Tecnologías del habla para el desarrollo del lenguaje

Carlos Vaquero, Óscar Saz, Eduardo Lleida

Grupo de Tecnologías de las Comunicaciones (GTC),
Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza
{cvaquero, oskarsaz, lleida}@unizar.es

***Abstract* — The work described in this paper aims to make use of speech technologies to help those who suffer from language pathologies. For this purpose, the Aragon Institute for Engineering Research (I3A), with the collaboration of the Public School for Special Education Alborada, has developed an application for computer aided speech therapy in Spanish language called Vocaliza, which will help the daily work of speech therapists who train the linguistic skills of Spanish speakers with different speech disabilities. Vocaliza is free to distribute and has been successfully installed in two schools for special education in Zaragoza.**

In addition, it is shown that Maximum Likelihood Linear Regression (MLLR) based speaker adaptation can be used to improve the performance of an Automatic Speech Recognition (ASR) system, when used by people with speech pathologies. This speaker adaptation technique was tested on the Alborada Corpus obtaining a relative improvement of 77.15% in ASR performance.

I. Introducción

Las tecnologías del habla tales como el Reconocimiento Automático del Habla (RAH) o la conversión Texto Habla ("Text To Speech", TTS) se están convirtiendo en un instrumento cada vez más común en nuestra Sociedad de la Información. Actualmente ya son habituales los interfaces orales hombre-máquina tales como los sistemas de telefonía móvil manos libres capaces de reconocer el habla de usuario y averiguar el contacto al que desea llamar, o los interfaces orales reconocedores de comandos, que permiten acceder y controlar mediante la voz el estado de ciertos dispositivos electrónicos. Sin embargo, a medida que se extienden estas tecnologías, el distanciamiento social que sufren aquellas personas que padecen alguna discapacidad en el habla se acentúa, ya que todos estos sistemas constituyen nuevas barreras para ellos, debido a la pronunciada degradación de las prestaciones de los mismos ante este tipo de voz.

Por otro lado, las Tecnologías del Habla pueden ser una herramienta muy útil para ayudar a aquellas personas que sufren patologías en el habla a mejorar su capacidad de comunicación. En efecto, la demanda de aplicaciones de apoyo a la logopedia ha crecido notablemente en los últimos años, conforme la fiabilidad de estos sistemas ha aumentado y la tecnología ha sido más asequible. A pesar de ello, no existe demasiada oferta en este ámbito. El sistema de apoyo a la logopedia comercial más popular ha sido la aplicación Speech Viewer de IBM, que no se encuentra disponible en castellano. En el ámbito de la investigación, se han llevado a cabo numerosos proyectos relacionados con el uso de las tecnologías del habla como apoyo a la logopedia, tales como los proyectos Orto Logo-Paedia [1], SPECO [2], ISAEUS [3] y HARP [4], pero todos ellos han tenido escasa transferencia de resultados aprovechables para la lengua castellana. Todo ello dificulta el uso de las aplicaciones disponibles por parte del colectivo de logopedas de nuestro país en sus labores cotidianas.

Por ello, el Instituto de Investigación en Ingeniería de Aragón (I3A) de la Universidad de Zaragoza, contando con la colaboración de profesionales en el ámbito de la logopedia del Colegio Público de Educación Especial Alborada (CPEE Alborada), presenta en este artículo un trabajo de investigación cuyo objetivo fundamental es el de poner las tecnologías del habla al servicio de aquellas personas que padecen discapacidades de comunicación. A continuación se comenta el trabajo que se ha realizado, organizado de la siguiente forma: En la sección 2 se describe una aplicación informática de apoyo a la logopedia en castellano y de libre distribución que se ha desarrollado en este trabajo. En la sección 3 se propone el uso de técnicas de adaptación al locutor para acercar el uso de reconocedores de comandos orales a personas con discapacidad y finalmente, en la sección 4 se comentan las conclusiones de este trabajo.

