DEFINICIÓN DE HERRAMIENTAS PARA LA MEJORA DE LA PRONUNCIACIÓN DEL ESPAÑOL COMO LENGUA EXTRANJERA: EL PROYECTO SAMPLE

DAVID ESCUDERO-MANCEBO, CÉSAR GONZÁLEZ-FERRERAS, VALENTÍN CARDEÑOSO-PAYO
Grupo de Investigación ECA-SIMM
Departamento de Informática. Universidad de Valladolid.

1.- INTRODUCCIÓN

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 o las actividades que muestra la Fundación para la Lengua Española 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 y la evaluación de la misma (Esquenazi, 2009).

En este artículo presentamos una revisión del estado del arte en el uso de tecnologías de habla para el desarrollo de herramientas CAPT (sección 2). A continuación presentamos un prototipo propio que se apoya en tecnología móvil Android (sección 3).

Los objetivos de nuestro grupo de investigación son más ambiciosos y buscamos el desarrollo de un sistema de diagnosis que incluya realimentación de los errores fonéticos y prosódicos que cometen los usuarios (sección 4). Finalizamos el artículo detallando las conclusiones y el trabajo futuro.

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 al 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, 2004). Los principales desafíos son el modelado de los hablantes no nativos y el manejo del habla que contiene errores (Wang, 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 esta tecnología en proyectos pedagógicos.

Con respecto al uso de síntesis de voz, encontramos excelentes revisiones (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 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, 2009). La síntesis basada en HMM apoyada en el vocoder STRAIGHT (Kawahara, 2006) ofrece síntesis de alta calidad integrando modelos prosódicos específicos. Otra técnica popular es el uso de morphing. Las pronunciaciones propias son transformadas como

si las dijera un nativo y viceversa (Kato, 2001). Por otro lado, también la técnica de ofrecer realimentación con tu propia voz, parece ofrecer buenos resultados (Hirose, 2003).

Pasando a las cuestiones relativas a la prosodia en la pronunciación de hablantes no nativos, primero decir que la prosodia hace referencia a cuestiones relativas a la entonación, acento y ritmo, medidos por variaciones de la frecuencia fundamental, la energía y la duración de los fonemas. La pronunciación de los hablantes no nativos suele contener varios tipos de desviaciones prosódicas que dependen de la nacionalidad del hablante y del idioma objetivo. Se han empleado diferentes métricas de la prosodia para estimar la calidad como las medidas de fluidez (Cucchiarini, 2002). Se han empleado propiedades prosódicas adicionales para estimar las competencias de los hablantes como el duration log-likelihood o el rate of speech o la combinación lineal de varios scores (Hirabayashi, 2010). También se ha valorado el acento léxico en términos de posición o modo (Minematsu, 2000). Otro aspecto que se ha valorado es el ritmo empleando diferentes métricas (Ramus, 2000).

De nuevo, se realiza la comparación de los patrones de entonación (+energía) entre alumnos y un modelo comparando entre palabras o usando como base diferentes tipos de unidades (Suzuki, 2008). Por último también se ha estudiado la realimentación correctiva con sistemas automáticos como los árboles de decisión o utilizando la propia voz del usuario (Hirose, 2003).

Con respecto a las bases de datos disponibles de locutores no nativos, en Raab, (2007) puede encontrarse una revisión del estado hasta la fecha. Se detallan 42 bases de datos en http://en.wikipedia.org/wiki/Non-native_speech_database de las cuales sólo dos contienen locuciones en español. Una de ellas es del proyecto TCSTAR que contiene grabaciones a locutores e intérpretes del parlamento europeo y otra es de uso militar.



Figura 1: Interfaz del sistema en modo reconocimiento de voz.

3.- PROTOTIPO BASADO EN RECONOCIMIENTO DE VOZ

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. El sistema se compone de cuatro módulos como son el módulo de definición de actividades, el módulo de síntesis de voz y el módulo de reconocimiento de voz. Aquí presentamos brevemente cada uno de estos módulos sin entrar en detalles técnicos. Estos detalles pueden ser encontrados en Escudero, (2013).

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.

El usuario puede necesitar escuchar la frase seleccionada antes de comenzar con la prueba. Un sintetizador de voz leerá la frase para que el usuario tenga un modelo de referencia. Al menos se ofrece el sistema de síntesis de Google, aunque otros dispositivos móviles tienen sus propios sintetizadores. 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.

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. 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. Este resultado es representativo de la calidad de su pronunciación.

