Guimarães - Portugal
acústica 2004

paper ID: 112 /p.1

Reconocimiento Fonético en Habla Contínua Usando Información de Segmentos Adyacentes

S. Fernández[a] and S. Feijóo[b]

[a] *Department of Phonetics & Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom, santi@phon.ucl.ac.uk*
[b] *Departamento de Física Aplicadal, Universidad de Santiago de Compostela, Facultad de Física, Campus Sur, 15782 Santiago de Compostela, España*

RESUMEN: Los modelos tri-fono son ampliamente utilizados para el reconocimiento automático de fonemas en habla contínua debido a su buen rendimiento en la mayoría de los sistemas. En base a estudios previos, se propone que los modelos acústicos de unidades fonéticas se beneficiarían de la inclusión de parte de la señal acústica adyacente. Por el contrario, los modelos tri-fono emplean, al menos teóricamente, sólo el segmento acústico correspondiente a cada fonema. Para probar la validez de dicha hipótesis, a partir de una gran base de datos de habla contínua en inglés británico, se construyeron modelos que incluyen el segmento fonético más la mitad de cada uno de los segmentos adyacentes. El porcentaje de acierto obtenido por el sistema de referencia basado en modelos tri-fono fue del 72%. El modelo propuesto fue probado en una segunda etapa a partir de la segmentacion obtenida con el sistema de referencia. No se observó una mejora significativa. En consecuencia, se propuso la hipótesis de que la segmentación obtenida con el sistema de referencia limitaba las posibilidades del modelo propuesto. Los resultados obtenidos a partir de la transcripción fonética suministrada con la base de datos fueron: 72% para los modelos tri-fono y 79% para el modelo propuesto (un 25% de reducción de la tasa de error).

ABSTRACT: Tri-phone models are the most widely used phonetic models for automatic recognition of continuous speech because of its good performance in most cases. On the basis of previous work it is argued that acoustic models of phonetic units would benefit from including part of the adjacent acoustic signal. This is in contrast with tri-phone models which, at least theoretically, rely only on the acoustic signal associated with a particular phonetic segment. In order to test this hypothesis, models including the phonetic segment plus half of the adjacent phonemes were trained and tested on a large corpora of spoken British English. Accuracy of the baseline tri-phone-based system was 72%. Recognition using the proposed models was carried out as a second stage based on the segmentation provided by the baseline system. No significant overall improvement was achieved. It was hypothesized that the tri-phone-based segmentation was limiting the possibilities of the new models. Results using the phonetic transcription provided with the database were 72% for tri-phones and 79% for the proposed models (25% error reduction).

1. INTRODUCCIÓN

El reconocimiento automático de fonemas en habla contínua es un problema complejo debido principalmente a la falta de correspondencia uno-a-uno entre segmentos acústicos y categorías fonéticas. Una primera aproximación consiste en emplear modelos mono-fono, llamados así porque cada modelo representa una categoría fonética determinada. En este caso, una

sentencia es representada como una sucesión de modelos mono-fono asociados a segmentos disjuntos y consecutivos de la señal acústica. La probabilidad para una secuencia determinada es obtenida asumiendo independencia entre modelos, es decir, multiplicando la probabilidad obtenida para cada mono-fono. Aquella sucesión con mayor probabilidad es seleccionada como correcta. El resultado de la decodificación consiste pues tanto en la secuencia de fonemas que forman la sentencia como en la segmentación de la señal acústica que proporciona dicha secuencia.

Dado que la producción del habla es un proceso dinámico, las propiedades acústicas de cada segmento fonético están influídas por los fonemas pronunciados anterior y posteriormente. Debido a ello es habitual emplear modelos de unidades (sub-)fonéticas para cada contexto. Los modelos más empleados son denominados tri-fono. Se define un modelo tri-fono para cada segmento fonético pronunciado en el contexto de un fonema inmediatamente anterior o posterior diferente. Una sentencia es, pues, representada como una sucesión de modelos tri-fono asociados a segmentos disjuntos y consecutivos de la señal acústica. La decodificación de la señal se lleva a cabo de forma similar a como se realiza con modelos mono-fono con la salvedad de que sólo son permitidas aquellas secuencias de modelos tri-fono en las que la identidad de cada modelo casa con el contexto en el que se encuentra.