II. La aplicación vocaliza

La aplicación de libre distribución desarrollada en el marco de este trabajo de investigación, llamada Vocaliza [5], tiene como objetivo fundamental el de servir de apoyo a la logopedia, aunque incluye igualmente un módulo de adquisición de datos y adaptación al locutor que facilita las labores de adquisición de corpora o bases de datos de habla patológica y la creación de modelos acústicos adaptados al usuario. A continuación se comentan las principales características de Vocaliza.

A. Interfaz de usuario

Puesto que muchos de los usuarios que van a utilizar la aplicación Vocaliza son niños o personas con pocos conocimientos de informática, a la hora de diseñar el interfaz de la aplicación, que puede apreciarse en la Figura 1, se ha pretendido conseguir una aplicación entretenida y a la vez sencilla de utilizar. Para ello Vocaliza ofrece juegos como técnicas de trabajo del habla, juegos que funcionan únicamente con la voz. Todos estos juegos tienen un funcionamiento similar: el usuario debe

Fig. 1. Interfaz gráfico de la aplicación Vocaliza: Ventana principal.

pronunciar, en base a una serie de pistas en forma imágenes, texto y sonidos, la palabra o frase que la aplicación le solicita. Si el usuario pronuncia correctamente dicha palabra o frase, superará el juego con éxito y éste volverá a comenzar, tras mostrar una animación que indica el éxito del usuario y su puntuación. De lo contrario, la aplicación esperará un tiempo mientras el usuario intenta pronunciar la locución correspondiente antes de volver a comenzar el juego.

De esta forma, una vez un usuario inicia un juego, sólo necesita un micrófono y su voz para llevarlo a cabo, no necesita utilizar ni el teclado ni el ratón, lo que permite que los usuarios sin conocimientos de informática puedan utilizar la aplicación sin supervisión, una vez ésta haya sido configurada por un logopeda para tratar la patología concreta del usuario.

B. Apoyo a la logopedia

Como aplicación de apoyo a la logopedia, Vocaliza permite trabajar 3 niveles del lenguaje: fonológico, morfo-sintáctico y semántico. Cada nivel del lenguaje se trabaja mediante un juego, que el logopeda puede configurar para trabajar con más precisión la patología concreta de cada usuario.

El nivel fonológico, que trata la correcta pronunciación de los fonemas y de los grupos de fonemas que aparecen en una lengua, se trabaja mediante el juego de pronunciación. Dicho juego obliga al usuario a pronunciar un conjunto de palabras concretas elegidas por el logopeda, indicando posteriormente mediante una calificación si lo ha hecho correctamente o no.

El nivel morfo-sintáctico trata la correcta construcción de sintagmas y oraciones en base a las reglas gramaticales existentes en una lengua, y Vocaliza permite trabajarlo mediante el juego de las frases. Dicho juego obliga al usuario a pronunciar una frase, tratando de dar al usuario conciencia de la función de cada elemento en la oración. Para superar el juego con éxito, el usuario debe pronunciar correctamente la frase, respetando el orden gramaticalmente correcto.

El nivel semántico, que trata la correcta asociación de locuciones (palabras) con ideas o significados, se trabaja mediante el juego de las adivinanzas, que plantea un acertijo al usuario ofreciendo varias posibles respuestas, de forma que, para superarlo con éxito, el usuario debe pronunciar adecuadamente la respuesta correcta.

Además de los juegos mencionados, Vocaliza ofrece el juego de la evocación, en el que el usuario puede pronunciar la palabra que desee siempre que ésta pertenezca a un grupo semántico previamente definido por el logopeda, y si lo hace correctamente, la aplicación mostrará una imagen relacionada con la palabra pronunciada en la pantalla. Dicho juego ofrece nuevas posibilidades de trabajo a nivel fonológico y semántico.

C. Perfil de usuario

Para facilitar el trabajo del logopeda en los casos en los que se utiliza un mismo equipo informático para trabajar con distintos usuarios, la aplicación Vocaliza permite la creación de perfiles de usuario. Estos perfiles permiten almacenar toda la información de usuario relacionada con la configuración de los distintos juegos, pero también permite almacenar grabaciones de locuciones del usuario, que posteriormente pueden ser utilizadas para analizar la evolución del usuario o en un proceso de adaptación al locutor.

