Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 121-128 recibido 1-05-2009; aceptado 5-06-2009

Detección de vocales mediante modelado de clusters de fonemas *

Vowel detection with phoneme cluster modelling

Iker Luengo
UPV/EHU
Alda. Urquijo s/n
Bilbao
iker.luengo@ehu.es

Eva Navas
UPV/EHU
Alda. Urquijo s/n
Bilbao
eva.navas@ehu.es

Jon Sánchez
UPV/EHU
Alda. Urquijo s/n
Bilbao
jon.sanchez@ehu.es

Inma Hernáez
UPV/EHU
Alda. Urquijo s/n
Bilbao
inma.hernaez@ehu.es

Resumen: La detección de regiones estables dentro de una señal de voz es necesaria en muchos sistemas de procesado del habla. Las vocales se corresponden precisamente con regiones de gran estabilidad, por lo que su detección automática puede ser muy conveniente. Este trabajo presenta un sistema de detección automática de vocales en la voz, mediante un identificador basado en modelos HMM de grupos fonéticos. Estos grupos fonéticos, creados según la similitud acústica de los fonemas, son la clave para el correcto funcionamiento del sistema en diferentes idiomas. Aunque los modelos han sido entrenados para euskera, las pruebas realizadas sobre bases de datos en euskera y en alemán demuestran que el sistema permite detectar las vocales y sus fronteras temporales con una precisión aceptable en ambos idiomas.
Palabras clave: segmentación, detección de vocales, estimación de ritmo

Abstract: Many speech signal processing systems require the detection of regions of stability within the signal. As vowels form great stability regions, a system capable of detecting them automatically in the speech is very convenient. This work presents such a system, which uses HMM models of phonetic clusters created according to the acoustic similarities among the phonemes. These clusters are the key element for the system to work correctly in different languages. Although models were trained in Basque, tests were carried out in both Basque and German speech databases showing that the system is able to detect the vowels and their boundaries with acceptable accuracy in both languages.
Keywords: segmentation, vowel detection, speech-rate estimation

1. Introducción

Muchos sistemas de procesado de voz requieren trabajar sobre segmentos de señal más o menos estables, de forma que las características de la misma se mantengan a lo largo del segmento. Esta estabilidad hace de esos segmentos el lugar más apropiado para poder realizar un análisis de la voz y sus características. De esta forma pueden obtenerse parámetros robustos, por ejemplo, para la caracterización de locutores (Mary y Yegnanarayana, 2008) o emociones (Ringeval y Chetouani, 2008).

Las vocales son un buen ejemplo de estas regiones. Suelen ser segmentos relativamente largos con respecto al resto de los fonemas, alcanzando hasta los 100 ms, con lo que proporcionan un extenso tiempo de análisis. Además, se trata de regiones bastante estables y con características acústicas muy definidas, a diferencia de las consonantes, mucho más cortas e inestables.

La mayoría de los sistemas de detección de vocales utilizan la curva de energía o la energía por bandas para localizar las regiones vocálicas. Puesto que las vocales contienen una gran energía, y además, ésta se concentra en las bandas bajas del espectro, esta aproximación ha sido muy utilizada. Por ejemplo, Pellegrino y Andre-Obrecht (1997) utilizan los picos de la curva SBEC (Spectral Band Energy Cumulating) para localizar las

* Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia dentro del proyecto AVIVAVOZ (TEC2006-13694-C03-02, www.avivavoz.es) y por el Gobierno Vasco en su subvención a grupos de investigación del sistema universitario vasco (IT-444-07).

vocales. Gracias a este sistema obtienen una precisión entre el 60 % y el 80 %, en función del idioma utilizado. Pfau y Ruske (1998) por su parte utilizan una aproximación similar, calculando la curva de volumen perceptual (loudness) modificada, cuyos picos también son utilizados para detectar las vocales con una precisión del 77 % sobre una base de datos en alemán.