Precisamente la dinámica del proceso de producción del habla implica que existe información acústica más allá del segmento correspondiente a un determinado fonema y que puede ser usada para identificar dicho fonema. De hecho, los oyentes emplean información acústica presente en la transición de un segmento fonético a otro para identificar los fonemas. Se han propuesto, pues, alternativas para modelar la trayectoria de los parámetros acústicos incluyendo la región de transición entre fonemas [1] o incluso teniendo en cuenta intervalos temporales del orden de la duración de una sentencia [2]. Varios sistemas de este tipo parten de una segmentación previa llevada a cabo, por ejemplo, con modelos tri-fono a partir de la cual se establecen los segmentos de señal que serán contrastados con los modelos propuestos. La segmentación obtenida con modelos tri-fonos usualmente no es óptima [3], en el sentido de que no establece satisfactoriamente las fronteras entre fonemas. Es por ello que generalmente los análisis se llevan a cabo empleando decenas de segmentaciones propuestas por un sistema basado en los modelos tri-fono. Los resultados muestran mejoras significativas del porcentaje de reconocimiento con respecto al sistema de referencia, aunque modestas. En algunos casos, dicha mejora sólo se obtiene incluyendo entre las posibles hipótesis la segmentación llevada a cabo manualmente.

Además de la información acústica debida a la dinámica de la producción del habla, otros factores presentes en el contexto en el que se encuentra un fonema ejercen una clara influencia en su identificación. van Son y Pols [4] mostraron que en sílabas CV la correcta identificación tanto de la consonante como de la vocal, depende de haber identificado correctamente el otro fonema presente en la sílaba. Una correcta identificación del contexto permite reconocer los fonemas con mayor precisión. Para sílabas VC no se encontró una relación similar. A la vista de los resultados obtenidos, concluyeron que las sílabas CV parecen ser procesadas de forma “integrada” por los oyentes, mientras que en las sílabas VC la identificación de los fonemas parece ser llevada a cabo de forma más independiente. La inter-dependencia entre fonema y contexto contrasta con la condición de independencia entre segmentos fonéticos que se emplea en los sistemas basados en modelos mono-/tri-fono.

Una posible alternativa para modelar dicha relación de inter-dependencia es el empleo de sílabas como unidades de reconocimiento. Los sistemas basados en unidades tipo-sílaba plantean varios problemas. Principalmente, la elección de un inventario de sílabas adecuado, el problema de sub-entrenamiento de los modelos y la eliminación de unidades durante el proceso de decodificación. Es necesario elegir un conjunto adecuado de sílabas, pero existe todavía cierta controversia sobre cómo dividir las palabras en sílabas e incluso en ciertos casos algunos fonemas pueden pertenecer a sílabas diferentes [5]. El número de sílabas puede llegar a ser extremadamente elevado, lo que implica problemas para obtener modelos representativos para cada unidad dado que no hay suficientes ejemplos en la base de datos. Ganapathiraju *et al.* [5] encontraron que el 80% de las señales empleadas para entrenar los modelos cubrían el 3% del total de modelos silábicos a emplear. Usar modelos silábicos dependiendo del contexto en que se encuentre la sílaba implica que millones de modelos han de ser entrenados. Además, dado que las sílabas son más complejas que los fonemas, es de esperar que sea necesario un mayor número de ejemplares. Ganapathiraju *et al.* [5] optaron por un sistema híbrido con sílabas y fonemas con resultados superiores a los obtenidos con tri-fonos a pesar de que los modelos sílabicos adolecían de un problema de sub-entrenamiento. Otro problema importante es que debido al modelado de segmentos tan largos (del orden de 200 ms) dependiendo del proceso de decodificación empleado puede eliminarse un número importante de sílabas cortas presentes en la señal. Una alternativa empleada consiste en usar como punto de partida una segmentación previa que detecte, por ejemplo, la presencia de vocales en la señal.

El modelo propuesto en este trabajo pretende incluir aquella información presente en segmentos adyacentes a los fonemas que se ha observado que tiene importancia para identificar dichos fonemas. Los modelos hacen enfásis en la identificación de fonemas, lo cual facilita mantener un inventario sencillo y coherente de unidades. Por último, se emplearán técnicas que permitan la integración de los modelos en el marco de los actuales sistemas de reconocimiento de habla.

