Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 101-108 recibido 05-05-12 revisado 30-05-12 aceptado 05-06-12

Diseño y desarrollo de un sistema de evaluación automática de la pronunciación para el euskara

Design and development of an automatic pronunciation evaluation system for Basque

Igor Odriozola
UPV/EHU
Urkixo zum., z/g
igor@aholab.ehu.es

Oliver Jokisch
TU Dresden
Chair for Sys. Theory and Speech Tech.
Oliver.Jokisch@tu-dresden.de

Inma Hernáez
UPV/EHU
Urkixo zum., z/g
inma@aholab.ehu.es

Rüdiger Hoffmann
TU Dresden
Chair for Sys. Theory and Speech Tech.
Ruediger.Hoffmann@tu-dresden.de

Resumen: En este artículo, se presentan los primeros pasos en el desarrollo de un sistema de enseñanza de la pronunciación asistida por ordenador (CAPT, *Computer-Assisted Pronunciation Teaching*) para el euskara. El punto de partida es un sistema estándar de reconocimiento automático del habla (ASR) basado en modelos ocultos de Markov (HMM) que maneja parámetros de confianza GOP (*Goodness of Pronunciation*) para la verificación de fonemas. Dicho ASR se integrará en AzAR, el software de entrenamiento de la pronunciación desarrollado para el alemán y varias lenguas eslavas. En este artículo se presentan los primeros pasos del diseño del currículum para el euskara, los problemas generados en la verificación por el uso de HMMs creados a partir de una base de datos de ASR, y algunos resultados iniciales.
Palabras clave: evaluación de la pronunciación, verificación de fonemas, parámetros de confianza GOP, sistema AzAR.

Abstract: In this paper, the first steps of the development of a computer-assisted pronunciation teaching (CAPT) system for Basque are introduced. The baseline is a standard automatic speech recognition (ASR) system based on hidden Markov models (HMMs) that manages GOP (goodness of pronunciation) scores for phoneme verification. This ASR will be integrated into AzAR, the pronunciation training software developed for German and other Slavonic languages. This paper presents the initial steps in the design of the curriculum for Basque, some verification problems caused by the use of HMMs created from an ASR database, and some preliminary results.
Keywords: pronunciation evaluation, phoneme verification, GOP confidence scores, AzAR system.

1 Introducción

Para desarrollar un sistema de enseñanza de la pronunciación asistida por ordenador o CAPT (*Computer-Assisted Pronunciation Teaching*), es una tarea esencial adaptar los algoritmos utilizados para el reconocimiento automático del habla (ASR, *Automatic Speech Recognition*). En este artículo presentamos los primeros pasos del proceso de integración del euskara en el sistema AzAR (*Automat zur Akzentreduktion* – Autómata para la reducción del acento), el cual es el resultado de un proyecto desarrollado en el IAS (*Institut für Akustik und Sprachkommunikation*) de la Universidad Técnica de Dresden (Jokisch et al., 2005).

El sistema AzAR se diseñó inicialmente para entrenar y mejorar la pronunciación de estudiantes de la lengua alemana (L2) cuya lengua nativa (L1) fuera alguna de las pertenecientes al grupo de lenguas eslavas. Dentro del proyecto de cooperación *Euronounce* (Demenko et al., 2009), el sistema se amplió para el polaco, eslovaco, checo y ruso

ISSN 1135-5948

como L2. La base de datos *Euronounce* incluye lecciones especiales para entrenar la pronunciación de unos fonemas concretos, y se incluyen ciertas frases para utilizarlas como referencia en la práctica de la entonación o prosodia. De todas maneras, este último aspecto no tiene un *feedback* automático por parte del sistema. En siguientes desarrollos, el concepto *Euronounce* fue testado también con estudiantes de alemán de origen chino (Ding, Mixdorff y Jokisch, 2010).