Otro método muy utilizado es la FBD (forward-backward divergence) (Andre-Obrecht, 1988). Se trata de un mecanismo para la detección de cambios en las características de la señal, que aproximadamente coinciden con las fronteras entre fonemas. Para ello utiliza una ventana deslizante de tamaño fijo y otra fija que va incrementando su tamaño, de forma que cuando las características de los segmentos enventanados sean diferentes, se detecta una frontera. Una vez conocidas las fronteras, se puede utilizar un identificador de fonemas para detectar qué segmentos se corresponden con las vocales. Esta es la aproximación utilizada por Ringeval y Chetouani (2008).

Este artículo presenta un sistema para la detección de vocales en señales de voz, utilizando modelos ocultos de Markov (HMM) de grupos fonéticos entrenados para el euskera. Dichos grupos han sido creados de forma automática utilizando árboles de regresión, en función de la similitud acústica entre los fonemas. Es precisamente esta agrupación fonética la que dota al sistema de estabilidad y precisión. Para comprobarlo se han realizado pruebas de detección de vocales en dos bases de datos, una en euskera y otra en alemán, ambas con grabaciones que incluyen estilos de habla emocional.

El objetivo del sistema es detectar las regiones vocálicas estables que proporcionen zonas adecuadas para el análisis de señal. Las regiones detectadas como vocales pueden posteriormete ser utilizadas para extraer parámetros que caractericen la voz en sistemas de identificación de locutores, emociones o idioma. Estas regiones también permiten estimar el ritmo del habla si se calcula el número de vocales detectadas por unidad de tiempo o la longitud de estos segmentos.

La primera sección del artículo describe el sistema y su arquitectura, así como la metodología utilizada para realizar la agrupación de los fonemas. Posteriormente se describen las pruebas y medidas realizadas y las bases de datos utilizadas. Por último se presentan los resultados de estas pruebas y las conclusiones obtenidas.

2. Descripción del sistema

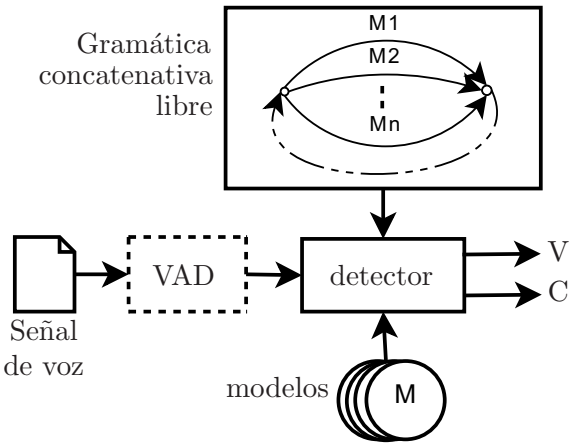

Figura 1: Esquema del detector de vocales propuesto

El esquema básico del sistema propuesto se presenta en la Figura 1. En su estructura más simple consta de un detector de fonemas basado en HMM de tres estados que trabaja con una gramática libre consistente en un bucle infinito de modelos concatenados. No se ha considerado utilizar una gramática más elaborada por limitaciones de tiempo en la realización del trabajo.

Para el entrenamiento de los modelos se ha utilizado la base de datos SpeechDat–EU (Hernáez et al., 2003), que contiene grabaciones en euskera de 1050 locutores realizadas a través de la red de telefonía fija y desde diversos entornos de ruido. Este entrenamiento se ha realizado utilizando las herramientas de HTK (Young et al., 2000) y siguiendo el sistema de reconocimiento de referencia RefRec (Lindberg et al., 2000). El número de componentes gaussianas de los modelos se ha decidido mediante pruebas empíricas de detección de vocales, estableciéndose en 1024 componentes. En cuanto a la parametrización, se han empleado 12 parámetros MFCC (Mel frequency cepstrum coeficients) calculados cada 10 milisegundos, junto con sus primeras y segundas diferencias.

Inicialmente se diseñó un sistema sencillo que constaba únicamente de un detector con un modelo diferente para cada uno de los sonidos del euskera. Posteriormente se añadió un módulo de detección de actividad vocal (VAD) para descartar los segmentos de

silencio. Por último se agruparon los fonemas según su similitud acústica con el fin de mejorar los resultados. En las siguientes secciones se describe cada uno de estos sistemas.