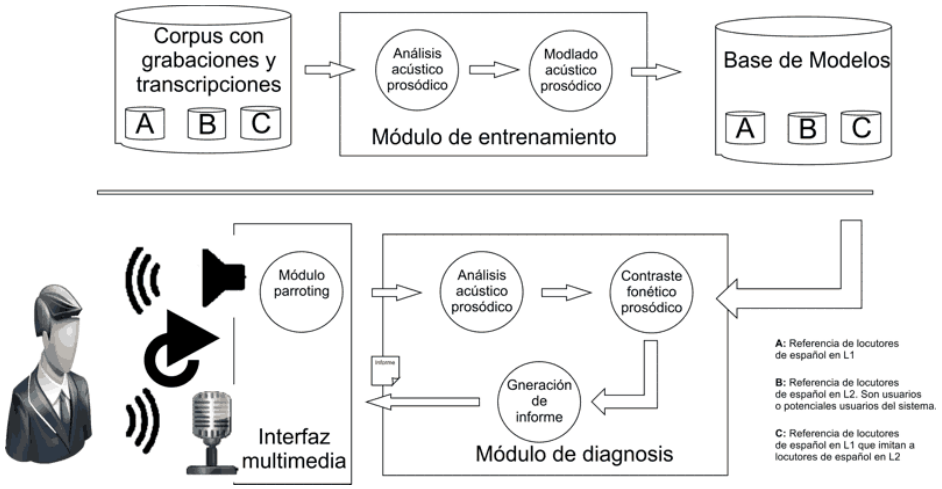


Figura 2: Esquema de sistema CAPT con módulo de diagnosis.

La tecnología empleada es Google Voice Search. Se trata de un reconocedor de voz que se realimenta con las locuciones de las que dispone esta empresa comercial. Puede dar alternativas de manera que a mayor número de alternativas, menor es la certidumbre del sistema y menor será la calificación. Esta incertidumbre está motivada principalmente por las deficiencias en la pronunciación y por lo tanto penalizan en la calificación.

4.- EL PROYECTO SAMPLE

4.1. Esquema del sistema CAPT

La figura resume el esquema operativo del proyecto SAMPLE de desarrollo de herramientas CAPT. El módulo de análisis acústico-prosódico extrae las propiedades más relevantes desde el punto de vista de la mejora de la pronunciación. La hipótesis que soporta la necesidad de este módulo es que, como hemos ya comprobado en nuestros trabajos de etiquetado prosódico (González Ferreras, 2012), la correcta selección de las propiedades prosódicas es determinante en el éxito de las etapas de modelado y de contraste posterior.

El módulo de modelado acústico prosódico recoge los aspectos más relevantes de la pronunciación de una serie de locutores de referencia. Permitirá crear una base de modelos de referencias de diferentes tipos de locutores. Los modelos deberán reflejar no sólo la variedad temporal de los contornos prosódicos y la caracterización acústica de los fonemas. También deberá correlacionar estos aspectos con las características lingüísticas de la entrada, principalmente de tipo léxico pero también sintáctico. También en trabajos previos se ha constatado la relevancia de estas variables en el modelado (Cardeñoso, 2004).

El módulo de contraste de la pronunciación entre diversas locuciones identificará las divergencias entre las locuciones del usuario con respecto a los modelos esperados. El contraste se hará teniendo en cuenta la evolución temporal de las características acústicas de la señal de entrada (la locución del usuario) y también las características lingüísticas del mensaje (léxico y sintaxis). La hipótesis básica es que si los modelos caracterizan correctamente a los locutores típicos, las distancias relevantes entre las locuciones y las referencias deberán ser identificadas como posibles problemas. La información léxica/sintáctica asociada al mensaje permitirá diagnosticar la divergencia como un error, previsiblemente repetible. La disponibilidad de modelos de locutores de L1, de L2 y de imitadores permitirá establecer si dicho error es un caso aislado o, por contra, se trata de un error sistemático y característico del locutor o tipo de locutor.

El interfaz para el uso del módulo de contraste se desarrollará para que facilite la interacción eficaz del usuario con el sistema. Aportará recomendaciones para la mejora constante de la pronunciación. También se adaptará al usuario en función de los problemas que manifiesta. La definición de un buen interfaz es fundamental para garantizar el

aprovechamiento del sistema por parte del usuario. La metáfora principal sobre la que nos vamos a apoyar es la del parroting. No obstante, el objetivo es que el interfaz ofrezca recomendaciones adaptadas al usuario para garantizar su mejor aprovechamiento.

Los modelos de referencia de locutores L1 y L2 se ampliarán con la elaboración de un corpus de imitadores. La hipótesis que justifica la necesidad de grabar este corpus es que la disponibilidad de un corpus de imitadores permitirá disponer de modelos de referencia que serán de gran utilidad para inferir los rasgos que son idiosincrásicos de los distintos tipos de locutores.