El euskara es una lengua aislada que no pertenece al grupo de lenguas indoeuropeas, al contrario que cabría esperar si nos fijamos en su ubicación geográfica (Hualde, 1991). Convive con dos lenguas vecinas de gran potencial demográfico: el español y el francés. La influencia que ejercen esas dos lenguas sobre el euskara es importante, ya que el euskara se encuentra en situación de diglosia tanto en su parte norte (bajo la administración francesa) como en la sur (bajo la administración española), y tiene un estatus jurídico y grado de oficialidad heterogéneo, bastante complejo, dependiendo de la región.

Tal y como se ha mencionado anteriormente, las lenguas vecinas tienen gran influencia en el euskara, sobre todo para las personas que la estudian como L2. Además, *Euskaltzaindia*, la Real Academia de la Lengua Vasca, no ha tomado aún una decisión sobre cuál debe ser la entonación o prosodia para el euskara estándar, debido a la variedad de acentuaciones y entonaciones que muestra el euskara de un dialecto a otro. Eso hace que los estudiantes de euskara como L2 no tengan una referencia clara de la entonación que deben utilizar, lo que conlleva que muchas veces se decanten por la de su propio L1.

Por otro lado, a nivel fonético, el euskara posee ciertos fonemas que no existen en el inventario de las lenguas vecinas, y deben ser, por tanto, objeto de atención en el desarrollo del sistema de evaluación de la pronunciación de fonemas.

Por tanto, la primera conclusión clara que se ha obtenido es que en el diseño de un sistema CAPT para el euskara hay que tener en cuenta dos aspectos: el segmental o relativo a un sólo fonema (en este caso, la realización fonética o modo de pronunciación) y el suprasegmental o relativo a un grupo de fonemas (en este caso, el acento y la entonación).

2 Diseño del currículum para el euskara

Tal y como se ha explicado en el apartado anterior, se va a implementar una parte para la evaluación segmental y otra para la suprasegmental. Por tanto, el currículum de referencia para el euskara o colección de las características principales de un idioma que un hablante debe adquirir y que son objeto de entrenamiento debe comprender ejercicios para evaluar ambos aspectos.

Los primeros pasos para la evaluación de la parte suprasegmental, más concretamente la prosodia, fue implementada en AzAR bajo el proyecto *Euronounce*, pero el sistema se basaba solamente en la percepción auditiva del alumno, ya que el alumno debía comparar la entonación de una frase repetida con una de referencia. Por tanto, no había ninguna evaluación automática, es decir, ningún *feedback* por parte del sistema. Para el euskara, se ha considerado que la evaluación de la parte suprasegmental es esencial y, por tanto, se va a añadir un módulo de análisis de la prosodia que constará de una representación gráfica de las curvas de frecuencia fundamental (f0) y una puntuación obtenida de forma automática comparando la curva de f0 grabada por el alumno con la de referencia.

En esta parte se evaluarán dos características:

- La acentuación a nivel de palabra.
- La entonación a nivel de frase.

La acentuación y entonación que se han escogido como de referencia son las correspondientes al dialecto central del euskara, por ser aquélla la más estable en un área mayor. El acento característico a nivel de palabra de esta área, el cual no es exclusivo del dialecto central, sigue mayoritariamente un patrón acentual [+2, -1] para palabras aisladas de más de tres sílabas (Álvarez, 1986); es importante tener en cuenta que el euskara es una lengua aglutinante posposicional y que las marcas de los casos gramaticales se añaden al sintagma nominal en forma de sufijos; las palabras que llevan dichas marcas también siguen el mismo patrón.

Esta parte del currículum consta, en total, de 20 palabras aisladas y 50 frases.

En lo que se refiere a la parte segmental, las características más destacables del euskara, tanto fonéticas como fonológicas, son las siguientes:

- Características fonéticas:
  - El sistema vocálico y sistema de diptongos e hiatos.
  - Fonemas que no existen en L1 (español y francés), como, por ejemplo, el /*ts`*/ (ver ref. *Sampa Basque*).
  - Diferenciación entre las 6 sibilantes sordas del euskara: /*s`*/, /*s*/ y /*S*/ (fricativas) y /*ts`*/, /*ts*/ y /*tS*/ (africadas).