2. MODELO ACÚSTICO

La información de contexto empleada por los oyentes a la hora de identificar los fonemas consiste en, primero, la información acústica debido a la dinámica del proceso de producción y, segundo, la identidad de dichos segmentos adyacentes. Hay que notar que los oyentes no necesitan la totalidad del segmento adyacente para hacer uso de la información de contexto. Ello implica que los oyentes no podrán, posiblemente, establecer con precisión la identidad de dicho segmento adyacente pero sí parte de sus características. En experimentos previos hemos encontrado diferencias en el uso del contexto a la hora de identificar consonantes fricativas cuando se añade parte de la vocal siguiente dependiendo de si dicha vocal es anterior o posterior. Aunque la identidad del segmento vocálico pueda no estar totalmente clara, los oyentes podrían emplear parte de las caracterísiticas que lo definen.

Los modelos tri-fono ya incluyen, en cierto modo, ambas fuentes de información pero se asume independencia entre los segmentos fonéticos. Al menos para sílabas CV esto no es cierto [4]. En [6] se ha comprobado que, en comparación con los modelos tri-fono, se

obtienen mejores resultados de clasificación y mejor correlación con la identificación de los oyentes cuando se modela de forma conjunta el segmento fonético y parte de los segmentos adyacentes.
Se propone, por lo tanto, incluir parte de la señal correspondiente al segmento fonético previo y posterior en los modelos para un segmento fonético particular, de forma que tanto la dinámica de los parámetros acústicos como determinadas características relacionadas con la identidad de los segmentos adyacentes son modeladas de forma conjunta con la información presente en el segmento fonético a identificar.
Los modelos propuestos son, pues, similares a tri-fonos y una sentencia es representada como una sucesión de modelos fonéticos con la salvedad de que ahora dichos modelos están asociados a segmentos de la señal que se solapan (ver figura 1). Esto dificulta la segmentación automática de la señal, de forma que a partir de una segmentación fonética previa realizada, por ejemplo, con modelos tri-fono, se seleccionará cada segmento fonético más la mitad de los segmentos anterior y posterior y esta señal acústica será la que se contraste con los modelos propuestos para determinar el fonema presente. El resto del proceso de decodificación se realiza de forma similar a como se hace con modelos tri-fono. Se usará el término *tri-fono ampliado* o *modelos ampliados* para referirse a los modelos propuestos.

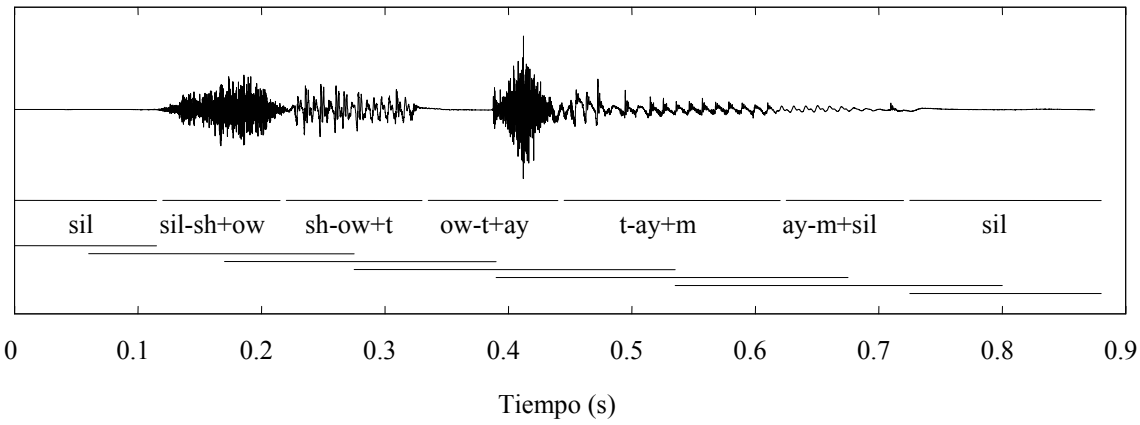

Figura 1 – *Segmentación de la señal "showtime" mediante modelos tri-fono (arriba) y mediante los modelos propuestos (abajo). En medio se muestran las etiquetas asociadas a cada modelo/segmento (fonema anterior-fonema+fonema posterior). En la decodificación sólo son válidas aquellas secuencias de modelos que casan de acuerdo con el contexto.*

3. MATERIALES Y MÉTODO

Las señales empleadas en los experimentos provienen de la versión inglesa de la base de datos Wall Street Journal (WSJCAM0). Las grabaciones seleccionadas han sido realizadas con un micrófono cercano de diadema. El conjunto de entrenamiento consiste en sentencias leídas por 92 hablantes para un total de 7861 sentencias. El conjunto de prueba esta formado por las sentencias leídas por 20 hablantes para un total de 368 sentencias (ref: SI_DT5A). La base de

datos incluye la segmentación fonética de las sentencias obtenida mediante reconocimiento forzado. El inventario fonético empleado incluye 44 categorías fonéticas más la categoría de "silencio".