D. Tecnologías del habla

Toda la funcionalidad que ofrece la aplicación Vocaliza es posible gracias al uso de diversas tecnologías del habla, concretamente a la integración de un sistema de RAH, un sistema de TTS, un sistema de adaptación al locutor y un sistema de Verificación de Pronunciación (“Utterance Verification” UV), cuyas características se explican a continuación:

El sistema de RAH constituye el núcleo de la aplicación Vocaliza. La aplicación utiliza el sistema para decidir que palabras o frases pronuncia el usuario en los juegos en cada momento, de forma que la aplicación puede determinar si éste ha tenido éxito o no en el juego. El sistema de RAH utilizado en la aplicación utiliza como entrada la señal de voz adquirida mediante el micrófono con una frecuencia de muestreo de 16 KHz y una precisión de cuantificación de 16 bits, que posteriormente es enventanada con una ventana de Hamming de 25 ms, con un solape de 15 ms. Los vectores de características que se utilizan

para el reconocimiento son vectores con 37 coeficientes Mel-cepstrum (MFCC), que consisten en 12 parámetros estáticos, sus primeras y segundas derivadas y la derivada del logaritmo de la energía. Para el modelado acústico se utilizan Modelos Ocultos de Markov (HMM), que se componen de 822 unidades dependientes del contexto, un modelo de silencio y un modelo de transición entre palabras, lo que supone un total de 824 unidades. Cada unidad se modela con un único estado y la función de densidad de probabilidad asociada a cada estado se modela con una mezcla de 16 gaussianas.

El sistema TTS permite a la aplicación ofrecer información auditiva al usuario sobre aquello que debe pronunciar en los distintos juegos, y ocasionalmente, puede servir de guía sobre la adecuada pronunciación de una palabra o frase.

El sistema de adaptación al locutor permite crear modelos acústicos adaptados al usuario, para que luego éstos puedan ser utilizados cuando el usuario trabaje con la aplicación, o bien puedan ser extraídos de la aplicación para su uso en otros sistemas, tales como interfaces orales para el control de entornos inteligentes, que faciliten la vida al usuario si éste padece, además, alguna discapacidad motriz. Vocaliza incluye un módulo de adquisición de datos que permite registrar el habla del usuario, asocia los datos adquiridos al usuario que esté utilizando la aplicación, y permite realizar un proceso de adaptación al locutor con los datos adquiridos. El sistema de adaptación al locutor se basa en el algoritmo del Máximo a Posteriori (MAP) [6], ya que permite obtener modelos adaptados al usuario que mejoran notablemente las prestaciones del sistema de RAH disponiendo de pocas muestras del habla del usuario. Esto es particularmente útil en este tipo de aplicaciones, en las que el usuario encuentra muy fatigoso el proceso de adquisición debido justamente a la patología presente en su habla.

El sistema de UV permite ofrecer una medida de la evolución de la pronunciación del usuario a partir de una locución que éste realice y de un modelo adaptado al mismo, de forma que los juegos pueden ofrecer una calificación de acuerdo con la pronunciación realizada por el usuario. El sistema de UV se basa en el Ratio de Verosimilitud (LR UV) [7] para determinar si la pronunciación del usuario está más próxima al modelo adaptado a su habla patológica o de un modelo de habla correcta, y en base a dicha medida, la aplicación mostrará una calificación orientativa para el usuario.

III. Reconocimiento de comandos orales

Cada vez más, los sistemas de RAH están demostrando unas muy buenas prestaciones, con tasas de error muy bajas, que hacen que su uso se esté extendiendo a la vida diaria de los usuarios finales. Estas tasas de error se mantienen en niveles muy bajos en situaciones de entorno controlado y locutor cooperativo; sin embargo, decaen dramáticamente en casos de entornos ruidosos o muy variables, o cuando el locutor es poco cooperativo con el sistema o presenta alguna deficiencia en su habla que lo aleja del habla más habitual. Esa situación se da en el caso de personas con algún tipo de discapacidad o patología en el habla, debido a que el modelado acústico se realiza para habla normalizada. Aún así, se pueden utilizar algoritmos de adaptación al locutor que permiten modelar el habla del usuario concreto que se pretende use el sistema, obteniendo unas tasas de reconocimiento más similares a las que se obtienen en la situación ideal. Por ello, se planteó la posibilidad de utilizar estas estrategias como un medio de cara a acercar a personas con deficiencias a los sistemas de interfaces orales.