- Características fonológicas:
  - El proceso de palatalización de las consonantes /*l*/ y /*n*/ en el contexto /*i*CV/ (V es cualquier vocal).
  - El proceso de pérdida de sonoridad del primer fonema de la palabra siguiente a la partícula negativa *ez* (*ez dator* > /es`tatorr/) o desaparición de la sibilante de la partícula negativa (*ez nator* > /enatorr/).

Esta parte del currículum consta de 60 pares de palabras (contrastes) y 125 frases.

La voz de referencia fue grabada con un locutor nativo de euskara. Las señales se grabaron en formato digital de 16 bits a 16 kHz, en el estudio de grabación del IAS.

3 Pruebas iniciales y adaptación

3.1 El ASR base

Para la parte segmental, se ha utilizado el sistema de verificación de fonemas para el euskara desarrollado por el grupo Aholab (Odriozola et al., 2012). La verificación se realiza mediante un ASR basado en modelos ocultos de Markov (HMM, *Hidden Markov Models*), mediante el procedimiento de alineamiento forzado. De este modo, el sistema produce un parámetro GOP (*Goodness of Pronunciation*) para cada fonema, que se utiliza como medida de confianza.

Por lo general, los sistemas CAPT utilizan grabaciones de hablantes nativos vs. no-nativos para evaluar las señales grabadas por los estudiantes. Por tanto, una base de datos desarrollada específicamente para la verificación de fonemas debe tener grabaciones de los dos tipos de hablantes. Además, utilizando conocimiento previo, se puede prever en cierta forma cuáles son algunos de los fonemas más conflictivos para el estudiante de un L1 concreto, y, por tanto, las bases de datos se completan habitualmente con grabaciones donde aparecen dichos fonemas (Demenko, Wagner, y Cylwik, 2010).

El euskara es una lengua con recursos limitados que no tiene bases de datos acústicas suficientes para desarrollar tecnologías del habla que puedan competir con sus lenguas vecinas. Actualmente, hay tres bases de datos acústicas diseñadas para crear sistemas de reconocimiento de voz: la base de datos *SpeechDat_eu* (Hernáez et al., 2003), grabada sobre la red de telefonía fija a una frecuencia de muestreo de 8 kHz; *SpeechDat_eu_M*, similar para telefonía móvil, y una base de datos *Speecon-like*, grabada con varios tipos de micrófonos a varias distancias a una frecuencia de muestreo de 16 kHz, estas dos últimas creadas bajo la financiación del Gobierno Vasco y cedidas para su uso en investigación.

La base de datos *Speecon-like* contiene grabaciones de hablantes nativos y no-nativos, así como habla dialectal y estándar. Contiene señales de audio de 230 hablantes, grabadas en diferentes partes de Euskal Herria. En cada sesión, a cada hablante se le preguntaba por su nivel lingüístico, a elegir entre las opciones "nativo/a", "nivel alto" o "nivel bajo". El subcorpus de hablantes nativos está compuesto por 149 hablantes, el subcorpus de hablantes con nivel alto por 56 hablantes, y el de hablantes con nivel bajo por 25 hablantes.

3.2 Pruebas iniciales de reconocimiento de la voz de referencia

Los modelos ocultos de Markov o HMMs utilizados en reconocimiento automático del habla por el laboratorio están entrenados utilizando los hablantes de las primeras 155 sesiones (dos tercios), ya que el tercio restante se reservó para el test. Dichos modelos se crearon para fonemas con contexto (trifonemas), utilizando vectores de 39 parámetros MFCC (*Mel Frequency Cepstral Coefficients*, coeficientes cepstrales en las frecuencias de Mel), coeficientes basados en la percepción auditiva humana. Con esos modelos se realizó una prueba de reconocimiento preliminar sobre la voz de referencia, para tener una primera idea de la validez de la voz de

referencia y una impresión preliminar sobre el funcionamiento del sistema de evaluación de la pronunciación.

