "Ha mejorado" y "Ha empeorado" aceptando como entrada cadenas *aproximadas* a algunas de ellas, como por ejemplo: la frase "crea que ha mejorado mucho" sería interpretada como "Ha mejorado" y la frase "pues ha empeorado bastante" como "Ha empeorado".

Como medida de las prestaciones del prototipo se ha calculado la tasa de aciertos de reconocimiento de respuesta. Esta tasa mide el tanto por cien de frases pronunciadas, en las que el prototipo identificó correctamente cuál era la respuesta escogida por el entrevistado. Hay que resaltar que el encuestado puede emitir la frase sin seguir exactamente la sintaxis de las respuestas posibles, y el prototipo, como resultado del reconocimiento de la frase, obtiene, gracias a la utilización de transductores, la transcripción de la respuesta seleccionada.

En las siguientes tablas 2 y 3 se muestra el acierto del prototipo ADET (en tanto por cien (%) y número absoluto (Abs)) en asignar la respuesta correcta en función de lo que dice el encuestado. Como ya se ha indicado no siempre el encuestado contesta exactamente siguiendo la sintaxis de la respuesta. Lógicamente esto es más frecuente en el corpus ODEC.

En la tabla 2 se muestran los mejores resultados, para cada corpus de voz, con la restricción de que los modelos utilizados por el reconocedor se generan automáticamente siguiendo la misma metodología. Éste sería el caso de un sistema totalmente automático exento de supervisión.

Estos resultados pueden ser mejorados añadiendo una mínima supervisión. Para algún tipo de pregunta los modelos que son más flexibles obtienen mejores resultados, mientras que para otros tipos ocurre al contrario. En la tabla 3 se muestran los mejores resultados utilizando en cada pregunta el modelo que se comporta mejor.

Como se observa, una mínima evaluación de los modelos generados automáticamente permitiría escoger los modelos apropiados para cada pregunta, repercutiendo en una mejora de las prestaciones del sistema.

5. CONCLUSIONES

El estudio de los ejemplos de encuestas proporcionadas por ODEC permite concluir que la automatización de encuestas telefónicas no es posible en un entorno en el que la persona encuestada no tiene ninguna restricción a la hora de expresarse. Los sistemas automáticos de reconocimiento del habla obtienen buenos resultados siempre y cuando el vocabulario con el que deben tratar esté limitado. Justamente en esta tarea estos sistemas pueden ser muy útiles, ya que el vocabulario viene restringido por el posible conjunto de respuestas de cada pregunta. Lo único necesario es que la encuesta se realice bajo unas condiciones determinadas, de tal forma que la persona encuestada se limite a responder con cierta flexibilidad alguna de las posibles respuestas. Como se ha visto en los resultados presentados, bajo estas condiciones, el prototipo desarrollado es capaz de transcribir correctamente entre el 83 y el 87 % de las respuestas (corpus ITI). En el caso del corpus de ODEC, en el que las respuestas pueden incluir una mayor espontaneidad, dentro de un margen razonable, el acierto del prototipo también es aceptable (entre un 69 y 78 %).

El prototipo desarrollado en este proyecto es capaz de realizar encuestas telefónicas generando los modelos necesarios de forma totalmente autónoma y automática. Este aspecto proporciona la posibilidad de que sea utilizado sin necesidad de tener conocimientos en reconocimiento automático del habla. Podría ser idóneo en aplicaciones ya incipientes, como son encuestas telefónicas automáticas que se realizan, por ejemplo, para obtener datos sobre la satisfacción de clientes ante un servicio dado.

6. BIBLIOGRAFÍA

[1] A. Sanchis, *Estimación y aplicación de medidas de confianza en reconocimiento automático del habla*, Ph.D. thesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, May 2004.

[2] A. Sanchis, A. Juan, y E. Vidal, "Improving utterance verification using a smoothed naive bayes model," in *IEEE International Conference on Acoustic, Speech and Signal Processing*. April 2003, vol. 1, pp. 592–595, IEEE Press.

[3] M. J. Castro, D. Llorens, J.A. S'anchez, F. Casacuberta, P. Aibar, y E. Segarra, "A fast version of the atros system," in *European Conference on Speech Communication and Technology. EUROSPEECH'99*, Budapest, September 1999, pp. 1299–1302.

[4] S. Young, J. Odell, D. Ollason, V. Valtchev, y P. Woodland, *The HTK Book: Hidden Markov Models Toolkit V2.1*, Cambridge Research Laboratory Ltd, Mar. 1997.

	Corpus ITI			Corpus ODEC		
Tipo	Nº Muestras	Acierto (%)	Abs	NºMuestras	Acierto (%)	Abs
1	90	94.4	85	69	76.8	53
2	60	81.7	49	49	75.5	37
3	78	80.8	63	38	55.2	21
4	78	71.8	56	24	70.8	17
5	78	89.7	70	32	62.5	20
6	12	75.0	9	15	60.0	9
TOTAL	396	83.8	331	227	69.2	157

Tabla 2. Resultados de evaluación del prototipo ADET para el caso de un sistema totalmente automático exento de supervisión.

	Corpus ITI			Corpus ODEC		
Tipo	NºMuestras	Acierto (%)	Abs	NºMuestras	Acierto (%)	Abs
1	90	98.9	89	69	95.7	66
2	60	88.3	53	49	87.8	43
3	78	87.2	68	38	55.2	21
4	78	73.1	57	24	70.8	17
5	78	89.7	70	32	65.6	21
6	12	75.0	9	15	60.0	9
TOTAL	396	87.4	346	227	78.0	177

Tabla 3. Resultados de evaluación del prototipo ADET utilizando en cada pregunta el modelo que se comporta mejor.

[5] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1998.

[6] Slava M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-35, pp. 400–401, Mar. 1987.