Aplicación de Tecnología de Voz en la Enseñanza del Español[1]

Ingrid Kirschning Nancy Aguas Angélica Ahuactzin

TLATOA - Grupo de Procesamiento Automático de Voz
ICT, CENTIA[2],
Universidad de las Américas- Puebla.
Sta. Catarina Mártir s/n, 72820 Cholula, Pue., México.
e-mail: ingrid@mail.udlap.mx
Tel: 2-29-2623, FAX: 2-29-2138

Palabras Clave: Tecnologías de Voz, Enseñanza de un segundo lenguaje, Verificación de Pronunciación, Reconocimiento de Voz, Redes Neuronales.

Resúmen

Este artículo presenta dos herramientas desarrolladas con tecnología de voz para la enseñanza del Español hablado en México. Estas son un diccionario Inglés-Español accesado por medio de la voz, y un nuevo método para la verificación de la pronunciación correcta de palabras o frases. Esta segunda herramienta es especialmente útil en los sistemas para la enseñanza de un lenguaje por medio de la computadora (CALL - Computer Aided Language Learning). Este método aprovecha las técnicas de reconocimiento de voz y las herramientas del CSLU Toolkit (del Center for Spoken Language Understanding, Oregon Graduate Institute) para reconocer la secuencia de sonidos emitidos por el usuario y marcar las partes mal pronunciadas. Cada frase o palabra pronunciada por el usuario puede evaluarse fonema por fonema detectando los errores en que incurrió el locutor. Para ello es necesario entrenar un sistema de reconocimiento de voz (una red neuronal en este caso) con los fonemas del lenguaje objetivo además de incluir los fonemas del idioma nativo del locutor. Esto se debe a que los errores de pronunciación de alguien que esta aprendiendo un nuevo idioma son causados por las costumbres de pronunciación del locutor en su lengua materna. A diferencia de otros sistemas existentes, con este método se puede proveer una retroalimentación explícita y clara al usuario, el cual podrá entonces estudiar e intentar pronunciar las frases que contengan esas partículas que requiere practicar.

1. Introducción

Gracias a los avances tecnológicos referente a la velocidad de procesamiento de las computadoras, espacio de almacenamiento y manejo de dispositivos de audio, las tecnologías de voz ya son una realidad en casi cualquier tipo de sistema. Las aplicaciones ya pueden utilizarse en computdoras personales, teléfonos celulares, etc. Esto hace a esta impresionante tecnología accesible a cualquiera.

[1] **Investigación realizada gracias al apoyo de CONACyT, proyecto # I28247-A**

[2] **Centro de Investigación en Tecnologías de Información y Automatización**

Entre las aplicaciones que más sentido tienen en la aplicación de interfaces de voz son los sistemas para la enseñanza de lenguajes. Ya existe una gran variedad de productos que utilizan la entrada de voz en cursos de idiomas, los cuales permiten a un estudiante practicar un nuevo idioma. Sin embargo, no se ha encontrado ninguno que provea una retroalimentación explícita y clara de los errores que hace el locutor. Los sistemas existentes sólo dan calificaciones generales para una palabra o muestran un par de ondas sonoras, una la que debió haber dicho y otra, la que dijo el locutor, sin indicar en dónde está el error o la razón de su baja calificación.

Este artículo presenta un método que utiliza un sistema de reconocimiento de voz para analizar explícitamente la pronunciación correcta de palabras o frases. Asimismo se introduce una interfaz de voz para accesar un diccionario bilingüe. Los resultados mostrados aquí pertenecen a dos trabajos de tesis de licenciatura de estudiantes del grupo de investigación en tecnologías del habla "TLATOA" (CENTIA) de la UDLA.