Se realizaron dos pruebas con diferentes diccionarios y diferentes gramáticas para así tener una visión más global de los resultados:

a) En la primera prueba, se evaluaron todos los ficheros, tanto los creados para la parte segmental como para la suprasegmental, 255 en total. El diccionario del sistema se creó con el lexicón completo del currículum: 921 entradas léxicas (EL). La gramática utilizada en este caso era un bucle de palabras sin modelado de lenguaje, donde cada palabra sucede a la siguiente con la misma probabilidad. Entre dos palabras, cabía la posibilidad de existir un silencio, pero sin modelar la coarticulación entre palabras.

b) En la segunda prueba, se evaluaron sólo los ficheros (60 en total) que contenían pares de palabras (contrastes entre sibilantes), ya que la capacidad de discernir entre fonemas es más manifiesta en este tipo de grabaciones. Para el diccionario, se utilizaron las palabras correspondientes a dichos ficheros (118 palabras diferentes), y se diseñó una gramática simple que modelaba la secuencia de dos palabras con silencios opcionales.

Los índices de error de palabra (WER, *Word Error Rate*) obtenidos en las dos pruebas de reconocimiento pueden verse en la Tabla 1.

TEST	DICC.	Nº FICH.	WER
1	921 EL	255	64,49 %
2	118 EL	60	76,67 %

Tabla 1: Resultados de reconocimiento de la voz de referencia, con modelos globales.

A primera vista, ante el hecho de que los resultados en el test 2 no eran tan buenos como los esperados, se dedujo que los modelos que habían sido creados para reconocimiento no eran capaces de discernir bien entre sibilantes. Por tanto, se pensó en repetir las pruebas con nuevos modelos creados utilizando sólo las grabaciones de hablantes nativos de la zona este de Euskal Herria, ya que en la zona oeste el fonema /s`/ ha sido históricamente asimilado por el fonema /*s*/ (actualmente, los dos se pronuncian como /*s*/). En total, se utilizaron 76 hablantes para entrenar nuevos modelos acústicos, y, tras repetir las pruebas, los resultados pueden verse en la Tabla 2.

TEST	DICC.	Nº FICH.	WER
1	921 EL	255	65,01 %
2	118 EL	60	81,86 %

Tabla 2: Resultados de reconocimiento de la voz de ref., con modelos de hablantes nativos.

Aunque los resultados son mejores con los nuevos modelos, siguen siendo más bajos de lo esperado para tareas tan sencillas. Por tanto, se realizó una prueba de verificación para analizar cómo se comportan los modelos a la hora de discernir un fonema. Para ello, se obtuvieron las distribuciones de los parámetros GOP de cada fonema, en dos situaciones diferentes: cuando el fonema está correctamente pronunciado, y cuando el fonema no está correctamente pronunciado. Para obtener la distribución de los fonemas incorrectamente pronunciados, se realizó una simulación de pronunciación incorrecta, realizando cambios controlados en el diccionario; es decir, sustituyendo un fonema de una posición concreta de cada palabra por otro del mismo grupo fonético (vocales, plosivas, nasales, líquidas y sibilantes), de forma aleatoria.

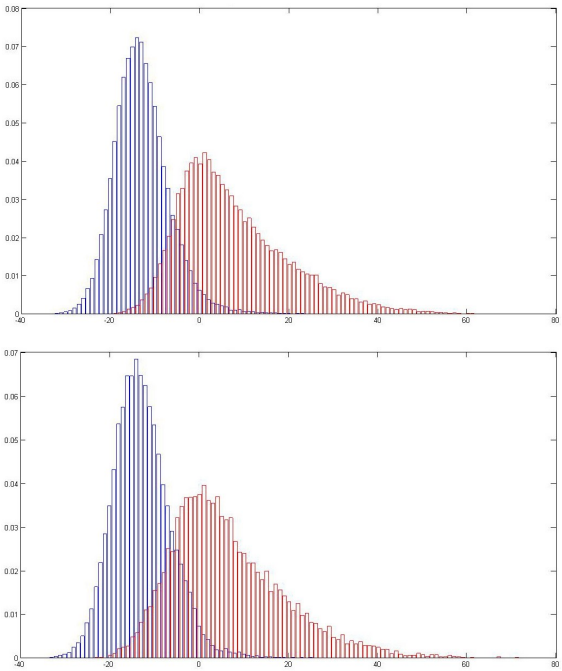

Figura 1: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /a/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

El par de distribuciones (correcto - incorrecto) se calculó para cada fonema, con los modelos del sistema reconocimiento y con los modelos entrenados usando sólo las grabaciones de los hablantes nativos.

