DESARROLLO DE UNA APLICACIÓN MÓVIL DE AYUDA A LA MEJORA DE LA PRONUNCIACIÓN DEL ESPAÑOL COMO LENGUA EXTRANJERA BASADO EN RECONOCIMIENTO DE VOZ

David Escudero-Mancebo

Valentín Cardeñoso-Payo

Universidad de Valladolid.

RESUMEN:

Esta comunicación presenta el desarrollo de un *App* para plataformas *Android* para la mejora de la pronunciación del español. El prototipo se basa en la metáfora del *parroting* e intenta corregir los errores de pronunciación con ejercicios donde los usuarios escuchan las frases de entrenamiento y deben repetirlas con la mayor corrección posible. Un sistema de reconocimiento de voz recibe las locuciones y las interpreta. En función de la respuesta del reconocedor, el sistema puede valorar su corrección fonética. En la comunicación se presenta una discusión sobre las capacidades de este tipo de sistemas para identificar errores típicos de determinados tipos de locutor. Se constata la necesidad de definir rutinas pedagógicas para explotar el sistema en entornos docentes prácticos.

1. Introducción

El aprendizaje de idiomas se beneficia del uso de las tecnologías de la información y de las comunicaciones desde hace ya varias décadas. En los ochenta era común que los cursos de idiomas distribuidos por las editoriales vinieran con su correspondiente cinta de audio en la que podían encontrarse ejercicios para escuchar diferentes diálogos y practicar la pronunciación. En los noventa, con la aparición de los CDs multimedia, la tecnología aporta una nueva dimensión al incluir ejercicios dinámicos que conjugan el audio con imágenes y vídeo. A finales de la década y comienzos del nuevo siglo, la multimedia entra en Internet merced al incremento del ancho de banda en la red. El abaratamiento de las TICs y la democratización del acceso a los servicios ha abierto un extenso abanico de posibilidades de las que se benefician también el aprendizaje de idiomas y las industrias de la lengua. Internet es hoy en día el principal medio de acceso y distribución de información[1]. Existen más de 2400 millones de usuarios de Internet a escala planetaria y el crecimiento anual es del 500%. El uso de *tablets* y *smartphones* ha demostrado una capacidad de crecimiento extraordinario. En 2011 apenas el 20% de usuarios de telefonía móvil usaba *smartphone* o *tablet*. Hoy en día, la proporción es más cercana al 100%[2]. Dentro del mercado de *smarphones*, se estima que *Android* seguirá siendo la tecnología predominante del mercado en 2015[3]. En este contexto, el español es el segundo idioma de comunicación internacional y el tercer idioma más

[1] Internet World Stats (http://www.internetworldstats.com/)
[2] RBC Capital Market Table and Smartphone Users vs. Other Markets (The potential for growth in huge) (http://rbc.com)
[3] Previsión del reparto de ventas según plataformas (http://www.journal3g.com/Dispositivos-moviles.htm)

estudiado como lengua extranjera. Parece pues que el desarrollo de herramientas para el aprendizaje del español que operen sobre tecnologías móviles *Android* como la que se defiende en este trabajo es una opción estratégica.

CALL (*Computer-Assisted Language Learning*) es el término que se emplea para referirnos al uso de los ordenadores como herramientas de apoyo en el aprendizaje de idiomas. Los servicios ofrecidos por el Instituto Cervantes en su aula virtual[4] o las actividades que muestra la Fundación para la Lengua Española[5] incluyen recursos CALL típicos como son ejercicios de vocabulario, ejercicios de gramática, lecturas, etc. Frente a las herramientas CALL convencionales surge más recientemente el concepto de herramientas CAPT (*Computer-Assisted Pronuntiation Trainning*). Mientras las herramientas CALL se orientan a potenciar la capacidad de lectura y/o escritura, así como la capacidad de comprensión oral por medio de vídeos y grabaciones, las herramientas CAPT se centran esencialmente en mejorar un aspecto fundamental de la capacidad comunicativa real de la expresión oral: la correcta pronunciación del idioma que se está aprendiendo (Eshani y otros 1997, Neri y otros 2002) y para la evaluación de la misma (Esquenazi, M. 2009; Zechner, 2007). La importancia radical del entrenamiento de la pronunciación fue defendida en (Bernstein y otros, 1990) al definir las capacidades comunicativas con la siguiente ecuación: $comm=pron*lex*(1+syn+rhet+prag+soc)$ donde *comm* con las capacidades comunicativas, *pron* es la pronunciación, *lex* es el control del vocabulario, *syn* es la sintaxis, *rhet* es la forma retórica, *prag* es la pragmática y *soc* es la sociolingüística. La conclusión es que las competencias de pronunciación afecta absolutamente a las capacidades comunicativas. Puede que una pronunciación como la de los nativos no sea necesaria, pero una pronunciación aceptable (comprensible) es deseable para que la comunicación tenga éxito.