Las siguientes secciones introducen la motivación y marco teórico para este trabajo. Luego se presenta el método basado en un reconocimiento con redes neuronales aplicado a la verificación de la pronunciación correcta del Español de México. Este método se implementó en un prototipo con una interfaz programada en Java, la cual invoca las herramientas de procesamiento de señales, el reconocedor y el sistema de verificación. Todo ello explicado a continuación, junto con los resultados obtenidos, demostrando el potencial de esta herramienta.

2. CALL (Computer Assisted Language Learning)

La instrucción asistida por computadora (CAI) se refiere al area de estudio sobre la forma a través de la cual una computadora asiste a un usuario en el proceso de aprendizaje. En los últimos 40 años se ha visto un crecimiento exponencial en las aplicaciones de enseñanza asistida por computadora. Durante esta rápida evolución tecnológica CAI se ha vuelto mas refinado como resultado de sus transiciones paradigmáticas del pensamiento conductista al cognitivo al constructivismo. Las computadoras ahora funcionan como medios sofisticados para la instrucción.

La inteligencia artificial simbólica (IA) ha propuesto esquemas interesantes en el área de ICAI (Intelligent CAI), pero las redes neuronales no se han utilizado mas que en muy pocas ocasiones. En general, el conocimiento manejado en CAI es representado más bien de manera explícita, por lo que las redes neuronales no han encontrado aplicación [1].

CAI da a lugar al modo de enseñanza y aprendizaje por descubrimiento, en dónde el estudiante esta involucrado en una exploración intelectual del conocimiento de una manera mucho más libre y auto-dirigida. Es entonces una herramienta a través de la cual el aprendizaje puede ocurrir. En los últimos años más y más sistemas CAI han incorporado herramientas multimodales, interfaces que manejan audio, imágenes video, y programas interactivos que invocan a diversas aplicaciones. Recientemente, la intergación de interfaces de voz a los sistemas CALL o ICALL (Intelligent CALL) ha creado un mundo nuevo de posibilidades para el desarrollo de herramientas para la enseñanza [2].

3. Tecnología de Voz en la Educación

Diversos grupos de trabajo han desarrollado aplicaciones interesantes para el area de enseñanza de lenguajes. Por ejemplo, el software llamado Pronunciation Power [3] presenta al usuario con herramientas que le ayudan a aprender sobre la pronunciación de un lenguaje. Muestra una pantalla con la forma escrita de los sonidos, le permite al usuario grabar su voz y comparar las representaciones gráficas de ondas sonoras de su grabación con una versión "correcta" (véase la figura 1).

Figura 1: Pronunciation Power (TM) muestra dos ondas de sonido, la del locutor y la "correcta" para que el usuario decida si lo que dijo estuvo bien o no.

La aparente desventaja de este enfoque es que, cuando existe un error en la pronunciación de una palabra, el sistema no provee una retroalimentación explícita. El usuario requiere de práctica para detectar en base a ondas de sonido si su grabación corresponde correctamente a la correcta o no. Además no sabe cómo debe corregir ese error porque no conoce la causa de las diferencias que encuentra. Adicionalmente, no se sabe bajo qué criterio se escogió la representación "correcta", ya que la apariencia de esta puede variar al provenir de un locutor hombre o de una mujer, debido a las diferencias naturales entre las voces humanas.

Otro ejemplo interesante es de la compañía "Language Connect" [4], la cual utiliza el ViaVoice de IBM. El software "escucha" cada palabra o frase que dice el usuario, incluso enunciados complejos. Luego responde con una calificación por toda la palabra o frase y dice si una persona cuya lengua materna es la que se está practicando. Este reconocedor es muy potente y de buen desempeño, sin embargo, esta calificación que recibe el estudiante por lo que dijo no le indica en dónde se equivocó ni cómo corregir su error.

En otro campo de aplicación, el Center for Spoken Language Understanding (CSLU) del Oregon Graduate Institute y de la Universidad de Colorado han estado colaborando con la Tucker Maxon Oral School [5], en un esfuerzo conjunto para enfocado al entrenamiento del lenguaje para niños hipoacúsicos (desde sordera ligera hasta sordera profunda).

