MODELADO ACÚSTICO DE LOS ERRORES EN LA REALIZACIÓN DE LOS FONEMAS PARA RECONOCIMIENTO AUTOMÁTICO DEL HABLA ALTERADA

Oscar Saz[1], Javier Simón[2], Eric Dallal[3], Eduardo Lleida[1], Richard C. Rose[3]

[1] Grupo de Tecnologías de las Comunicaciones (GTC),
Instituto de Investigación en Ingeniería de Aragón (I3A),
Universidad de Zaragoza, Zaragoza, España
[2] Departamento de Lingüística General e Hispánica,
Universidad de Zaragoza, Zaragoza, España.
[3] Department of Electrical and Computer Engineering,
McGill University, Montréal, Québec, Canada.

ABSTRACT

En esta comunicación se plantea el problema de una tarea de Reconocimiento Automático del Habla (RAH) de palabras aisladas para un conjunto de jóvenes locutores con trastornos en el habla y el lenguaje. En esta tarea, los errores en la realización de los fonemas por estos locutores (un 17.61% de los fonemas) provocan que las unidades acústicas habituales en tareas de RAH sean insuficientes para modelar la señal de voz. Los primeros resultados obtenidos en esta tarea con un conjunto de más de 700 unidades acústicas dependientes del contexto (32% de error) muestran la dificultad que supone para los sistemas de RAH enfrentarse a este tipo de habla.
Contra esto, se definen siete nuevas unidades acústicas que representan características básicas de la producción de voz (vocales, glides, oclusivas, fricativas, nasales, líquidas y vibrantes) que se usan para crear un nuevo diccionario personalizado para cada locutor, mezcla de las unidades contextuales y las genéricas. La mezcla se basa en la predicción que de los errores de realización de cada locutor da un árbol de decisión binario. Así pues, los fonemas en los que se prevé error son sustituidos por las unidades genéricas para modelar mejor la variabilidad del habla producida por los errores del locutor. Los resultados obtenidos con este nuevo diccionario no pierden prestaciones (32% de error) con respecto a los obtenidos utilizando las unidades acústicas contextuales, y muestran que las unidades acústicas planteadas en el trabajo modelan los errores de realización, abriendo nuevas vías en las tareas de RAH para este tipo de trastornos en el habla.

Sección temática: Síntesis y reconocimiento automático.

1. INTRODUCCIÓN

El colectivo de discapacitados está siempre necesitado de ayudas que faciliten su integración en la sociedad. Las nuevas tecnologías deberían ser uno de estos elementos de integración, mediante el desarrollo de sistemas que les permitan superar sus barreras. En concreto, las Tecnologías del Habla pueden ser de gran ayuda a grandes discapacitados físicos para superar sus limitaciones [1][2]. Sin embargo, muchas veces, los discapacitados físicos y psíquicos presentan también alteraciones en el habla y en el lenguaje que dificultan su posibilidad de usar sistemas de Reconocimiento Automático del Habla (RAH). Algunas de estos trastornos del habla se manifiestan como errores graves en la realización de algunos fonemas específicos por parte del paciente; esto es, confusión de un fonema por otro o eliminación de algún fonema de la palabra [3].

Este tipo de errores son muy perniciosos para los sistemas de RAH, ya que estos sistemas aprenden la estructura acústica de la señal de voz mediante un modelado acústico generalmente basado en la estructura fonética canónica de las palabras. Sin embargo, dado que cuando se enfrentan a situaciones especiales, los sistemas de RAH necesitan de adaptación al entorno para dar las mejores prestaciones, se podría plantear la posibilidad de aprender también la forma en que el usuario del sistema comete dichos errores y utilizar una serie de alternativas diferentes para esos casos.