2.1. Primera aproximación

La primera aproximación del sistema consistía solamente en el detector de fonemas, sin utilizar el VAD ni la agrupación fonética. El sistema utilizaba un modelo diferente para cada fonema y el silencio. En las pruebas preliminares se comprobó que este sistema comete una gran cantidad de errores en la detección de los silencios, tanto por falsas inserciones como por falsas omisiones. Estos errores provocaban a su vez un incremento en los errores de detección de vocales, hasta dejar la precisión del sistema por debajo de un umbral aceptable.

2.2. Uso de detector de actividad vocal

Debido a los problemas en la detección de los silencios, se decidió añadir un sistema VAD basado en el propuesto por Ramirez et al. (2004) previo al detector, tal y como puede verse en la Figura 2 . El VAD permite separar los segmentos de voz y los de silencio con mayor precisión. De esta forma, al detector fonético sólo entran segmentos de voz, por lo que el modelo de silencio es innecesario. Para hacer que los modelos fueran consistentes con el nuevo método, se aplicó el VAD a la base de datos de entrenamiento y se reentrenaron todos los modelos fonéticos, esta vez sin incluir el silencio.

2.3. Agrupación de fonemas

Con una arquitectura tan sencilla y una gramática libre en forma de bucle infinito sin restricciones, la probabilidad de cometer errores en la secuencia de fonemas detectados y en la precisión temporal de las marcas creadas es muy grande. El sistema tiene demasiadas alternativas a la hora de decidir el fonema correspondiente en cada momento, con lo que aumenta la probabilidad de tomar una mala decisión, y con ello, el error en la detección de los fonemas y sus fronteras.

Para reducir la complejidad del sistema se ha llevado a cabo una agrupación de fonemas, de tal forma que se crea un único modelo para cada grupo. De esta forma el sistema no ha de decidir entre todos los fonemas existentes, sino sólo a qué grupo pertenece. Al reducir el número de alternativas, se espera que también se reduzca el error del sistema. Además, cada grupo de fonemas se entrena utilizando todos los ejemplos de los fonemas que agrupa, lo que quiere decir que cada modelo se entrena con más ejemplos, dando como resultado modelos más robustos.

Por supuesto, para que esta técnica funcione es necesario que los fonemas agrupados tengan características fonéticas similares. Para conseguirlo se ha decidido realizar un clustering ciego de los fonemas y utilizar los grupos resultantes.

Se ha creado un modelo de componentes Gaussianas (GMM) inicial para cada fonema, de una única gaussiana. Estos modelos se han entrenado utilizando las grabaciones de la base de datos SpeechDat–EU. Al igual que en el sistema final, se ha utilizado parametrización MFCC con primeras y segundas diferencias. Estos modelos permiten conocer las características acústicas de cada fonema de forma compacta, y han sido utilizados para realizar el clustering fonético mediante árboles de regresión. La Figura 2 muestra el dendrograma con el clustering resultante. El punto de poda del árbol está representado por la línea discontinua, y ha sido seleccionado teniendo en cuenta tanto la similitud acústica de los grupos como el número de ejemplos de entrenamiento disponibles para cada grupo, con el objetivo de asegurar poder entrenar modelos robustos. Cada grupo de fonemas está representado por un color diferente, excepto los grupos que constan de un único fonema, que están representados en negro.

Se puede observar que los grupos de fonemas resultantes se corresponden aproximadamente con los diferentes modos de articulación, de tal forma que se pueden nombrar como el grupo de fonemas fricativos y africados o el grupo de nasales. Las vocales forman cada una su propio grupo, por lo que cada vocal será modelada de forma independiente. Igualmente el fonema aproximante /D/ ha quedado formando su propio grupo separado del resto de aproximantes /G/ y /B/. El grupo denominado *L y similares* contiene fonemas que en principio no comparten modo de articulación (/L/ es líquida, /jj/ es fricativa y /gj/ es africada), pero sin embargo no es un grupo heterogéneo, ya que todos ellos tienen un sonido bastante similar hasta tal punto que hay personas que no distinguen entre ellas y las pronuncian igual. De ahí que