Nuestro sistema de referencia fue desarrollado empleando el programa HTK (http://htk.eng.cam.ac.uk/). Las señales fueron caracterizadas calculando sobre ventanas de 25ms, solapadas 15ms, 12 coeficientes MFCC más la componente de energía y los coeficientes Δ y ΔΔ de estos 13 parámetros. En primer lugar un conjunto de 44 modelos ocultos de Markov mono-fono (más un modelo para el silencio) fueron creados. Los modelos son del tipo izquierda-a-derecha con 3 estados. Las probabilidades de observación fueron modeladas mediante una sola gaussiana usando una matriz de covarianza diagonal. A partir de estos modelos se crearon modelos tri-fono del tipo izquierda-a-derecha con 3 estados y probabilidades de observación modeladas por una sola gaussiana. La técnica de agrupamiento basada en árboles de decisión fue utilizada en este punto para ligar aquellos estados similares y estimar así de manera robusta los modelos. El árbol de decisión es similar al incluído en HTK pero adaptado a las categorías fonéticas presentes en nuestro inventario. 17287 modelos tri-fono fueron obtenidos para representar al total de 85185 modelos tri-fono posibles con 44 categorías fonéticas más la categoría silencio. Finalmente, el número de gaussianas fue incrementado hasta ocho.

El proceso de creación de los modelos tri-fono ampliado se llevó a cabo de manera similar con la diferencia de que, mientras los tri-fonos son obtenidos mediante un proceso de entrenamiento incrustado (*embedded training*), los tri-fonos ampliados son obtenidos mediante entrenamiento aislado. Los segmentos acústicos que los tri-fonos ampliados han de modelar se extraen de las sentencias de acuerdo con la segmentación de referencia. Los modelos ampliados constan de 7 estados, con el objeto de representar el estado inicial, medio y final del segmento fonético central, más las regiones de transición y estable de la porción incluida de los dos segmentos fonéticos adyacentes. El número total de modelos despúes del agrupamiento es de 36585 para representar un total, de nuevo, de 85185 modelos. En este caso, el agrupamiento de estados implica que, para un modelo dado, la identidad de los fonemas adyacentes puede no estar inequivocamente determinada.

La decodificación de las señales de prueba con los modelos tri-fono se realiza de la forma habitual (salvo que se indique lo contrario) utilizando el algorítmo de Viterbi mediante *paso de testigo*. Para los modelos ampliados, la decodificación parte de una segmentación previa de la que se extraen los segmentos de señal que incluyen parte de los fonemas adyacentes. La decodificación de dicha sucesión de segmentos se realiza con un decodificador propio usando el algorítmo de Viterbi. En cualquier caso el reconocimiento es puramente fonético, sin emplear ningún tipo de gramática u otra información aparte de las probabilidades de observación de las unidades fonéticas.

4. RESULTADOS

El porcentaje de acierto (teniendo en cuenta inserciones, sustituciones y eliminaciones) obtenido con los modelos tri-fono es del 71.86%. La segmentación obtenida con los tri-fonos fue usada para decodificar la secuencia empleando los modelos ampliados. El porcentaje de

acierto con los modelos ampliados es del 71.33%. Muy similar al obtenido por los tri-fonos. No se encontraron diferencias significativas entre los porcentajes de reconocimiento de las diferentes categorías fonéticas obtenidos por cada modelo. Dado que la segmentación obtenida con tri-fonos es óptima en cuanto a la decodificación de la señal con modelos tri-fono, es habitual que nuevos modelos no obtengan una mejora significativa. Generalmente, se extraen decenas de segmentaciones con modelos tri-fono y se re-evaluan con los nuevos modelos, eligiendo aquella que proporciona los mejores resultados. Debido a limitaciones de tiempo, se realizó el reconocimiento a partir de la segmentación proporcionada con la base de datos. Dado que dicha segmentación fue obtenida de forma automática mediante reconocimiento forzado, los resultados proporcionarán una estimación de cuál sería la mayor diferencia encontrada entre ambos tipos de modelos si se realizara el procedimiento de re-evaluación descrito anteriormente.