De este modo, se vio que, por ejemplo, los modelos acústicos del fonema /*a*/ dan como resultado dos pares de distribuciones casi idénticos (ver Figura 1), lo cual indica que las diferencias entre las realizaciones fonéticas de dicho fonema son casi iguales para los hablantes nativos de euskara y no-nativos de la base de datos. Además, los gráficos muestran pares de distribuciones parcialmente separados, con lo cual se deduce que, a priori, se puede establecer un umbral de decisión sin mucha dificultad y, por lo tanto, la detección de errores de pronunciación podría dar buenos resultados.

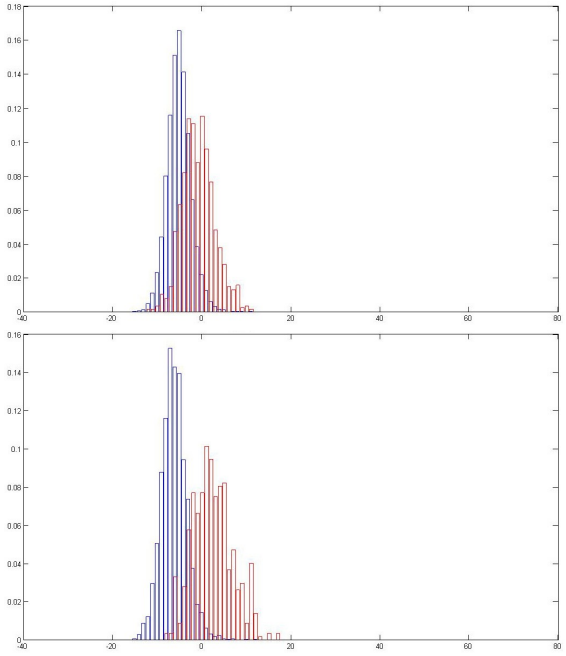

Figura 2: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /ts`/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

Por otra parte, en cuanto al fonema /*ts`*/, el cual no existe en español, los pares de distribuciones experimentan un cambio para ambos grupos de modelos (ver Figura 2). Con los nuevos modelos, como cabía esperar, las distribuciones están más separadas, lo cual quiere decir que, en principio, se obtendrán mejores resultados en la verificación de fonemas.

Por último, se analizó la variación de los pares de distribuciones del fonema /*ts*/, el cual es problemático hoy en día en todo el territorio en donde se habla euskara, y se constató que los pares de distribuciones están muy solapados, tanto para los modelos de reconocimiento como para los creados para verificación (ver Figura 3). Este fonema se neutralizó a /ts`/ en la zona oeste de Euskal Herria, mientras que en la parte central y este parece que está siendo neutralizada actualmente a /*tS*/.