En el aprendizaje del español como lengua extranjera (ELE) o en el aprendizaje de cualquier otra segunda lengua, la competencia fonológica constituye una competencia lingüística esencial para conseguir que el usuario llegue a alcanzar un nivel aceptable de competencia comunicativa. En efecto, para conseguir que los usuarios no nativos desarrollen una capacidad suficiente para comunicarse con hablantes nativos, la pronunciación se nos muestra como un aspecto esencial que hay que entrenar por cuanto de ella depende hacerse inteligible, y así se destaca en el "Marco común europeo para las lenguas: aprendizaje, enseñanza y evaluación" y en "Plan curricular del Instituto Cervantes". El aprendizaje del español, como el de la mayoría de las lenguas, requiere conocer y articular correctamente los sonidos que constituyen su sistema y utilizar con eficacia los elementos segmentales.

Siempre han existido ejercicios que animaban al usuario a escuchar grabaciones e intentar repetirlas. La innovación importante que estamos aportando en los últimos tiempos es el uso de las tecnologías del habla para ayudar al entrenamiento de la pronunciación. Esto supone que el ordenador pueda analizar la voz del alumno para valorarla y aconsejar al estudiante de cara a su corrección y mejora. Existen trabajos que ofrecen realimentación visual de cómo mejorar la articulación al hablar en el idioma que se está aprendiendo (Badim y

[4] Aula virtual del Instituto Cervantes (http://cvc.cervantes.es/).
[5] Fundación de la Lengua Española (http://www.fundacionlengua.com).

otros, 2010). Sin embargo, la tecnología del habla aplicada hoy en día, puede actuar tanto en los aspectos segmentales (mejora de la pronunciación de los fonemas) como en la prosodia (tonos, acento, entonación y patrones rítmicos).

En este artículo presentamos el desarrollo de una herramienta de apoyo a la mejora de la pronunciación del español. En la sección 2 se hace una revisión del estado del arte que permite ubicar nuestro producto en el contexto tecnológico actual. En el apartado 3 se presentan los componentes del sistema: módulo de definición de actividades, módulo de reconocimiento de voz, módulo de síntesis y módulo social. En la discusión se defienden las aportaciones presentes y futuras de nuestro sistema. Finalizamos el artículo con las conclusiones.

2. Estado del arte

La tecnología predominante para mejorar los aspectos segmentales de la pronunciación se apoya en los sistemas de reconocimiento de voz. Un módulo de análisis acústico extrae las propiedades que son potencialmente útiles de cara a la diagnosis. Estas propiedades se suponen relacionadas con la articulación y presentan la dificultad de su estimación robusta en entornos reales. DFT, *Mel-band filter bank*, IDFT son las propiedades que se emplean porque han demostrado servir para distinguir entre los distintos tipos de fonemas (Rabiner y Juang, 1992). En sistemas CAPT las principales tareas son el alineamiento segmental del habla, la detección de errores de pronunciación y la valoración (Tsubota y otros, 2004; Esquenazi y otros, 2007). Los principales desafíos son el modelado de los hablantes no nativos y el manejo del habla que contiene errores (Wang y otros, 2009). Para acotar el problema se acotan también las frases que el usuario puede decir, simplificando así las tareas de reconocimiento. El hecho de acotar el conjunto de frases que el usuario pueda decir limita mucho el uso de proyectos pedagógicos.