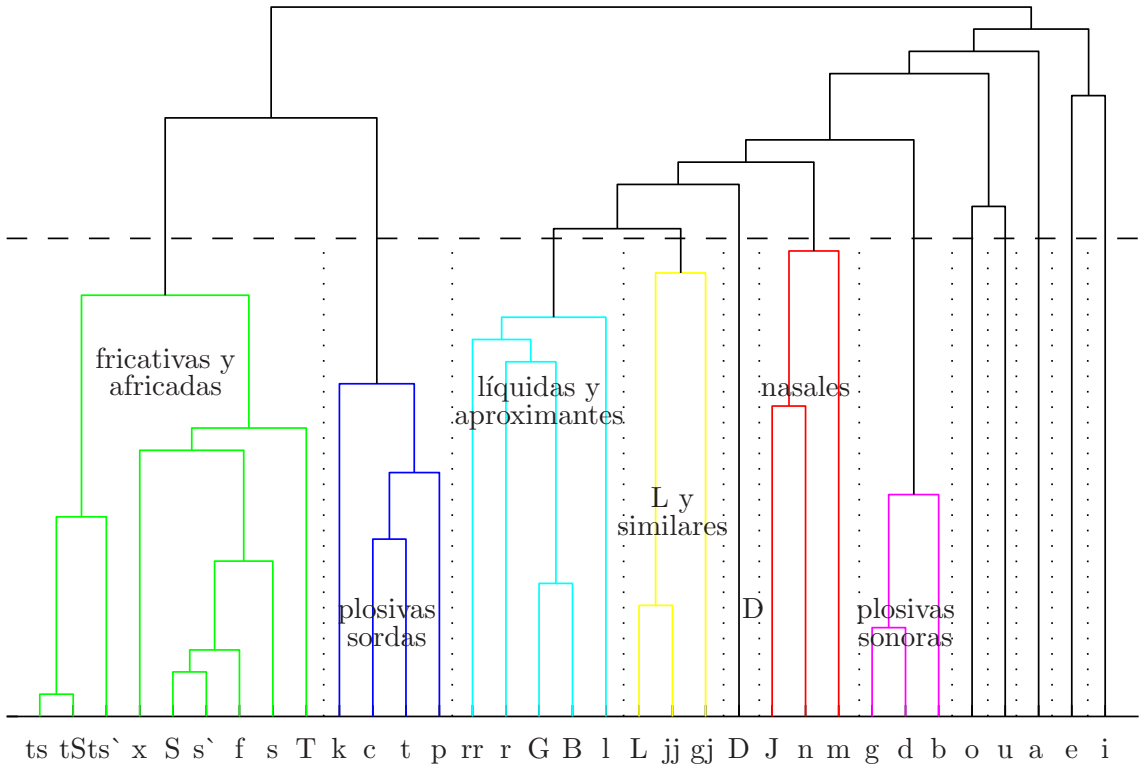

Figura 2: Dendrograma de la salida del clustering de fonemas

hayan quedado en el mismo grupo.

Una vez realizados los grupos de fonemas, se volvió a repetir el entrenamiento, en este caso creando un único modelo para cada grupo. Con este sistema se redujo significativamente el número de errores en la detección de vocales. En la sección 5 puede verse la comparación de resultados entre el sistema con y sin agrupación de modelos.

3. Medidas de precisión

Se han realizado varios experimentos y medidas para comprobar la eficacia del sistema propuesto. Para ello se han utilizado bases de datos para las que se dispone de un etiquetado fonético de referencia con el que comparar las marcas de vocal automáticas. Una vez aplicado el sistema a estas bases de datos, se han calculado los siguientes parámetros:

- La tasa de acierto, definida como el porcentaje de vocales correctamente detectadas sobre el número total de vocales en la base de datos
- La tasa de falsa inserción, definida como el porcentaje de falsas vocales detectadas sobre el número total de vocales detectadas.
- La tasa de falsa omisión, definida como el porcentaje de vocales no detectadas sobre el número total de vocales en la base de datos.
- La precisión del sistema, definida según la expresión:

$$Acc = \frac{N_{ref} - E_{ins} - E_{omi}}{N_{ref}} \times 100 \quad (1)$$

  Dónde N_{ref} es el número total de vocales en el etiquetado de referencia, E_{ins} es el número de errores de inserción y E_{omi} es el número de errores de omisión.
- La precisión temporal de la detección, expresada como el porcentaje de vocales que han sido etiquetadas con un error inferior a 20 ms con respecto a las marcas de referencia. En este caso sólo se consideran las vocales correctamente detectadas, y se ha calculado por separado para la marca de inicio de la vocal y para la de final.

La precisión temporal de las marcas creadas es un parámetro importante, puesto que

no se trata sólo de detectar la existencia de una vocal, sino también sus fronteras, de forma que sea posible utilizar el segmento vocálico para el cálculo de parámetros relativos a la voz. El valor aquí calculado (número de marcas con un error inferior a 20 ms) permite obtener resultados comparables a otros sistemas, pues se trata de la medida más comúnmente utilizada para esta precisión. Por ejemplo, podemos decir que los etiquetadores humanos llegan al 95 % de marcas con diferencias menores de 20 ms, mientras que en un sistema automático, un valor superior a 80 % se considera bueno (Hosom, 2000).

Puesto que el objetivo del sistema es realizar una detección de vocales y sus fronteras, y no identificar cuál es cada una de esas vocales, no se han distinguido las vocales entre sí a la hora de realizar esta comparación. Es decir, una /a/ detectada como /i/ no se considera un error, puesto que no deja de ser una vocal.

4. Bases de datos de pruebas

Se han utilizado dos bases de datos de diferentes características para realizar estas pruebas de precisión:

DB–Emozio: Consiste en una base de datos de habla expresiva en euskera, grabada por dos actores profesionales, un hombre y una mujer, que simulaban siete estados emocionales diferentes: Enfado, miedo, sorpresa, felicidad, tristeza, asco y neutro. El hecho de ser habla expresiva permitirá medir el comportamiento del algoritmo con diferentes tipos de voz. Esta base de datos contiene 702 frases por locutor y emoción. Las grabaciones fueron realizadas en una sala de grabación con un micrófono BeyerDynamic MC740. Saratxaga et al. (2006) proporciona la descripción completa de esta base de datos.

Berlin: Esta base de datos (Burkhardt et al., 2005) contiene grabaciones de habla expresiva en alemán, realizada por cinco actores y otras tantas actrices que simulaban siete estados emocionales: Enfado, miedo, aburrimiento, felicidad, tristeza, asco y neutro. Estas emociones son las mismas que las consideradas en *DB–Emozio*, excepto la sorpresa que se sustituye por aburrimiento. En total contiene 535 grabaciones microfónicas realizadas en una sala anecoica.

En este caso se pretende únicamente comprobar la eficacia del algoritmo con un idioma tan diferente al euskera como es el alemán, cuyo sistema vocálico es distinto. Mientras que el euskera contiene 5 vocales /a,e,i,o,u/, el alemán contiene 9 vocales cortas y 7 vocales largas, siendo la duración de las vocales largas aproximadamente el doble que el de las cortas. De estas vocales, 4 cortas y 2 largas no tienen correspondencia en el euskera, con lo que la detección de las mismas se complica. No es objetivo de este trabajo conseguir una herramienta multilingüe.

No se han realizado pruebas sobre la propia base de datos SpeechDat–EU, utilizada para el entrenamiento, debido a que dicha base de datos no tiene una segmentación revisada que poder utilizar de referencia en la comparación.

5. Resultados

La Tabla 1 resume los resultados de las pruebas de detección de vocales sobre las diferentes bases de datos utilizadas. Los diferentes campos de la tabla representan:

Total: El número total de vocales en la base de datos (según las transcripciones de referencia).

Aciertos: La tasa de aciertos respecto al total de vocales de la base de datos.