Ellos elaboraron un conjunto de herramientas que incorporan reconocimiento de voz y producción de voz, así como un agente conversacional animado, llamado Baldi [6; 7; 8]. Este agente esta representado por una cara animada en 3D que produce habla visual, es decir, produce expresiones faciales sincronizadas con el sintetizador [9]. Las expresiones faciales incluyen movimiento de labios, quijada, cejas, ojos y lengua, haciendo posible que un niño pueda leer los labios de Baldi y entender lo que esta diciendo. Baldi es también capaz de mostrar emociones de tristeza, alegría, molestia, etc. Los niños juegan con la interfaz de voz, en dónde se pueden diseñar las más variadas lecciones con imágenes, video, sonido y voz, reconocimiento y producción (véase la figura 2).

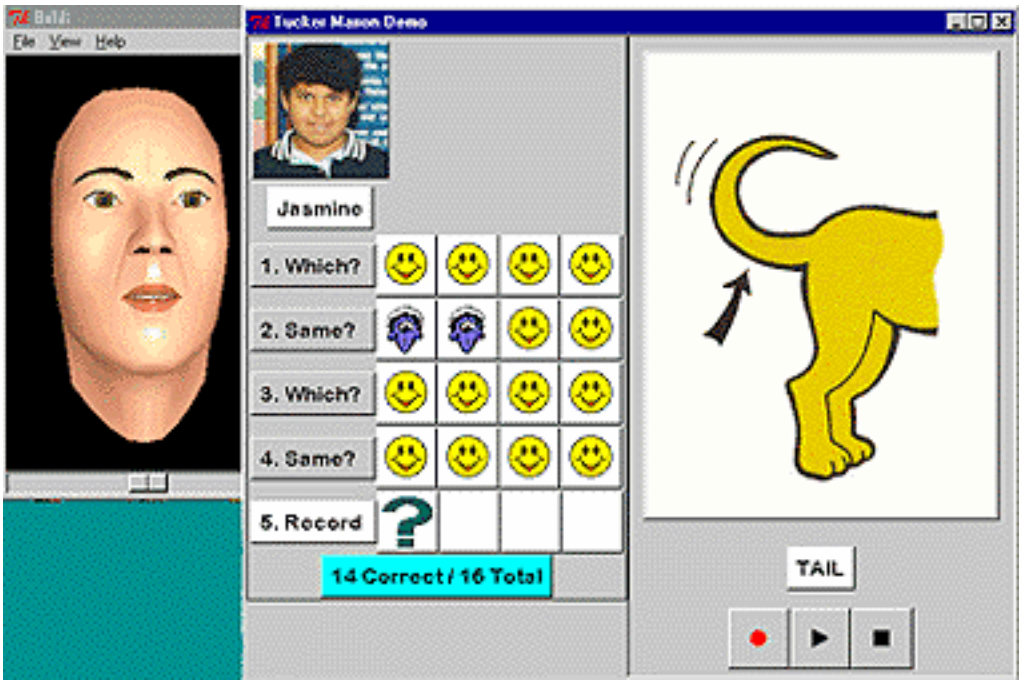

Figura 2: Un ejemplo de una sesión de trabajo con el CSLU Toolkit

Además de las aplicaciones mencionadas anteriormente existe una gran variedad de herramientas, programas y juegos que tienen la intención de ayudar en el aprendizaje de un lenguaje. La mayoría de estas herramientas son bastante impresionantes y cubren diversas necesidades de los usuarios.

TLATOA esta colaborando desde 1997 con el CSLU en la construcción de laversión en Español Mexicano del CSLU Toolkit[10]. En la Udla-P, así como en otras instituciones a través del país, los estudiantes extranjeros invierten el verano en cursos para aprender Español. Un gran porcentaje de estos estudiantes proviene de los E.E.U.U. El constante aumento en la demanda de cursos de Español nos hacen pensar en la necesidad de proveer ambientes de enseñanza adecuados. Por ello se inició un proyecto para el desarrollo de un ambiente para el aprendizaje del Español.