4.2. Corpus de hablantes no nativos

Para realizar los modelos de pronunciación de locutores españoles se dispone del corpus Glissando (Garrido 2013). Este corpus consta de una sección de noticias leídas y de otra de diálogos. Cuatro actores y cuatro locutores profesionales fueron grabados en condiciones de estudio profesional. El corpus está transcrito, segmentado y etiquetado prosódicamente de forma parcial.

Este corpus es la referencia para nuestro corpus de habla no nativa. Un total de 14 informantes de nacionalidades china, estadounidense y japonesa han leído una parte del corpus de noticias. También se hizo una selección de un conjunto de frases balanceadas fonéticamente. Una sección de habla espontánea y otra de lectura de la fábula de Esopo titulada *El viento y el sol* completan el corpus. A fecha de finalización de la escritura de este documento disponemos de 220 minutos de grabación. El corpus no está cerrado y se persigue incrementar su volumen con nuevos locutores de otras nacionalidades.

En la actualidad el corpus se está procesando por un sistema de reconocimiento automático del habla. Confiamos obtener con ello una lista de errores típicos y una tipificación automática de la calidad de los informantes. Una campaña de evaluación subjetiva de estas muestras de voz servirá para contrastar las métricas objetivas con los juicios de una serie de expertos que valorarán las locuciones de los hablantes no nativos.

5.- CONCLUSIONES Y TRABAJO FUTURO

Con respecto al prototipo de mejora de la pronunciación que utiliza el sistema de reconocimiento de voz, disponemos de una primera versión operativa que está siendo actualmente probada con usuarios reales. La evaluación de la misma en escenarios reales dará cuenta del potencial de la misma. En paralelo trabajamos en la definición de una 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.

Con respecto al esquema de trabajo para la evaluación diagnóstica de la pronunciación no nativa, se trata de un trabajo en vías de desarrollo a diversos niveles. En particular, se ha solicitado financiación a la Junta de Castilla y León para poder avanzar en estos asuntos y mientras tanto disponemos ya de propuestas de contraste de perfiles de F0 entre locutores. Contamos con poder cruzar dicha información con la relativa a etiquetado prosódico con prontitud. Disponemos de un corpus de informantes que está permitiendo identificar errores típicos de los locutores no nativos.

La correlación de la evaluación automática con la evaluación objetiva de las muestras de voz servirá para poner en valor los juicios del sistema automática que estamos desarrollando.

BIBLIOGRAFÍA

Cucchiarini C. et alii, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech". *The Journal of the Acoustical Society of America* 111. (2002)

Cardeñoso V. y D. Escudero. "A strategy to solve data scarcity problems in corpus based intonation modelling". *Proceedings of ICASSP* 1. (2004): 665-668.

Escudero, D. y V. Cardeñoso, "Desarrollo de una aplicación móvil de ayuda a la mejora de la pronunciación del español como lengua extranjera basado en reconocimiento de voz". Actas del III Congreso Internacional del Español, 2013.

Eskenazi, M. "An overview of spoken language technology for education". *Speech Communication* 51. 10 (2009): 832-844.

Garrido J. M. et alii, "Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan". *Language Resources and Evaluation* (2013): 1-27.

González Ferreras, **C** et alii. "Improving automatic classification of prosodic events by pairwise coupling". *Audio, Speech, and Language Processing, IEEE* (2012)

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

Hirabayashi K. y S. Nakagawa, "Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques". *Proc. Eurospeech* (2010): 598-601.

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

Kato, S. et alii, "Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing". *Proc. SLaTE* (2011)

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

Luo, D. et alii, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences". *Proceedings of SLaTE* (2009).

Minematsu N. y S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations". *Proc. Integration of Speech Technology into Learning* (2000): 130–137.

Raab M. y R. Gruhn y E. Noeth, "Non-native speech databases, in: Automatic Speech Recognition & Understanding". ASRU. IEEE Workshop on, IEEE, (2007): 413-418.

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

Ramus F. y M. Nespor y J. Mehler, "Correlates of linguistic rhythm in the speech signa". *Cognition* 75. 1 (2000).

Suzuki M. et alii. "Automatic evaluation system of English prosody based on word importance factor". *Journal of Systemics, CyberneTIC and InformaTIC* 6. 4 (2008): 83-90.

Tsubota,Y. y M. Dantsuji y T. Kawahara, "An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors". *ReCALL* 16. 1 (2004): 173-188.

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