Omisiones: La tasa de omisiones respecto al total de vocales de la base de datos.

Inserciones: La tasa de inserciones respecto al total de vocales detectadas.

Precisión: La precisión total del sistema según la expresión (1).

Inicial <20ms : La tasa de vocales correctamente detectadas con un error del instante de inicio inferior a 20 ms.

Final <20ms : La tasa de vocales correctamente detectadas con un error del instante de final inferior a 20 ms.

Centrando la atención sobre los resultados de la base de datos en euskera *DB–Emozio*, se comprueba que el detector de vocales tiene una precisión global ligeramente superior al 85 %, con más del 90 % de las vocales existentes detectadas correctamente. Además estas vocales detectadas tienen una buena precisión temporal en las marcas de inicio y final, con más de un 80 % de las marcas con un error inferior a 20 ms.

Respecto a la base de datos en alemán *Berlin*, como cabe esperar, la precisión del

sistema cae hasta un 69 %, ya que en este caso a la dificultad de estilos se suma el problema de la diferencia de idioma. Recordemos que el alemán tiene un sistema de vocales mucho más complejo que le euskera. También se aprecia un incremento considerable de las vocales insertadas, en parte debido a la existencia de las vocales largas, que el sistema tiende a etiquetar como dos vocales seguidas.

	DB-Emo	**Berlin**
Total	271976	6454
Aciertos	90,30 %	92,21 %
Omisiones	7,79 %	9,70 %
Inserciones	6,58 %	18,89 %
Precisión	85,71 %	69,27 %
Inicial <20ms	88,73 %	78,34 %
Final <20ms	83,47 %	73,47 %

Tabla 1: Resultados de las pruebas de detección de vocales en cada una de las bases de datos.

Para analizar más en profundidad el efecto de los estilos en la detección de vocales, la Tabla 2 desglosa la precisión del sistema para cada estilo. En ambas bases de datos el estilo neutro es en el que se alcanza la máxima precisión, con gran diferencia respecto a las demás emociones. Este resultado es comprensible, puesto que la base de datos de entrenamiento del sistema contiene señales precisamente en ese estilo neutro. Por el contrario, los casos de miedo y sorpresa en *DB–Emozio* así como tristeza y asco en *Berlin* tienen las tasas de precisión más bajas.

Para comprobar la eficacia de la agrupación fonética utilizada, se han comparado estos resultados con dos niveles de agrupación extremos. Se han hecho pruebas con una agrupación nula, es decir, cuando cada fonema se modela por separado; y con una agrupación total, en donde sólo se consideran dos modelos, uno que agrupa todas las vocales y otro que modela todas las consonantes. Los resultados de estas pruebas pueden verse en las Tablas 3 y 4 respectivamente.

Cuando no se utiliza ninguna agrupación fonética se alcanza un 76,62 % de precisión en *DB–Emozio*, frente al 85,71 % al utilizar el agrupamiento. En el caso de *Berlin* los resultados son aún más significativos, ya que sin agrupación se llega a obtener un -38,46 % de precisión, debido a que los errores de inserción ascienden a 71,20 % y los de omisión a 73,87 %. La razón de tener estas tasas de error tan grandes es otra vez la existencia de fonemas en alemán que no tienen correspondencia en el euskera. Mientras que un sistema que modela grupos de fonemas similares puede ser capaz de detectar que un sonido pertenece a un grupo de fonemas, aunque nunca haya sido entrenado con muestras del mismo, un sistema basado en modelos independientes para cada fonema no tiene esa capacidad de generalización, y no es capaz de clasificar adecuadamente los sonidos no vistos durante el entrenamiento.