Como una parte de este proyecto se enfocaron dos trabajos de tesis en el desarrollo de una herramienta para verificar la pronunciación de palabras y/o frases y también para accesar un diccionario y consultar la pronunciación de palabras o frases por vía de la voz.

El aprender un lenguaje es ser capaz de hablarlo y ser entendido, por ello el entrenamiento de una pronunciación correcta es obviamente una actividad importante.

4. Verificación de la Pronunciación Correcta

El principal objetivo de este trabajo es proveer un medio para poder practicar un lenguaje hablado y poder detectar los errores de pronunciación, dando una retroalimentación explícita sobre el desempeño del locutor.

Para ello se diseñó una interfaz capaz de invocar las herramientas de reconocimiento y síntesis de voz del CSLU Toolkit. Estas herramientas son en

parte las que provee el mismo Toolkit en su versión estándar (reconocedor de propósito general y sintetizador de voz en Español Mexicano e Inglés generadas en TLATOA e integradas al sistema con excepción de los sistemas para el Inglés) y otras, como el reconocedor específico para la verificación de la pronunciación se crearon especialmente para este trabajo.

4.1 Entrenamiento del Reconocedor

Como primer paso para poder verificar la pronunciación, el sistema deberá poder reconocer la secuencia exacta de fonemas que haya dicho el usuario. Sin embargo, cuando se crea un reconocedor para un idioma en particular, este rechazará cualquier fonema que no pertenezca al idioma o tratará de buscar dentro de sus fonemas el que más se le parezca. Esto nos pone ante el problema de que debemos saber de antemano los sonidos que pudiera producir el locutor al tratar de pronunciar algo en Español. Debido a esto primero se limitó el proyecto a buscar verificar la pronunciación del Español hablado en México, cuando lo trata de hablar una persona de origen norteamericano. Entonces, se definieron los fonemas que provienen del inglés norteamericano que no pertenecen a los que se utilizan en el Español. Este paso es necesario para seleccionar los datos con los que se entrenará el reconocedor.
Una vez que se tienen suficientes elementos se entrena una red neuronal capaz de reconocer todos los fonemas del Español, más los del Inglés.

4.2 Vocabulario y Gramática

Un reconocedor requiere de que se defina un vocabulario y la gramática de las palabras que buscará reconocer. Para hacer que el reconocedor reconozca secuencias de palabras a pesar de estar mal pronunciadas y aceptar errores de pronunciación sin tratar de corregirlos fue necesario definir primero como vocabulario las letras del Español y sus posibles pronunciaciones, correctas e incorrectas, como se muestra en la tabla número 1.

Letras	Transcripciones fonéticas
a	{ a | ay | e | ey | uh }
b	{ bc b | B }
e	{ e | ey | i }
j	{ x | h }

Tabla 1: Ejemplo de las transcripciones de algunas letras (correctas e incorrectas desde el pnto de vista del Español)

La gramática es, por otro lado, una que permite cualquier secuencia de fonemas. Pero si se deja que el reconocedor trate de reconocer todos los sonidos que lleguen al micrófono se generarían demasiados errores de inserción por ruido ambiental, etc.
Para evitar este problema se utilizó una inspirada en el proceso para el etiquetado automático de voz, llamado forced alignment [11; 12; 13]. Esta técnica utiliza el grafo de pronunciaciones de una sola palabra o frase (la que se pretende reconocer) y utilizando el reconocedor busca mapear cada sonido con alguna de las posibles secuencias de fonemas para esa palabra o frase específica.
Por ejemplo, si la palabra a reconocer es "abeja" las posibles pronunciaciones se generan a partir de las pronunciaciones por cada letra (tabla 1):

ABEJA = {{ a | ay | e | ey | uh }{ bc b | B }{ e | ey | i }{ x | h }{ a | ay | e | ey | uh }};

Con esto se crea un grafo de pronunciaciones y la red neuronal

analizará la señal de entrada frame por frame tratando de encontrar por cada letra lo manera en que el usuario las pronunció [14].
Si el usuario pronunció mal la 'j' como 'h', entonces el reconocedor debe gererar como resultado la secuencia {a bc b e h a}. Esta salida se puede comparar con la versión correcta {a bc b e x a} e indicarle al usuario el punto en dónde se equivocó.