Esta contribución se estructura de la siguiente forma: En el apartado 2 se presenta el corpus de habla alterada que se ha usado en este trabajo. Posteriormente, en el apartado 3 se presentan las técnicas de predicción basadas en árboles de decisión binarios usadas en este trabajo para predecir los errores en la realización de los fonemas y en el apartado 4 se hace un resumen de las técnicas más habituales de modelado acústico mostrando el sistema fonético empleado y su uso en la creación de los diccionarios del sistema de RAH. Finalmente, se dan los resultados del trabajo y su discusión en el apartado 5 y se extraen las conclusiones en el apartado 6.

2. CORPUS

El trabajo presentado en esta contribución se ha llevado a cabo sobre un corpus compuesto de alocuciones de jóvenes locutores entre 11 y 21 años que presentan diferentes tipos y grados de alteraciones en el habla y el lenguaje. El corpus contiene el habla de 14 locutores (7 chicos y 7 chicas) repitiendo en cuatro sesiones diferentes las 57 palabras del Registro Fonológico Inducido (RFI) [4], para un total de 3192 palabras aisladas. Por

otro lado, también se ha adquirido un corpus con habla infantil y juvenil normalizada, disponiéndose de 168 locutores entre 11 y 18 años repitiendo cada uno de ellos una sesión del RFI, para un total de 9576 palabras aisladas. Para las grabaciones se ha contado con la colaboración del Colegio Público de Educación Especial (C.P.E.E.) “Alborada”, del Colegio de Educación Infantil y Primaria “Río Ebro” y el Instituto de Enseñanza Secundaria “Tiempos Modernos”, todos situados en Zaragoza, y a los que los autores agradecen su colaboración para llevar a cabo este trabajo. La colaboración con el C.P.E.E. “Alborada” ha fructificado ya en diversos trabajos orientados a la aplicación de las tecnologías del habla a la logopedia [5].

Como trabajo adicional a la adquisición de este corpus, se ha llevado a cabo una anotación manual por parte de un grupo de 10 evaluadores independientes de la forma en que los locutores han realizado los fonemas respecto a la pronunciación canónica de la palabra. Cada serie del RFI dicha por cada locutor ha sido sometida a la evaluación de tres de estos evaluadores que han dado una nota a cada fonema dentro de la palabra (0, 1 ó 2) dependiendo de si el fonema ha sido eliminado, incorrectamente realizado o correctamente realizado respectivamente. De acuerdo a esta anotación, un 17.61% de los fonemas han sido eliminados o incorrectamente realizados de media entre todos los locutores. Los resultados completos se presentan en la Tabla 1. La tasa de acuerdo entre cada pareja de anotadores es de un 85.81%.

Loc.	#001	#002	#003	#004	#005	#006	#007
Err.	1.11%	21.57%	5.22%	3.17%	43.49%	0.68%	12.93%
Loc.	#008	#009	#010	#011	#012	#013	#014
Err.	30.82%	8.22%	21.49%	6.76%	25.68%	56.42%	8.99%

Tabla 1: Porcentaje de errores fonéticos realizados por cada locutor.

3. ÁRBOLES DE DECISIÓN BINARIOS

Para llevar a cabo la predicción de los errores de pronunciación, se ha aplicado una estrategia de árboles de decisión binarios [6]. Los árboles de decisión binarios se han venido usando como una forma sencilla de realizar predicciones basadas en un entrenamiento previo. En nuestro caso, se plantea usar los árboles de decisión como el medio de predecir el resultado de la anotación manual sobre el corpus de habla alterada de que se dispone para este trabajo.

Los árboles de decisión binarios se basan en una estructura de árbol de tal forma que en cada rama del árbol se realiza una pregunta cuya respuesta sólo puede ser ‘Sí’ o ‘No’ y se sigue descendiendo por el árbol de acuerdo a la respuesta obtenida en la última rama hasta que se llega a una hoja, que no es sino una rama en la que ya no hay más preguntas, dándose un resultado final. El método llevado a cabo en este trabajo consiste en entrenar un árbol de decisión binario para cada locutor a partir de la transcripción canónica y de la anotación manual realizada sobre el corpus. Las preguntas de cada rama se basan en conocer el modo de articulación del fonema a evaluar o del contexto inmediatamente anterior o posterior (contexto izquierdo o derecho). Un ejemplo del funcionamiento de los árboles de decisión se da en la Fig. 1.