A pesar de heredar buena parte de la tecnología, las herramientas CAPT difieren de los sistemas ASR (Automatic Speech Recognition). En un sistema ASR, para un X dado (X: señal de entrada), hay que encontrar el W (W: lista de palabras) que maximiza la probabilidad condicionada p(W|X). Se resuelve maximizando p(W) ∗ p(X|W) y se entrena un modelo para cada fonema p(x|w). En CAPT por contra, W viene dada y X no es fiable. Se resuelve ahora el problema aplicando una segmentación con un alineamiento forzado, pero debe incluirse un módulo de detección de errores dado que puede aparecer un W’ en lugar de un W de manera que p(X|W′) > p(X|W). La valoración de la pronunciación se hace calculando p(X|W), pero el problema radica entonces en conocer cómo entrenar el modelo para locutores no nativos. El uso de corpus y de técnicas de aprendizaje automático servirá para obtener una representación automática es una tendencia como se evidencia en las publicaciones (Meng y otros, 2009).

Con respecto al uso de síntesis de voz, encontramos excelentes revisiones en (Black 2007 y Handley 2005). Existen tres modelos en el estado del arte: como modelo de pronunciación para entrenar la pronunciación requiriendo que la pronunciación sea extremadamente buena; como máquina de leer, usando el sistema para la

Figura 1: Interfaz del sistema.

práctica del dictado o sistemas de *shadowing*, requiriendo que la pronunciación sea buena; o como interlocutor en un sistema de diálogo donde los requisitos de calidad no son tan altos (Luo y otros, 2009). La síntesis basada en HMM apoyada en el *vocoder STRAIGHT* (Kawahara, 2006) ofrece síntesis de alta calidad integrando modelos prosódicos específicos. En Kubo (1998) se experimenta con el uso de síntesis *STRAIGHT* para trabajar la mejora de la distinción entre /l/ y /r/. Otra técnica popular es el uso de *morphing*. Las pronunciaciones propias son transformadas como si las dijera un nativo y viceversa (Kato y otros 2001). Por otro lado, también la técnica de ofrecer realimentación con tu propia voz, parece ofrecer buenos resultados (Hirose y otros 2003).

3. Componentes del sistema

La figura 1 muestra el interfaz básico del sistema. Se presentan cuatro opciones para introducir nuevos ejercicios, elegir alguno de los disponibles y ver resultados. Una vez elegido un ejercicio, la frase correspondiente puede oírse e intentar también pronunciar. A continuación explicamos cada uno de los módulos que componen el sistema.

3.1 Formulación de actividades

El usuario puede elegir entre un conjunto de actividades predefinidas o incluir nuevas actividades en función de sus necesidades. Una actividad nueva se configura introduciendo el texto con el que después, el alumno practicará su pronunciación.

El programa permite añadir en formato libre los textos que el usuario considere oportuno. Un uso responsable de la herramienta supone que sea el profesor o tutor quien sugiera los textos en función de las limitaciones observadas en sus estudiantes. El alumno podrá introducir o elegir las frases en las que observe mayor dificultad.

Para gestionar los usuarios y las actividades que tienen programadas empleamos el sistema gestor de base de datos *SQLite* disponible en el kit de desarrollo de *Android*. La figura 1 muestra el interfaz de introducción de actividades. Una vez introducidos los textos, el usuario está en condiciones de escuchar las frases y repetirlas bajo demanda.

3.2 Módulo de síntesis

El usuario puede necesitar escuchar la frase seleccionada antes de comenzar con la prueba. Para esto el sistema ofrece la posibilidad de usar un sintetizador de voz. El sintetizador de voz leerá la frase para que el usuario tenga un modelo de referencia.

En el momento en el que el usuario elige esta opción, el software se conecta con la máquina de síntesis para reproducir la frase seleccionada. La aplicación busca en el terminal los sintetizadores disponibles y deja que el usuario elija uno de ellos. Al menos se ofrece el sistema de síntesis de *Google*, aunque otros dispositivos móviles tienen sus propios sintetizadores. La calidad puede variar considerablemente entre los distintos servicios del sistema por lo que consideramos importante dar esta opción al usuario.