4.3 Interfaz

Para probar este sistema se diseñó un prototipo con una interfaz gráfica que permite al usuario grabar su voz, verificar la pronunciación y realizar otras consultas utilizando menús y la voz.

En la figura 3 se muestra como se ve la interfaz, en la cual existen las siguientes opciones:
El usuario puede seleccionar una frase o palabra que desea practicar a través de un menú de tópicos. Dentro de cada tópico existe una larga lista de palabras e incluso frases. Al escoger alguna, el sistema automáticamente le presenta una imagen (si esta disponible) y le da un ejemplo de pronunciación tocando un archivo .wav. El usuario también puede consultar el diccionario a través de la voz y se le responde utilizando el sintetizador en el idioma seleccionado (en este caso el de Español)[15]. La pantalla además presenta la traducción en forma escrita y la información que le indica cómo se ha de pronunciar una palabra.

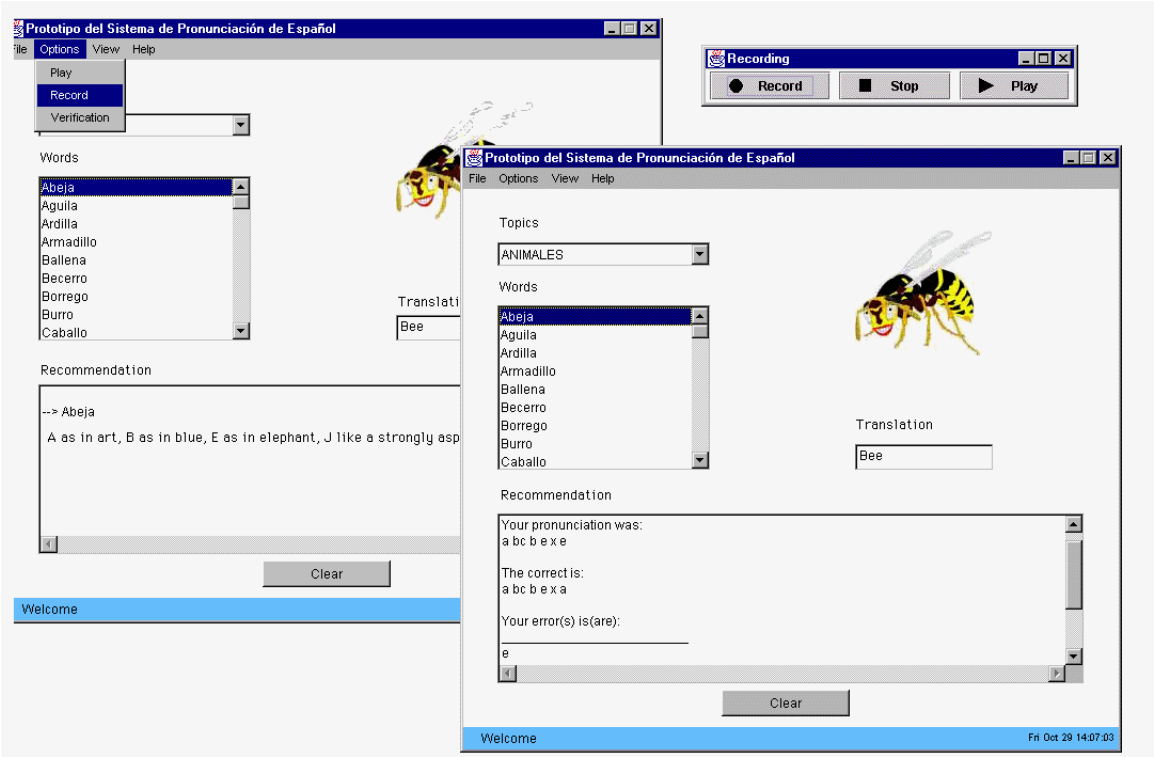

Figura 3: Interfaz del prototipo, que permite grabar una palabra o frase, escuchar un ejemplo de pronunciación, verificar lo grabado y obtener información de dónde hubo un error.

Cuando el usuario decide grabar una palabra y verificarla, el sistema le responde con lo que se reconoció y lo que debía haber dicho el locutor, marcando los errores.
Toda la parte de la interfaz fue programada en Jacl[16], el cual es Java con una interfaz que permite invocar rutinas escritas en Tcl. Esto fue necesario ya que las utilerías del CSLU Toolkit están todas escritas en C y Tcl.

5. Resultados y Conclusiones

Para determinar los ejemplos de sonidos que se debían incluir en el conjunto de entrenamiento de la red se estudiaon los errores más comunes que tienen estudiantes norteamericanos al pronunciar el Español. Estos son principalmente en las vocales y los sonidos que no existen en su idioma, como el de la 'r', 'y', y la 'll'.
En una base de datos se guardaron las palabras con las que puede practicar el estudiante, junto con su(s) pronunciación(es) correcta(s), así como las recomendaciones para su pronunciación.
El sistema entonces accesa esta información para obtener la pronunciación correcta y la información relacionada.