Los resultados usando con ambos tipos de modelos la segmentación proporcionada con la base de datos fueron 71.99% para los tri-fonos y 78.85% para los tri-fonos ampliados. Como puede observarse, los tri-fonos no obtienen una mejora apreciable con respecto a la decodificación habitual. Ello es muestra de que los modelo tri-fono obtienen el mejor resultado posible con el proceso usual de decodificación. Los modelos tri-fono ampliado reducen, por el contrario, la tasa de error en un 25% con respecto a los tri-fonos. En la figura 2 pueden observarse las diferencias para cada categoría fonética por separado. El número de fonemas correctos para cada modelo es significativamente diferente (χ^2=96.48, gl=44, $p<8.5\cdot10^{-6}$). No se observa, sin embargo, un patrón claro respecto qué fonemas resultan más beneficiados de la inclusión del contexto. Mientras que para las vocales la mejoría es generalizada (χ^2=53.47, gl=19, $p<4.0\cdot10^{-5}$) para las consonantes no existe una diferencia significativa (χ^2=12.36, gl=23, p<0.96). Torre-Toledano *et al.* [3] observaron que la segmentación obtenida con modelos tri-fono muestra que éstos modelan el segmento fonético y parte del contexto adyacente o bien sólo parte del segmento fonético (dado que modelos consecutivos se corresponden con segmentos consecutivos y disjuntos de la señal acústica). La inclusión de parte del contexto tiene lugar especialmente para aquellos fonemas menos

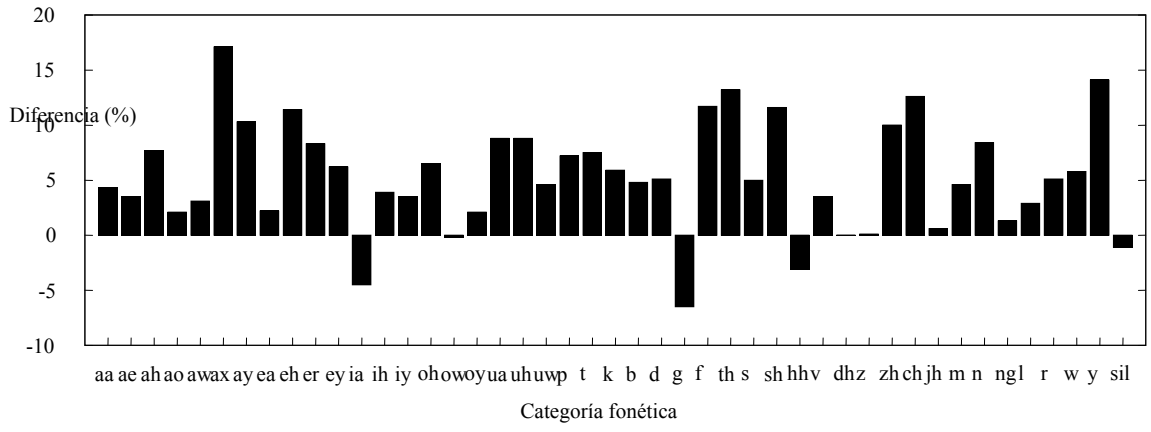

Figura 2 – *Diferencia entre los porcentajes de reconocimiento obtenidos para cada categoría fonética con los modelos propuestos (tri-fono ampliado) y con los modelos de referencia (tri-fono). Una diferencia positiva indica que los modelos propuestos obtienen mejor resultado.*

estacionarios (plosivas, africadas, etc.). De hecho este efecto resultó ser sistemático, y la segmentación obtenida con tri-fonos pudo ser corregida en gran medida para obtener una segmentación fonética más precisa. Por lo tanto, el uso del contexto en el reconocimiento fonético es necesario. Los modelos tri-fono emplean parte de los segmentos adyacentes de forma sistemática para mejorar el reconocimiento pero, obviamente, dentro de las limitaciones impuestas (los segmentos acústicos modelados han de ser consecutivos y disjuntos). En nuestro caso, se permite el solapamiento entre los segmentos acústicos modelados, haciendo un mejor uso del contexto que mejora el reconocimiento fonético.