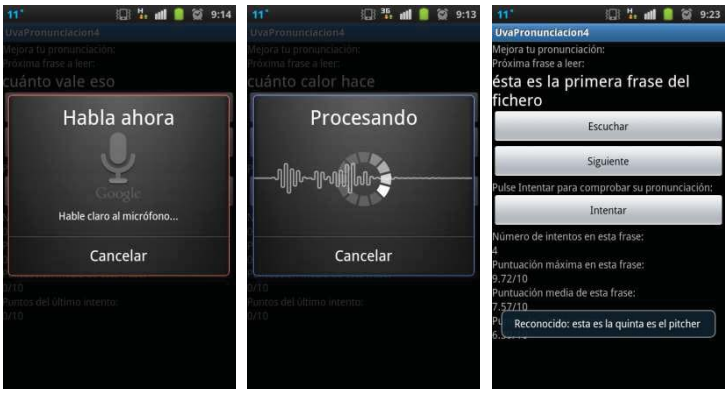

Figura 2: Interfaz del sistema en modo reconocimiento de voz

Utilizamos la tecnología *Android* disponible en la librería *android.tts*, en particular la clase *TextToSpeech*. Se crea un objeto *locale* donde se especifica el idioma y el país. El método de síntesis se llama *speak*. Este método recibe un texto, que puede estar enriquecido con información prosódica tipo *SABLE*. Cuando las muestras de audio están disponibles, el método las lanza al altavoz del dispositivo.

La calidad de la voz sintética puede no ser muy buena para determinadas frases. En estos casos puede pensarse que el beneficio de esta opción es dudoso. Para nosotros la importancia de disponer de este recurso no es despreciable porque desambigua algunas opciones como son la de posicionar el acento adecuadamente.

La alternativa a la síntesis de voz hubiera sido la grabación de la locución por parte de un actor o de un locutor profesional. Sin embargo, esta opción no es posible debido a que la definición de actividades, y por lo tanto de los textos a sintetizar, es realizada de forma dinámica por el propio usuario en función de sus necesidades. Al no disponer de un repertorio estático de locuciones, no es posible tener una grabación previa de las mismas.

3.3 Módulo de reconocimiento automático del habla

Una vez elegido el texto a reproducir (una palabra, una expresión o una frase o frases), comienza la fase de entrenamiento y prueba. El usuario deberá pronunciar el texto de la forma más natural posible. Como resultado, el sistema le asignará una puntuación que dependerá de la calidad de su pronunciación.

Una ventana avisa al alumno de que puede comenzar a emitir su locución (ver figura 2). Las herramientas de *Google Voice* se apropian del micrófono del dispositivo. Desde este momento, la voz del usuario es digitalizada y enviada al servidor central de *Google* que devuelve el texto reconocido. La aplicación toma el texto devuelto por *Google Voice* y computa una distancia entre el resultado devuelto y el resultado esperado. El resultado esperado depende exclusivamente del ejercicio elegido por el alumno. El resultado devuelto, de la calidad de su pronunciación.

Pongamos por ejemplo la siguiente frase: “El perro de Juan es negro”. Un alumno de nacionalidad china, obtuvo el siguiente resultado: “El pelo de Juan es nuevo”. En otra ocasión, la locución fue tan poco fluida que el reconocedor sólo reconoció la palabra “Juan”. A más distancia entre el resultado devuelto por *Google Voice* y la frase original peor calificación. En este caso, el primer alumno obtuvo un 6/10 y el segundo un 2/10.

La tecnología empleada es *Google Voice Search*. Se trata de un reconocedor de voz muy potente que se realimenta con las locuciones de las que dispone esta empresa comercial. Actualmente está disponible para más de cien idiomas, entre ellos el español Empleamos la clase *RecognitionIntent* de *android.speech.* Puede limitarse la gramática de búsqueda, pero nosotros empleamos el modo FREE_FORM para permitir cualquier locución. El resultado de *Google Voice* es una cadena de caracteres que incluye posibles alternativas.

También utilizamos las alternativas para valorar la distancia entre las frases predichas y las frases esperadas. A mayor número de opciones menor es la certidumbre del sistema de reconocimiento de voz y menor será la calificación. Esta incertidumbre está motivada principalmente por las deficiencias en la pronunciación y por lo tanto penalizan en la calificación.