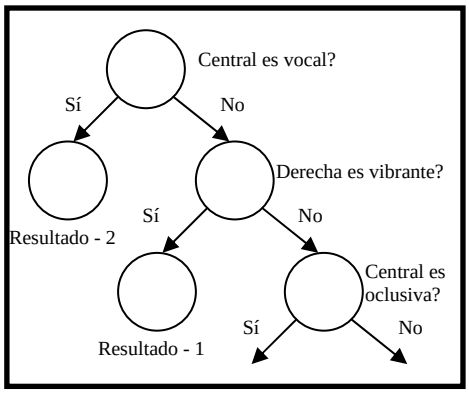

Figura 1: Ejemplo de árbol de decisión binario para la tarea expuesta.

4. MODELADO ACÚSTICO

El modelado acústico en sistemas de RAH es la forma en que se recoge la estructura de la señal de voz para poder comparar la señal de entrada con dicho modelo matemático de la voz. Como se explica de forma profunda en [7] el modelado acústico más usado hoy en día se basa en cadenas de Markov que modelan las características aleatorias variantes con el tiempo de la señal de voz. En el modelado acústico actual, la voz se considera la concatenación de un conjunto de unidades acústicas que representan partes del discurso. De esta forma, estas unidades acústicas pueden ser tan grandes como la palabra o tan pequeñas como el fonema (o incluso menores) pasando por la sílaba o la demisílaba. Sin embargo, ya se ha comprobado en varios trabajos[8] que este tipo de modelado acústico tan cerrado puede demostrarse poco útil en casos de trastornos en el habla y el lenguaje.

El modelado acústico que se propone como base en este trabajo se basa en una estructura fonética de 25 fonemas, que son los 23 fonemas definidos según la estructura de Alarcos [9] junto con la inclusión de los dos alófonos /j/ y /w/ que representan las glides de la /i/ y la /u/. Dado que en muchas ocasiones se han obtenido mejores resultados en RAH haciendo las unidades más específicas y bajando a niveles inferiores al fonema que recojan el contexto, el diccionario definitivo se obtiene utilizando un conjunto de más de 700 de estas unidades. La transcripción de la palabra 'árbol', contenida en el RFI, según este conjunto de unidades se puede ver en la Tabla 2.

[# < a] [a] [a > ɾ] [a < ɾ] [ɾ] [ɾ > B] [ɾ < B] [B] [B > o] [B < o] [o] [o > l] [o < l] [l] [l > #]
[# < vocal] [vocal] [vocal > vibrante] [vocal < vibrante] [vibrante] [vibrante > oclusiva] [vibrante < oclusiva] [oclusiva] [oclusiva > vocal] [oclusiva < vocal] [vocal] [vocal > liquida] [vocal < liquida] [liquida] [liquida > #]
[# < a] [a] [a > ɾ] [a < vibrante] [vibrante] [vibrante > B] [ɾ < B] [B] [B > o] [B < o] [o] [o > l] [o < liquida] [liquida] [liquida > #]

Tabla 2: Diccionarios de la palabra 'árbol' para unidades específicas, genéricas y mezcla.

Por otra parte, la propuesta que se realiza en este trabajo es que unidades acústicas más genéricas que en principio modelarían peor la señal de voz para RAH puede ser muy útiles en el caso en el que nos encontramos en el que los locutores cometen errores graves en la realización de los fonemas, pero pueden mantener algunas características de comunes relativas al proceso de producción de la voz. Así pues, toda la estructura fonética de 25 fonemas se reduce a 7 unidades que representan los posibles modos de articulación de los sonidos, esto es: vocal, glide, oclusiva, nasal, fricativa, líquida o vibrante. Esto es traduce en un conjunto de 40 unidades acústicas que mantienen la posibilidad de recoger el contexto, la Tabla 2 muestra así la transcripción de la palabra 'árbol' en estas unidades.