Aunque no disponemos de una segmentación manual de la base de datos para contrastar con la segmentación obtenida con modelos tri-fono, se ha observado en la segmentación obtenida mediante reconocimiento forzado, que los modelos tri-fono correspondientes a consonantes incluyen habitualmente parte de los segmentos adyacentes. Dichos segmentos adyacentes (habitualmente vocales) son, pues, modelados usando sólo parte del segmento fonético correspondiente. Ello explicaría por qué la mejora con los modelos ampliados es generalizada para las vocales y más irregular en el caso de las consonantes.

Asumiendo que el uso del contexto es especialmente importante para la robustez del sistema, se realizaron experimentos en condiciones ruidosas empleando los modelos entrenados sobre señales limpias. Al conjunto de prueba se le añadió ruido blanco con una relación señal/ruido de 20 dB. El sistema de referencia empleando modelos tri-fono obtuvo un porcentaje de acierto del 42.96%. Usando la segmentación obtenida con este sistema los modelos tri-fono ampliado obtuvieron un porcentaje de acierto del 45.46%. En este caso, la segmentación con modelos tri-fono sí da lugar a una reducción de la tasa de error.

Si se usa la segmentación proporcionada con la base de datos, el reconocimiento con modelos tri-fono mejora con respecto al caso anterior: 47.71%; pero el incremento es todavía más sustancial para los modelo tri-fono ampliado: 56.90%. Por lo tanto, se consigue una reducción de la tasa de error del 17% con respecto al sistema basado en modelos tri-fono. Un mejor uso de la información de contexto que, en este caso, probablemente implica además el uso de información presente en regiones con una mejor relación señal-ruido (por ejemplo, para el reconocimiento de las consonantes se empleará en muchos casos parte de las vocales adyacentes), contribuye a reducir la tasa de error en presencia de ruido.

5. CONCLUSIÓN

En base a la información de contexto que se ha observado que es relevante para la percepción de los fonemas se han propuesto nuevos modelos fonéticos. Estos modelos hacen uso de la información presente en los segmentos acústicos adyacentes al segmento fonético que se pretende modelar: a) información acústica debida a la dinámica de producción del habla, y b) identidad de los segmentos adyacentes. Los modelos mantienen un inventario de unidades

reducido y se emplean técnicas que podrían permitir su inclusión en el marco matemático que forma actualmente la base de los sistemas de reconocimiento automático.
Los resultados muestran una reducción de hasta el 25% de la tasa de error. Los modelos propuestos mejoran también el reconocimiento fonético de señales en presencia de ruido, con una reducción de hasta el 17% de la tasa de error para una relación señal-ruido de 20 dB.
Sería adecuado, a partir de este punto, establecer un sistema previo de segmentación óptimo para el empleo de los modelos o la modificación de los algorítmos de decodificación para permitir que la segmentación y el reconocimiento se llevaran a cabo de forma conjunta.

RECONOCIMIENTOS

S. Fernández ha sido financiado mediante una beca Marie Curie del programa de la Comunidad Europea "Improving the Human Research Potential and the Socio-Economic Knowledge Base" bajo contrato número HPMF-CT-2002-02129.

REFERENCIAS

[1] K. Reinhard y M. Niranjan; *Diphone subspace mixture trajectory models for HMM complementation.* Speech Communication, vol. 38, págs. 237-265, 2002.

[2] L. Deng y J. Ma; *Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics.* Journal of the Acoustical Society of America, vol. 108, págs. 3036-3048, 2000.

[3] D. Torre Toledano, L. A. Hernández Gómez y L. Villarrubia Grande; *Automatic phonetic segmentation.* IEEE Transactions on Speech and Audio Processing, vol. 11, págs. 617-625, 2003.

[4] R. J. J. H. van Son y L. C. W. Pols; *Perisegmental speech improves consonant and vowel identification*. Speech Communication, vol. 29, págs. 1-22, 1999.

[5] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski y G. R. Doddington; *Syllable-based large vocabulary continuous speech recognition.* IEEE Transactions on Speech and Audio Processing, vol. 9, págs. 358-366, 2001.

[6] S. Fernández y S. Feijóo; *Modelos acústicos de sílabas consonante-vocal para el reconocimiento de fricativas.* European Acoustics Symposium, Guimaraes, 2004.