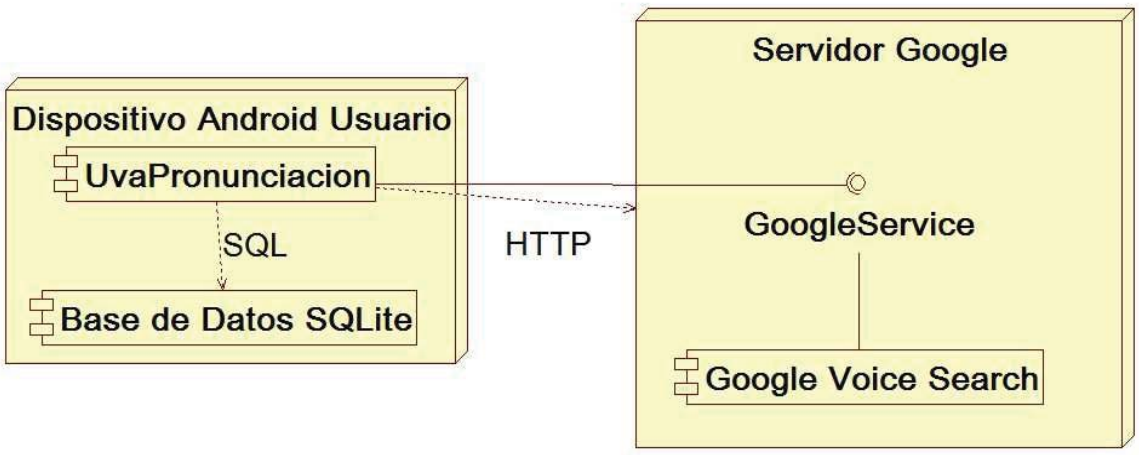

Figura 3: Esquema de la arquitectura del sistema.

3.4 Extensión social

En el esquema presentado en la figura 3, la base de datos de usuarios y ejercicios está en el propio terminal. Los datos relativos al usuario que realiza los ejercicios y los ejercicios en sí, se almacenan de forma local, de manera que sólo pueden ser accedidos desde el propio dispositivo en el que han sido introducidos. Una evolución de la herramienta utiliza una base de datos externa al dispositivo móvil.

La ventaja de utilizar una base de datos remota es que dicha base de datos puede ser compartida por diferentes usuarios y accesible desde distintos terminales. De esta manera pueden surgir nuevas oportunidades de cara al aprendizaje de idiomas, una de ellas, la versión social de la herramienta.

La versión social de la aplicación es un "juego serio" donde más de un usuario puede realizar simultáneamente los mismos ejercicios. En términos prácticos, varios alumnos pueden competir a ver quién de ellos pronuncia con mayor precisión la misma frase. En el escenario social, se ofrece un servicio en el que el usuario puede conocer a otros usuarios conectados al sistema. Los usuarios participan en competiciones en las que el desafío es pronunciar mejor una frase dada. Los beneficios tienen que ver con el incremento de la motivación personal de cada usuario para mejorar. El resultado esperado es un incremento del uso del sistema con la consiguiente mejora de la pronunciación.

Además, el poder acceder a una base de datos común desde distintos terminales da la posibilidad a los profesores de encargar de forma remota ejercicios a sus alumnos utilizando el sistema. Supone el paso de una herramienta de autoaprendizaje a una herramienta que puede permitir al profesor asistir el autoaprendizaje de sus alumnos.

4. Discusión

4.1 Aportaciones pedagógicas

La definición de actividades puede personalizarse y adaptarse a las necesidades de los usuarios. Estamos utilizando un sistema de síntesis y reconocimiento general, no limitado a un número de frases prefijado. Esto aporta una flexibilidad que permite incorporar nuevas actividades y nuevos usuarios bajo demanda. En términos prácticos, esto se proyecta en el hecho de que antes un alumno con dificultades para pronunciar determinado fonema, por ejemplo /r/ o /x/. Un profesor puede confeccionar un conjunto de actividades que trabajen este particular.

Tanto el alumno como el profesor pueden ver la progresión del aprendizaje gracias a la evaluación suministrada por el sistema. De especial interés es la capacidad del sistema de cara al autoaprendizaje del alumno. El propio estudiante puede plantear ejercicios. En este contexto, es el propio estudiante el que puede trabajar determinado aspecto de su interés.