Por último, el experimento final que se quiere realizar es el uso conjunto de las unidades específicas y genéricas que se mezclan según la predicción que de los errores de pronunciación da un árbol de predicción binario como se ha visto en el apartado 3. En este caso, donde el árbol de predicción indique que no habrá error se usarán las unidades específicas, mientras que los fonemas en los que se prediga un error se sustituirán por las unidades genéricas. De

esta forma, la transcripción de la palabra ‘árbol’ según este diccionario para el locutor #005 al que su árbol de decisión binario predice errores en el fonema /ɾ/ y en el /l/ se puede ver también en la Tabla 2.

5. RESULTADOS Y DISCUSIÓN

Los resultados de RAH para los 14 locutores del corpus utilizando modelos acústicos de 700 unidades específicas entrenados con habla infantil normalizada son de un 31.42% de media para los 14 locutores. El uso de sólo las 7 unidades genéricas aumenta la tasa de error a un 64.19% de media (104.30% de pérdida de prestaciones); mientras que la combinación de ambos tipos de unidades según los árboles de decisión binarios mantiene el error en un 32.65% de promedio (3.91% de pérdida). Los resultados individualizados por locutor se presentan en la Tabla 3.

Loc.	#001	#002	#003	#004	#005	#006	#007
Err. 1	15.35%	22.37%	10.09%	4.39%	56.14%	5.70%	35.96%
Err. 2	46.49%	71.49%	61.84%	39.47%	93.41%	39.47%	69.74%
Err. 3	15.35%	23.25%	10.09%	4.39%	60.96%	5.70%	38.16%
Loc.	#008	#009	#010	#011	#012	#013	#014
Err. 1	52.63%	22.37%	38.16%	9.65%	65.79%	78.95%	22.37%
Err. 1	79.39%	49.12%	72.37%	48.25%	82.90%	93.42%	51.32%
Err. 3	53.07%	22.37%	41.67%	9.65%	66.67%	83.77%	21.93%

Tabla 3: Tasa de error en RAH por cada locutor con unidades específicas (Err.1), genéricas (Err. 2) y mezcla (Err. 3)

Sin embargo, un resultado interesante para discusión puede ser comparar los resultados que se obtiene introduciendo las unidades genéricas según nos indica el árbol de decisión binario frente a introducir las unidades genéricas de una forma aleatoria. Para ello, realizaremos un experimento con el locutor #005 consistente en lo siguiente:

- Identificar el porcentaje de fonemas erróneamente realizados según la predicción del árbol de decisión (56.14% para el locutor #005).
- Generar 11 diccionarios distintos introduciendo de forma incremental las unidades genéricas, desde un 0% (sólo unidades específicas) a un 100% (sólo unidades genéricas) de 10% en 10%. Las unidades genéricas se introducen de forma totalmente aleatoria en el diccionario sustituyendo a las unidades específicas según el porcentaje indicado.

- Obtención de los resultados de reconocimiento para esos 11 diccionarios. Los resultados para el Locutor #005 se muestran en la Fig. 2. Posteriormente, se obtiene la recta de regresión que la caracteriza:

$$WER_{pol}(\%) = 0.001 \cdot x^2 + 0.263 \cdot x + 56.85 \quad (1)$$

donde x es el porcentaje de unidades genéricas en el diccionario.