En la siguiente tabla se muestra un ejemplo de la salida del verificador de pronunciación. En la primera columna esta la palabra que el usuario desea practicar, en la segunda su transcripción correcta y en la tercera columna lo que se obtuvo de la pronunciación del usuario.

Palabra	Transcripción Correcta	El Usuario dijo
ABEJA	a bc b e x a	a bc b **ey** x a
CUATRO	kc k w a tc t r o	kc k w a tc t r **ow**

Tabla 2: Un ejemplo de errores detectados.

Cuando el sistema obtiene la pronunciación dicha por el usuario y la compara con la versión correcta se puede encontrar con diferentes tipos de error. Tradicionalmente en los sistemas de reconocimiento de voz se ecuentran errores de inserción, eliminación o sustitución.
El error de inserción es cuando el usuario dice más de los que debía decir. El error de eliminación ocurre cuando el usuario omite alguna letra. Estos dos errores no puedes ocurrir en nuestro caso ya que se forza la pronunciación a un conjunto limitado de fonemas. El error de sustitución por otro lado es el que se encuentra en estos casos.
El sistema trata de mapear lo reconocido con la pronunciación correcta y cuando no coinciden almacena el carácter equivocado (lo marca) y siguie comparando.
Para probar el desempeño de este sistema se invitaron a 5 estudiantes extranjeros a grabar 12secuencias de palabras pertenecientes a diversos tópicos (alimentos, animales, etc.). Primero estas secuencias fueron evaluadas por evaluadores humanos y luego por el sistema. Los evaluadores notaron 86 errores (un 6.72%) mientras que el sistema sólo marcó 62 errores . No todos los erores marcados por el sistema correspondían a los que marcaron los evaluadores, un 98.28% de los casos correspondían a las evaluaciones hechas por humanos. Un 5.16 % de fonemas mal pronunciados no fueron detectados por el sistema.
De esto se concluye que la técnica utilizada para verificar la pronunciación es prometedora pero requiere de más pruebas. En el proceso de prueba se encontró que algunos fonemas y contextos del inglés no estaban contemplados en el conjunto de

entrenamiento, por lo que el sistema fue incapaz de reconocerlos.