Para ello, el primer paso fue la adquisición de un corpus de habla patológica infantil en colaboración con el CPEE Alborada. Este corpus contiene el habla de 14 alumnos del centro (7 chicos y 7 chicas) de entre 11 y 21 años y con diversas afecciones en su desarrollo mental que llevan asociadas alteraciones en su lenguaje de carácter fonológico, morfológico/sintáctico y semántico-pragmático. El corpus se compone de 4 repeticiones por parte de cada alumno de las 57 palabras del Registro Fonológico Inducido (RFI) [8], para un total de 3192 alocuciones en el corpus. Para evaluación y referencia, se ha adquirido también un corpus de habla infantil y juvenil en colaboración con el Colegio de Educación Infantil y Primaria (CEIP) “Río Ebro” y el Instituto de Enseñanza Secundaria (IES) “Tiempos Modernos” de la ciudad de Zaragoza. El corpus consta de 168 locutores, chicos y chicas de entre 10 y 18 años repitiendo cada uno de ellos todas las palabras del RFI una vez, para un total de 9576 realizaciones de voz. Todo el corpus se adquirió en las instalaciones de los tres centros educativos colaboradores, con un micrófono tipo close-talk inalámbrico modelo AKG C444L.

Para este trabajo, se planteó la utilización del algoritmo Maximum Likelihood Linear Regresion (MLLR) [9] como algoritmo de adaptación al locutor por sus buenos resultados en otras tareas. Se pretendía estudiar la influencia del tipo de unidad acústica empleada y la influencia de la cantidad de material de voz usado para realizar la adaptación. Como unidades acústicas a estudio se planteó el uso de unidades de palabra y de unidades subfonéticas contextuales. Para cada una de estas posibles unidades se plantea el entrenamiento con una sola serie de 57 palabras, dos series (114 palabras) o tres series (171 palabras). En todos los casos se usan el resto de series para obtener los resultados de RAH, manteniendo el criterio de independencia entre los datos de entrenamiento y de test.

El resultado de la tarea de reconocimiento propuesta para estos locutores (57 palabras aisladas) nos da una tasa de error de un 52,22% de promedio entre los 14 locutores. El peor de los locutores obtiene un 89,47% de tasa de error, mientras que el mejor de los locutores tiene un 13.16% de error. El habla de referencia infantil y juvenil obtiene una tasa de error del 15,99% por el desajuste existente debido a que los modelos usados en reconocimiento son entrenados con bases de datos de habla adulta (Spanish SpeechDat-Car). La línea de base es la misma para ambos tipos de unidades, ya que los modelos de palabra se obtienen inicialmente como concatenación de las unidades subfonéticas que forman la palabra.

Los resultados promedios para los usuarios de habla patológica (Figura 2) muestran que las unidades subfonéticas presentan una tasa de error menor que las unidades de palabra para todos los casos de tamaño del material de entrenamiento,