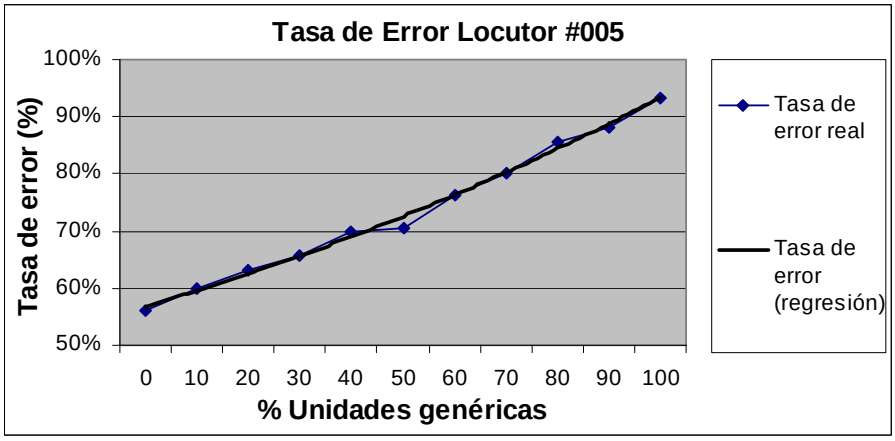

Figura 2: Tasa de error usando las unidades genéricas de forma aleatoria

- Finalmente, cálculo del resultado de esta recta para el porcentaje de fonemas erróneos obtenido en el primer punto (56.14%), que para nuestro caso es de 74.77%.

Así pues, comparando este valor (74.77%) con el obtenido con el diccionario creado a partir del árbol de decisión (60.96%) se aprecia una ganancia del 22.65% que nos indica cómo las unidades genéricas pueden servir para modelar los errores en la realización de los fonemas cuando se es capaz de predecirlos de alguna forma.

6. CONCLUSIONES

En esta comunicación se ha presentado la problemática existente en el RAH para locutores con algún tipo de alteración en el habla y el lenguaje. En concreto, sobre un corpus de 14 jóvenes locutores se obtiene una tasa de error del 31.42% sobre unos modelos adaptados al habla infantil y juvenil. La anotación llevada a cabo sobre ese corpus nos muestra que los locutores cometen una media de un 17.61% de errores respecto a la realización fonética canónica de la palabra. Los resultados muestran como el empleo de

unidades acústicas genéricas producen una gran pérdida de prestaciones, mientras que la mezcla de unidades específicas y genéricas, utilizando estas últimas como sustitutos de las específicas en los casos en que un árbol de decisión binario prediga la presencia de un error, mantienen las prestaciones sobre la línea de base, aunque mejoran la situación que se da en el caso de usar las unidades genéricas sin usar ningún tipo de predictor.

7. AGRADECIMIENTOS

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

8. REFERENCIAS

[1] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, M. Parker: "Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers". 8th European Conference on Speech Communication and Technology. Ginebra (Suiza), 2003.

[2] J.M. Alcubierre, J. Mínguez, L. Montesano, L. Montano, O. Saz, E. Lleida: "Silla de ruedas inteligente controlada por voz". Primer Congreso Internacional de Domótica, Robótica y Teleasistencia para Todos. Madrid (España), 2005.

[3] P. Perellò. "Trastornos del Habla y Trastornos del Lenguaje". Ed. Científico Médica, Barcelona. 1984.

[4] M. Monfort, A. Juárez-Sánchez. "Registro Fonológico Inducido (Tarjetas Gráficas)". Ed. Cepe, Madrid, 1989.

[5] C. Vaquero, O. Saz, E. Lleida, W.R. Rodríguez: "E-inclusion Technologies for the speech handicapped". 2008 International Conference on Acoustic, Speech and Signal Processing. Las Vegas (NV), USA, Abril 2008.

[6] P.H. Swain, H. Hauska: "The decision tree classifier: Design and potential". IEEE Transactions on Geoscience Electronic, vol. 15, pp. 142—147, Julio 1977.

[7] X. Huang, A. Acero, H.-W. Hon: "Spoken Language Processing: A Guide to Theory, Algorithm and System Development". Ed. Prentice-Hall, Mayo 2001.

[8] J.R. Deller Jr., D. Hsu, L.J. Ferrier: "On the use of Hidden Markov Modelling for recognition of dysarthric speech". Computer Methods and Programs in Biomedicine, vol. 35, pp. 125—139. 1991.

[9] E. Alarcos: "Fonología Española". Ed. Gredos, Madrid. 1950.