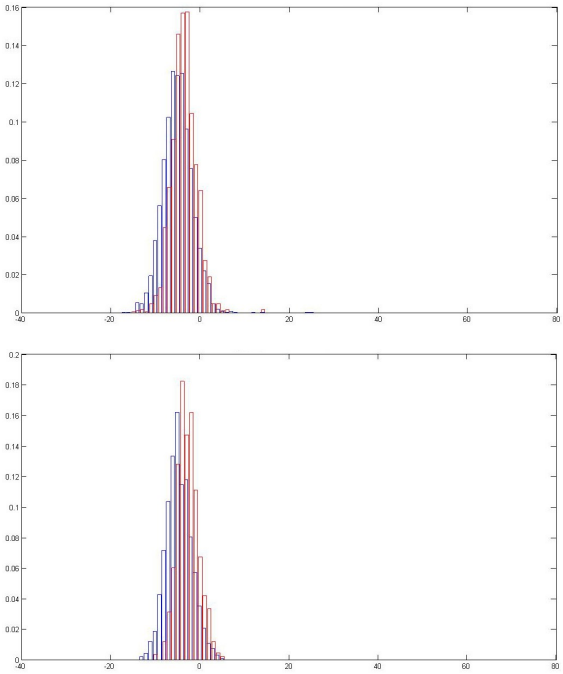

Figura 3: Pares de distribuciones (correctos: izda. vs. incorrectos: dcha.) de los GOP para el fonema /ts/, con modelos globales del sistema de reconocimiento (arriba) y modelos creados con hablantes nativos (abajo).

Para ver si realmente este modelo causa un empeoramiento de la tasa de error en nuestro sistema, se eliminó del currículum la parte donde se encontraba este fonema, y se repitieron las pruebas 1 y 2. Los resultados se muestran en la Tabla 3.

TEST	DICC.	Nº FICH.	WER
1	871 EL	235	65,70 %
2	100 EL	50	90,00 %

Tabla 3: Resultados de reconocimiento de la voz de ref., con modelos de hablantes nativos, y eliminando el fonema /ts/ del currículum.

Tras analizar las distribuciones del fonema /*ts*/, se concluye que hace falta entrenar mejor los modelos acústicos correspondientes a dicho fonema, ya que las señales que se disponen en la base de datos no tienen una buena correspondencia con sus transcripciones fonéticas, con pronunciaciones relativamente alejadas de las teóricamente esperadas.

3.3 Medida de la calidad de la segmentación de la voz de referencia

Las señales de audio se introdujeron en el sistema de reconocimiento y, por medio del procedimiento de alineamiento forzado, se obtuvo una segmentación a nivel de fonema. Para evaluar la calidad de la segmentación automática, los ficheros de audio se segmentaron también manualmente, a nivel de fonema.

La evaluación se realizó trama a trama. En total se analizaron un total de 584.612 tramas (todos los ficheros de la parte segmental), y se obtuvo un porcentaje de coincidencia de etiquetas del 88,08 %.

A primera vista, se observó que las mayores diferencias suceden en los finales de palabra, donde el sistema de reconocimiento necesita una cantidad de tramas mayor para salir del último HMM cuando éste no es el correspondiente al modelo de silencio. Esto es debido a ecos y reverberaciones que, aunque sean mínimas en las grabaciones de estudio, existen y tienen cierta presencia en la señal que el ser humano clasificaría como silencio.

Para solventar este problema, se pueden crear modelos de silencio más robustos. Sin embargo, se puede considerar que la segmentación de los fonemas es de muy buena calidad.

3.4 La entonación

Tal y como se ha explicado anteriormente, el análisis y evaluación de la prosodia se realizará a dos niveles: a nivel de palabra y a nivel de frase.

Como primer paso para la integración de la evaluación automática de la prosodia en el sistema AzAR, se ha creado una representación gráfica de las curvas de f0 de la señal de referencia y de la que el alumno ha grabado, alineadas y normalizadas. De esta forma, además de tener una referencia auditiva, el alumno tiene también una referencia visual que le permita evaluar mejor las diferencias entre las dos señales y aplicar así las correcciones necesarias.

4 Trabajo futuro

En la actualidad, el sistema de evaluación de la pronunciación para el euskara está en la fase de integración. La interfaz del software AzAR, desarrollada por el instituto IAE de la TU de Dresden, está siendo adaptada y modificada, y el sistema verificador de fonemas y de extracción de parámetros desarrollado por el grupo Aholab de la UPV/EHU está siendo integrado en el sistema AzAR.

Tras la integración, se realizarán varias pruebas de evaluación con estudiantes de euskara (L2) cuyo L1 es el español. En la actualidad se están diseñando los ejercicios y pruebas pertinentes para dicha tarea. La evaluación se realizará comparando los resultados obtenidos de forma automática con los proporcionados por un experto lingüista.