También se comprueba que para la base de datos en alemán, no sólo falla la detección de las vocales, sino que aquellas que han sido detectadas como tales tienen una gran imprecisión temporal. No así en la base de datos en euskera, donde aunque la precisión de la identificación disminuye, la precisión temporal se

Base de datos	**Precisión**
DB–Emozio	85,71 %
Enfado	*86,96 %*
Miedo	*81,94 %*
Sorpresa	*83,92 %*
Neutro	*88,92 %*
Felicidad	*85,65 %*
Tristeza	*86,57 %*
Asco	*86,00 %*
Berlin	69,27 %
Enfado	*66,11 %*
Miedo	*71,62 %*
Aburrimiento	*72,90 %*
Neutro	*80,50 %*
Felicidad	*67,72 %*
Tristeza	*62,09 %*
Asco	*61,39 %*

Tabla 2: Valores de precisión obtenidos para cada base de datos, según la expresión (1).

	DB-Emo	**Berlin**
Omisiones	18,32 %	73,87 %
Inserciones	5,84 %	71,20 %
Precisión	76,62 %	-38,46 %
Inicial <20ms	88,15 %	27,99 %
Final <20ms	80,96 %	27,63 %

Tabla 3: Resultados de las pruebas de detección de vocales sin la agrupación de fonemas.

	DB-Emo	Berlin
Omisiones	12,27 %	10,94 %
Inserciones	5,67 %	14,03 %
Precisión	82,46 %	74,53 %
Inicial <20ms	82,15 %	78,59 %
Final <20ms	75,63 %	72,92 %

Tabla 4: Resultados de las pruebas de detección de vocales con sólo un modelo de vocal y otro de consonante.

mantiene bastante estable.

En el caso de usar sólo un modelo de vocal y otro de consonante el sistema obtiene una precisión del 82,46 % en la base de datos *DB–Emozio*, lo que sigue siendo inferior al 85,71 % conseguido con la agrupación de fonemas por similitud acústica. En este caso la precisión temporal de las marcas es más baja, posiblemente porque al tratarse de modelos tan genéricos que agrupan fonemas de diferentes características el sistema tiene dificultades para distinguir cuándo acaba un fonema y empieza otro. Sin embargo esta generalidad beneficia al caso de la base de datos en alemán ya que al no tratarse de modelos tan específicos el sistema es capaz de detectar vocales no vistas durante el entrenamiento. De esta forma se alcanza un 74,53 % de precisión en *Berlin*, lo cual es incluso superior a lo obtenido con el sistema propuesto.

6. Conclusiones y trabajos futuros

En este artículo se ha presentado un sistema de detección y segmentación automática de vocales basado en modelos ocultos de Markov de grupos fonéticos. Estos segmentos pueden ser posteriormente utilizados para el análisis de las características de la voz, aprovechando que las vocales representan segmentos relativamente estables si se comparan con las características de los segmentos asociados a las consonantes. El sistema desarrollado es capaz de detectar las partes estables de las vocales, con cerca de un 85 % de las fronteras detectadas con un error inferior a 20 ms respecto a las fronteras reales. Hay que destacar que un error de 15 ms supone tan solo un desplazamiento de trama y media de análisis en las marcas, lo que no es muy importante para muchas aplicaciones. De hecho, este sistema y los segmentos detectados están siendo utilizados en la actualidad con éxito para el cálculo de parámetros prosódicos en sistemas de identificación de emociones en la voz y sistemas de verificación de locutor por nuestro grupo de trabajo.

Aunque el detector de vocales ha sido entrenado utilizando una base de datos en euskera en estilo neutro, se ha comprobado que la precisión del sistema se mantiene en valores aceptables para otros estilos de habla. También se ha querido comprobar el funcionamiento del método en otros idiomas que no sean el de entrenamiento, para lo que se han realizado experimentos con una base de datos en alemán. Si bien la precisión en este caso ha sido significativamente menor, ha dado resultados que pueden ser utilizados para realizar sistemas automáticos de análisis de voz, sobre todo si se restringen al estilo neutro. Teniendo en cuenta que las diferencias entre el sistema vocálico del euskera y del alemán son considerables, se estima que la precisión del sistema con otros idiomas más similares al de entrenamiento (por ejemplo, el castellano o catalán) puede ser bastante mayor.