También puede diagnosticarse la capacidad de pronunciar los diferentes sonidos por parte del alumno. Pueden programarse ejercicios que den cobertura fonética y contrastar las calificaciones obtenidas. La posibilidad de definir tus propios ejercicios, en combinación con el hecho de que el sistema opere en plataformas móviles, convierte a la herramienta en un sistema de gran interés para el autoaprendizaje.

Autoaprendizaje no significa aislamiento ni independencia del profesor. La extensión social del sistema permite al alumno estar en contacto permanente con su tutor y con otros alumnos. Esto es capital para incentivar su uso, incrementando la motivación del alumno con la definición de pruebas que suponen un desafío a afrontar.

4.2 Evaluación

Una etapa fundamental del desarrollo de todo producto software es la etapa de pruebas y evaluación. En un enfoque clásico de pruebas, se somete al programa a una batería exhaustiva de pruebas que pretende reflejar la casuística que puede darse durante la interacción de los usuarios. En la documentación del proyecto se recogen los tests de caja blanca y de caja negra realizados.

Más importante en este caso es la evaluación de la respuesta de los usuarios ante el sistema. Hemos probado la usabilidad del sistema teniendo en cuenta la facilidad de uso del producto cuando se enfrenta a un usuario nuevo. Sin embargo falta por probar, y asumimos como una debilidad del trabajo, la eficacia desde el punto de vista pedagógico del sistema.

En una evaluación de la eficacia del producto, debemos atender a cuestiones tales como la mejora de la pronunciación de un usuario que se somete a los ejercicios programados. Además también será necesario evaluar las tareas pedagógicas que puedan encomendarse. Ante un resultado negativo (nula mejora de la pronunciación

del alumno) cabe preguntarse si nuevas rutinas de uso del sistema pueden contribuir en la consecución de la deseada mejora. También hemos elaborado una versión social ¿hasta qué punto es eficaz en contraste con el uso aislado del sistemas por parte del alumno?

Confiamos en que la colaboración con profesores de ELE nos permita profundizar en la obtención de resultados que resuelvan los interrogantes planteados en el párrafo anterior para plantear nuevas publicaciones en el futuro.

4.3 Aportaciones tecnológicas

Desde un punto de vista tecnológico defendemos un producto que opera en un entorno como es *Android*, soportado por una gran cantidad de plataformas móviles (teléfonos y tabletas), y que está ampliamente extendido. Los servicios *Google Voice* son gratuitos y mejoran día a día por lo que la operatividad del servicio es no sólo extensa sino eficaz tal y como ha sido defendida.

El sistema presentado es completamente extensible a otros idiomas. Aunque nuestro objetivo principal era el de desarrollar una herramienta para la mejora de la pronunciación del español, la extensión a otras lenguas es inmediata, dependiendo exclusivamente de la disponibilidad de tecnología *Google Voice* para dicho idioma. Actualmente la compañía estadounidense ofrece servicios en más de cien idiomas y sus expectativas y recursos les permiten pensar en un crecimiento ilimitado.

5. Conclusiones y trabajo futuro

En este trabajo se ha presentado el estado del arte en el uso de tecnologías del habla en el desarrollo de herramientas CAPT de mejora de la pronunciación. Se ha documentado cómo desarrollar una herramienta CAPT para dispositivos móviles tipo *Android* basada en las herramientas *Google Voice*. La herramienta permite a un usuario practicar su pronunciación a la vez que obtiene una puntuación sobre la calidad de la misma.

En la actualidad se está trabajando en la extensión de la herramienta en dos vías diferentes. Primero se quiere mejorar el interfaz y la portabilidad empleando tecnología AIR de *Adobe*. Por otro lado se quiere hacer la invocación a los servicios *Google Voice* desde un *script* alojado en un servidor externo, en lugar de hacerlo desde el propio terminal. Esta opción abre la posibilidad de almacenar el audio y realizar un procesamiento en paralelo del mismo para así poder aplicar filtros sobre la señal de voz, que permiten añadir funcionalidad a la herramienta. Nuestro objetivo principal es visualizar la información prosódica o de segmentación fonética, mejorando la realimentación que reciba el usuario. Confiamos poder aplicar la experiencia del grupo en etiquetado y procesamiento prosódico (Gonzalez Ferreras y otros, 2012; Escudero y otros, 2002) en sistemas para la mejora del acento en estudiantes de español como lengua extranjera[6].