Los fonemas particularmente difíciles de distinguir para el sistema fueron la /x/, la pronunciación de la 'j', y /dZ/, que es la pronunciación de la 'll'.

El diccionario hablado, por otro lado gozó de gran aceptación y se reportó un reconocimiento del 92% de reconocimiento de las entradas hechas por los cinco usuarios de prueba y una respuesta siempre correcta de las traducciones solicitadas.

El sistema de verificación de pronunciación para el Español Mexicano permite detectar exactamente en cual letra el usuario hace un error. Esto le permite enfoca sus estudios a los puntos débiles de manera exacta para progresar en la forma hablada del lenguaje.

Este trabajo es una primera etapa de un sistema que pretende proveer un ambiente de aprendizaje del Español utilizando tecnología de voz.

Bibliografía

[1] Ayala, G., Comunicación personal, 1999.

[2] S. Bull, "Student Modelling for Second Language Acquisition", Computers and Education, 1994.

[3] Pronunciation Power, http:// www.englishlearning.com/.

[4] Language Connect, http:// shop. languageconnect.com

[5] Tucker Maxon Oral School, http://www.oraldeafed.org/schools/tmos/index.html.

[6] Center for Spoken Language Understanding , http:// cslu.cse.ogi.edu /tm/.

[7] R.Cole, T.Carmell, P.Connors, M.Macon, J.Wouters, J.de Villiers, A.Tarachow, D.Massaro, M.Cohen, J.Beskow, J.Yang, U.Meier, A.Waibel, P.Stone, G.Fortier, A.Davis, C.Soland, "Intelligent Animated Agents for Interactive Language Training", STILL:ESCA Workshop on Speech Technology in Language Learning, Estocolmo Suecia, mayo 1998.

[8] R.Cole, D.Massaro, J.de Villiers, B. Rundle, K. Shobaki, J.Wouters, M.Cohen, J.Beskow, P.Stone, P.Connors, A.Tarachow, D. Solcher, New Tools for Iinteractive speech and language training: Using animated conversational agents in the classroom of profoundly deaf children, Proc.ESCA-Matisse ESCA/Socrates Workshop on Method & Tool Innovation for Speech Science Education, London, UK, abril 1999.

[9] A. Black, P.Taylor, Festival Speech Synthesis System: System documentation (1.1.1.), Human Communication Research Centre Technical Report HCRC/TR-83, Edinburgh, 1997.

[10] B.Serridge, R.Cole, A.Barbosa, N.Munive, A.Vargas, Creating a Mexican Spanish version of the CSLU Toolkit, Proc. of the International Conference on Spoken Language Processing, Sydney, Australia, noviembre, 1998.

[11] S. Young, J. Odell, D.Ollason, V.Valtchev, P.Woodland, HTK Book, 1997. http://ceres.ugr.es /HTKBook/ HTKBook.html

[12] A. Olivier, I. Kirschning, Evaluación de métodos de determinación automática de una transcripción fonética", Memorias del II. Encuentro Nacional de Computación 1999, ENC'99, Septiembre 1999, Pachuca, Hidalgo, México.

[13] J.P. Hosom, Thesis Proposal: Accurate Determination of Phonetic Boundaries Using Acoustic-Phonetic Infomation in Forced Alignment,

CSLU, Oregon Graduate Institute, Agosto 1998.
[14] N. Aguas, Verificación de pronunciación para un ambiente de aprendizaje basado en tecnología de reconocimiento de voz, Tesis de licenciatura, Depto. de Ing. en Sistemas Computacionales, Universidad de las Américas, Puebla, Diciembre 1999.
[15] A. Ahuactzin, Desarrollo de un Diccionario Inglés-Español utilizando Tecnologías de Voz, Tesis de licenciatura, Depto. de Ing. en Sistemas Computacionales, Universidad de las Américas, Puebla, Diciembre 1999.
[16] http://scriptics.com/products