Para alcanzar estos resultados ha sido necesario incluir un detector de actividad vocal que permita distinguir adecuadamente entre segmentos de voz y silencio, sin el cual el sistema resultaba poco fiable. También ha sido necesario crear los grupos fonéticos mediante un clustering ciego en función de las similitudes acústicas entre los fonemas. Utilizar modelos de grupos fonéticos en lugar de modelar los fonemas por separado o agrupar todas las vocales y todas las consonantes en sendos modelos ha sido clave para lograr un sistema estable y con una precisión aceptable.

Se ha comprobado mediante pruebas empíricas que utilizar un modelo separado para cada fonema dificulta el trabajo del detector debido a que el sistema comete mayores errores cuando se le presentan demasiadas opciones de clasificación. Además, utilizar modelos tan especializados supone grandes problemas en caso de que se presenten fonemas no considerados durante el entrenamiento, tal y como han certificado las pruebas sobre grabaciones en alemán. Por el contrario, utilizar un modelo de vocal y otro de consonante es una aproximación muy simplista que forma modelos muy heterogéneos y poco especializados, lo que también provoca un incremento de errores. Aunque este método ha demostrado ser muy robusto frente a la

aparición de fonemas no previstos, los resultados reflejan que el utilizar modelos de grupos fonéticos creados a partir de su similitud acústica es una compromiso que proporciona las mayores ventajas.

Una posible línea de continuación del trabajo sería utilizar una gramática más elaborada, teniendo en cuenta la estructura silábica del idioma. Esta gramática restringiría las posibilidades del sistema, favoreciendo la correcta detección de las vocales.

Por otro lado, la capacidad del sistema para generalizar a sonidos no utilizados en el entrenamiento es limitada. En vista de la pérdida de precisión al cambiar de idioma, se deduce que en el caso de querer aplicar el sistema para otro idioma sería conveniente realizar el entrenamiento de los modelos acústicos con grabaciones del idioma objetivo.

Bibliografía

Andre-Obrecht, Régine. 1988. a new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1):29–40, Jan.

Burkhardt, Felix, A. Paeschke, M. Rolfes, Walter F. Sendlmeier, y B. Weiss. 2005. A database of german emotional speech. En *Interspeech*, páginas 1517–1520, Lisbon. Portugal, Sep.

Hernáez, Inma, Iker Luengo, Eva Navas, Maren Zubizarreta, Iñaki Gaminde, y Jon Sánchez. 2003. The basque speech–dat (ii) database: A description and first test recognition results. En *Eurospeech*, páginas 1549–1552, Geneva, Sep.

Hosom, John Paul. 2000. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. Ph.D. tesis, Oregon Graduate Institute of Science and Technology.

Lindberg, Borge, Finn T. Johansen, Narada Warakagoda, Gunnar Lehtinen, Zdravko Kacic, Andrej Zgank, Kjell Elenius, y Gianpiero Salvi. 2000. A noise robust multilingual reference recogniser based on speechdat(ii). En *ICSLP*, volumen 3, páginas 370–373, Beijing, Oct.

Mary, Leena y B. Yegnanarayana. 2008. Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50(10):782–796, Oct.

Pellegrino, François y Régine Andre-Obrecht. 1997. From vocalic detection to automatic emergence of vowel systems. En *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'97)*, volumen 3, páginas 1651–1654, Apr.

Pfau, T y G Ruske. 1998. Estimating the speaking rate by vowel detection. En *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'98)*, páginas 945–948.

Ramirez, Javier, Jose C. Segura, Carmen Benitez, Angel de la Torre, y Antonio Rubio. 2004. Efficient voice activity detection algorithms using long term speech information. *Speech Communication*, 42:271–287, Apr.

Ringeval, Fabien y M. Chetouani. 2008. Exploiting a vowel based approach for acted emotion recognition. *Lecture Notes on Computer Science*, 5042:243–254.

Saratxaga, Ibon, Eva Navas, Inma Hernáez, y Iker Luengo. 2006. Designing and recording an emotional speech database for corpus based synthesis in basque. En *Language Resources and Evaluation Conference (LREC)*, páginas 2126–2129, Genoa, Italy, May.

Young, Steve, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, y Phil Woodland. 2000. *The HTK Book*. Cambridge University, Cambridge, Inglaterra.