[6] Agradecimientos: Diego Vallejo y David Soler por sus aportaciones en la programación de las primeras versiones de la herramienta. Mª Pilar Celsa porque el apoyo que nos está dando en el desarrollo de estas investigaciones es un estímulo impagable. El proyecto del Ministerio Glissando FFI2011-29559-C02-01

BIBLIOGRAFÍA

BADIM, P. y A. Ben Youssef, G. Bailly, F. Elisei, T. Hueber (2010), *Visual articulatory feedback for phonetic correction in second language learning*, Proceedings SLATE.

BERNSTEIN, J. y M. Cohen, H. Murveit, D. Rtischev, M. Weintraub (1990), *Automatic evaluation and training in English pronunciation*, First International Conference on Spoken Language Processing.

BLACK, A. W. (2007) Speech synthesis for educational technology, in: Proc. ISCA ITRW SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA.

EHSANI, F. y J. Bernstein, A. Najmi, O. Todic (1997), *Subarashii: Japanese interactive spoken language education*, Fifth European Conference on Speech Communication and Technology.

ESCUDERO, D. y C. González, V. Cardeñoso (2002) *Quantitative evaluation of relevant prosodic factors for text-to-speech synthesis in Spanish* Proceedings of ICSLP

ESKENAZI, M. (2009) *An overview of spoken language technology for education*, Speech Communication 51 (10) 832–844.

ESKENAZI, M. y A. Kennedy, C. Ketchum, R. Olszewski, G. Pelton (2007) *The native accent pronunciation tutor: measuring success in the real world*, Proceedings of SLaTE.

GONZÁLEZ-FERRERAS, C y D Escudero-Mancebo, C. Vivaracho Pascual, V. Cardeñoso-Payo (2012) *Improving automatic classification of prosodic events by pairwise coupling* Audio, Speech, and Language Processing, IEEE.

HANDLEY, Z. y M.-J. Hamel (2005), Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (call), Language Learning & Technology 9 (3) 99–120.

HIROSE, K. y F. Gendrin, N. Minematsu (2003), *A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice*, Proc. EUROSPEECH, Vol. 4, pp. 3149–3152.

KATO, S. y G. Short, N. Minematsu, C. Tsurutani, K. Hirose (2011), *Comparison of native and non-native evaluations of the naturalness of japanese words with prosody modified through voice morphing*, Proc. SLaTE.

KAWAHARA H. (2006), *Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,* Acoustical science and technology 27 (6) 349–353.

KUBO, R. /r/-/l/ *perception training using synthetic speech generated by straight algorithm*, in: Proc. Spring Meeting of Acoust. Soc. Japan, 1998, pp. 383–384.

LUO, D. y N. Minematsu, Y. Yamauchi, K. Hirose (2009), *Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences*, Proceedings of SLaTE.

MENG, H. y W.-K. Lo, A. M. Harrison, P. Lee, K.-H. Wong, W.-K. Leung, F. Meng (2009), *Development of automatic speech recognition and synthesis technologies to support Chinese learners of English: The cuhk experience*, Proc. of APSIPA.

NERI, A. y C. Cucchiarini, H. Strik, L. Boves (2002), The pedagogy-technology interface in computer assisted pronunciation training, Computer Assisted Language Learning 15 (5) 441–467.

RABINER, L. y B.-H. Juang, (2009) *Fundamentals of speech recognition.* Prentice Hall.

TSUBOTA,Y. y M. Dantsuji, T. Kawahara (2004), *An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors, ReCALL* 16 (01) 173–188.

WANG, H. y C. J. Waple, T. Kawahara (2009), *Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition*, Speech Communication 51 (10) 995– 1005.

ZECHNER, K. y D. Higgins, X. Xi (2007) *Speechrater: A construct-driven approach to scoring spontaneous non-native speech*, Proc. SLaTE.