En publicaciones anteriores (Odriozola et al., 2012) se vio que la tasa de acierto (SA, *Scoring Accuracy*) del sistema, tanto en la evaluación de fonemas bien pronunciados como mal pronunciados, está alrededor del 80 %. Ese sistema proporciona sólo dos niveles a la salida, dependiendo de un umbral: si el parámetro de confianza GOP está por encima del umbral, se considera que el fonema está bien pronunciado; de lo contrario, está mal. Los sistemas actuales utilizan una salida con varios umbrales que permiten puntuar o evaluar una realización fonética con varios niveles de corrección, por lo general 3 ó 5. Este tipo de sistemas son más efectivos a la hora de interactuar con el usuario, ya que no sólo consideran que un fonema está "bien" o "mal" pronunciado (cosa que incluso para un humano puede llegar a ser una tarea compleja), sino que se tienen en cuenta graduaciones intermedias. El *feedback* suele darse con un sistema de colores donde, generalmente, el color verde corresponde a un fonema correctamente pronunciado, el color rojo a un fonema incorrectamente pronunciado, y colores intermedios para niveles intermedios. La impresión que un usuario puede llegar a percibir de un sistema puede mejorar notablemente si se implementa dicha característica.

5 Conclusiones

En este artículo se han presentado los primeros pasos para la integración del euskara en el sistema tutor de la pronunciación AzAR. Se ha concluido que, para el euskara, hacen falta dos tipos de análisis: la verificación de fonemas (para el entrenamiento de la parte segmental) y el análisis de la prosodia (para la parte suprasegmental).

La parte segmental está muy desarrollada en el sistema AzAR, pero necesita de modelos acústicos adecuados para cada lengua. En cuanto a la parte suprasegmental, se han dado los primeros pasos para implementar un sistema de evaluación automática de la prosodia en AzAR, ya que actualmente no cuenta aún con ningún *feedback* creado de forma automática. Esto constituye toda una línea de investigación que será desarrollada a corto plazo entre los dos grupos de investigación implicados en el proyecto del presente artículo.

Bibliografía

Alvarez, J.L. (Txillardegi). 1986. Proposamen bat azentuari buruz. *Euskera XXXI*, páginas 341–348 (en euskera).

Demenko, G., A. Wagner, N. Cylwik. 2010. The Use of Speech Technology in Foreign Language Pronunciation Training. *Archives of Acoustics*, 35(3), páginas 309–329.

Demenko, G., A. Wagner, N. Cylwik, O. Jokisch. 2009. An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody. En *Proc. of 2nd ISCA Workshop on Speech and Language Technology in Education*, *SLaTE*, Wroxall Abbey Estate (Reino Unido).

Ding, H., H. Mixdorff, O. Jokisch. 2010. Pronunciation of German syllable codas of Mandarin Chinese speakers. En *Proc. Konferenz Elektronische Sprachsignal-verarbeitung*, *ESSV*, páginas 281–287, Berlin (Alemania).

Hernáez, I., I. Luengo, E. Navas, M. Zubizarreta, I. Gaminde, J. Sanchez. 2003. The Basque speech_dat (II) database: a description and first test recognition results. En *Proc. of Eurospeech-2003*, páginas 1549–1552. Ginebra (Suiza).

Hualde, J. I. 1991. *Basque Phonology*. Routledge, London & New York.

Jokish, O., U. Koloska, D. Hirschfeld y R. Hoffman. 2005. Pronunciation learning and foreign accent reduction by an audiovisual feedback system. En *Proc. of 1st Intern. Conf. on Affective Computing and Intelligent Interaction*, *ACII*, páginas 419–425, Pekín (China).

Odriozola, I., E. Navas, I. Hernaez, I. Sainz, I. Saratxaga, J. Sánchez, D. Erro. 2012. Using an ASR database to design a pronunciation evaluation system in Basque. En *Proc. of 8th Inter. Conf. on Language Resources and Evaluation*, *LREC*, Estambul (Turquía), páginas 4122–4126.

University of the Basque Country (UPV/EHU), Aholab Signal Processing Laboratory's website: http://aholab.ehu.es/sampa_basque.htm