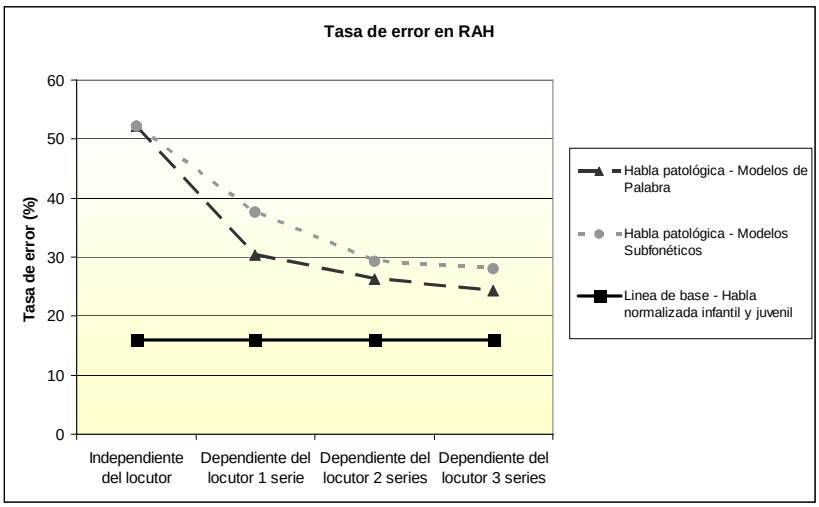

Fig. 2. Tasa de error dependiendo del tamaño de la adaptación

especialmente en el caso de menor cantidad de material para la adaptación (37,63% de error con palabras frente a 30,27% de error con subfonemas). Se puede apreciar como la mejora incremental disminuye según se aumenta la cantidad de material para la adaptación, y como en el caso de unidades subfonéticas se obtiene una tasa de error de 24,27% (un 77,15% más cercano a la línea de base del habla normalizada).

IV. Conclusiones

Como resultado de este trabajo, se ha desarrollado una aplicación de apoyo a la logopedia libre y en castellano, que cubre las necesidades de los profesionales de la logopedia en lengua castellana, trabajando el lenguaje a nivel fonológico, morfo-sintáctico y semántico. La aplicación está actualmente instalada en dos centros de educación especial de la comunidad de Aragón, y en breve estará disponible mediante una página web para cualquier logopeda o usuario que pueda necesitarla.

La gran acogida que ha tenido Vocaliza en los centros de educación especial en los que se ha instalado es muy alentadora para seguir trabajando en esta línea de investigación, y en los próximos meses se pretende desarrollar una serie de juegos que complementen la aplicación trabajando niveles del lenguaje más triviales, como la respiración o la entonación.

Por otra parte, se ha probado que existen técnicas que permiten a los usuarios más jóvenes con patologías en el lenguaje aproximarse de forma fructífera al uso de sistemas de reconocimiento de comandos orales, decrementando la tasa de error relativa que obtendrían ante este tipo de sistemas en un 77,15% mediante el uso de técnicas de adaptación al locutor.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia de España a través del Proyecto Nacional TIN 2005-08660-C04-01.

Referencias

[1] A. Protopapas, A. Öster, D. House and A. Hatzis, “Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia)”, *TMH-QPSR*, vol. 44, Fonetik, 2000.

[2] A. Öster, Z. Kacic, Z. Barczikay, K. Vicsi, P. Roach, and I. Sinka, “SPECO – a multimedia multilingual teaching and training system for speech handicapped children”, *Tech. Rep, Eurospeech, 6th Conference on speech Communication and Technology, Interspeech,* 1999.

[3] García Gómez et al., “Isaeus, speech training for deaf and hearing-impaired people”, SPECO – a multimedia multilingual teaching and training system for speech handicapped children”, *Tech. Rep, Eurospeech, 6th Conference on speech Communication and Technology*, *Interspeech*, 1999.

[4] “Harp – an autonomous speech rehabilitation system for hearing-impaired people”, *Final Report, HARP (TIDE project 1060),* May 1996.

[5] C. Vaquero, O. Saz, E. Lleida, “Vocaliza, an Application for Ciomputer-Aided speech therapy in Spanish Language”, *IV jornadas en tecnologías del habla, Zaragoza, España,* pp. 321-326, 2006.

[6] J. L. Gauvain and C. H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains”, *IEEE Transactions on speech and Audio Processing*, vol. 2, nº 2, pp. 291-298, 1994.

[7] E, Lleida and R. C. Rose, “Utterance Verification in continuous speech recognition: decoding and training procedures”, *IEEE Transactions on speech and Audio Processing*, vol. 8, nº 2, pp. 126-139, 2000.

[8] M. Monfort, A Juárez Sánchez, “Registro Fonológico Inducido (Tarjetas Gráficas)”, *Ed. CEPE*, Madrid, 1989.

[9] C.J. Legetter and P.C. Woodland, “Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models”, *Computer Speech and Language*, vol. 9, pp. 171—185, 1995.