ESTUDIO DEL USO DE INFORMACIÓN PROSÓDICA EN RECONOCIMIENTO DE LOCUTOR EN ÁMBITO FORENSE

Rubén Fernández Pozo[1], Carlos Fombella Mourelle[1], Doroteo Torre Toledano[2], Eduardo López Gonzalo[1] y Luis Hernández Gómez[1]

[1]Grupo de Aplicaciones de Procesado de Señales (Universidad Politécnica de Madrid),
[2]Área de Tratamiento de Voz y Señal (Universidad Autónoma de Madrid)
luis@gaps.ssr.upm.es

RESUMEN

En este trabajo se realiza un estudio de un conjunto de técnicas que hacen uso de la información prosódica aplicadas a la verificación de locutor dentro del campo de la Acústica Forense. Esta información ha sido aplicada siguiendo dos enfoques. Por un lado, se han implementado herramientas de modelado de lenguaje de la secuencia de segmentos obtenidos de la evolución de f0 y energía. Por otro lado, se han estudiado las posibilidades de mejorar un sistema básico de GMMs mediante una partición del espacio acústico fundamentada en la prominencia silábica. Estas técnicas se aplican sobre una base de datos de ámbito forense. Asimismo, se presentan los resultados obtenidos utilizando estas técnicas, y se muestra como se consigue una mejora de hasta un 10% en la tasa de error, al combinarlas con un sistema acústico de referencia.

1. INTRODUCCIÓN

En este articulo se presentan los resultados de una primera etapa de estudio de técnicas de utilización de la información de alto nivel que proporciona la evolución de las curvas de frecuencia fundamental y energía de la señal de voz, aplicadas al campo del reconocimiento de locutor en el ámbito forense. En este campo se tiene por objeto afrontar la problemática de la identificación del hablante presente en unas grabaciones ante un juez. Este escenario presenta una doble vertiente: Por un lado, disponer de una tecnología de identificación de locutor capaz de obtener tasas de error reducidas; y por otro, presentar de forma adecuada los resultados de la identificación ante el juez. Nuestro trabajo se enmarca dentro del primero de los enfoques presentados anteriormente, y tiene como objetivo principal el empleo de un conjunto de métodos recientes que hacen uso de la información de evolución de f0 y energía en una aplicación que se fundamenta en una base de datos especialmente desarrollada para acústica forense.

El estado del arte actual en los sistemas de identificación de locutor está representado por técnicas acústicas de bajo nivel, que hacen uso de información espectral en el dominio cepstral, representando los parámetros mediante GMMs, y adaptando los modelos mediante distintas técnicas a partir de un modelo universal [1]. Sin embargo, utilizando solamente este conjunto de técnicas, resulta difícil mejorar resultados actuales. En este sentido, los sistemas acústicos pueden verse favorecidos por la incorporación de técnicas de alto nivel, que aportan información incorrelada con la anterior, mejorando el rendimiento global del sistema. La información de alto nivel, además, presenta la ventaja de verse afectada en menor medida por el ruido y los efectos del canal. De este modo, al introducir elementos de nivel superior al acústico, se logra mayor robustez en el sistema, interesante también de cara a la aplicación planteada, pues las grabaciones en acústica forense no siempre se pueden realizar en condiciones idóneas [2].

En este trabajo se estudian las posibilidades para mejorar el rendimiento del sistema básico de GMMs a través de dos técnicas que hacen uso de la información prosódica. La primera de ellas se basa en los resultados presentados en [3], según los cuáles se hace uso de las trayectorias temporales de la f0 y la energía a través de una segmentación automática de la señal de voz en base a los máximos y mínimos de los parámetros citados, para caracterizar al locutor mediante un modelado de lenguaje de la secuencia de *tokens* obtenida. Con este enfoque se pretende utilizar esta información adicional de forma complementaria al sistema acústico de referencia a través de una fusión de ambos a nivel de resultados.

Por otra parte, proponemos también un nuevo esquema que utiliza la energía y frecuencia fundamental para identificar sílabas dentro de la señal de voz y asignar a cada una de ellas un valor de prominencia, es decir, un indicador de cómo se perciben estas unidades lingüísticas con respecto a las de su entorno. El objetivo de esta aproximación es analizar si es posible mejorar un sistema básico de GMMs modelando de forma controlada el espacio acústico en función de diferentes tipos de sílabas como se propone en [4], pero no clasificadas a través de su contenido fonético sino por su nivel de intensidad. Para ello se emplean técnicas de modelado acústico de forma similar a lo que se hacía en [4], pero utilizando en este caso información prosódica en vez de información fonética. Con esta metodología vamos a estudiar si la voz asociada a distintos niveles de prominencia contribuye también de manera diferente a la discriminación de voces en un sistema de verificación de locutor de la misma forma que lo hacían las distintas clases fonéticas [5].

Estas técnicas han sido aplicadas sobre la base de datos BDRA, muy adecuada para la tarea, dado que está constituida por grabaciones telefónicas de casos reales. Esta base de datos supone un escenario ideal para su uso en experimentos de identificación de locutor en Acústica Forense.

El artículo está estructurado de la siguiente forma: En la sección 2 se determinan las características de la base de datos empleada. En la sección 3 se describen las técnicas utilizadas y en la sección 4 se revisan los resultados obtenidos. Por último la sección 5 contiene las conclusiones y líneas futuras basadas en los resultados.

2. BDRA

Uno de los elementos fundamentales de este trabajo es la Base de Datos sobre la que se han realizado las pruebas. Los experimentos descritos en este artículo utilizan voz procedente de la base de datos B.D.R.A (Base de Datos de Registros Acústicos), integrada por voz extraída de casos forenses reales ya cerrados, y cuyo material grabado (intervenciones telefónicas, registros en contestadores, grabaciones ocultas, etc.) pasa a formar parte de la base de datos con fines de investigación mediante autorización judicial. Gracias al uso de esta base, aportada por la Guardia Civil dentro del Proyecto del Plan Nacional TIC2003-09068, las pruebas se han podido realizar en un ámbito muy similar al que podría darse en una situación real.

Esta base de datos contiene muchas horas de conversaciones que necesitan de un trabajo previo de selección y etiquetado para su uso en sistemas de identificación automática de locutor. Por ello, para las pruebas elaboradas en este trabajo se ha hecho una selección de 50 locutores diferentes, cada uno de ellos con al menos 13.5 minutos de conversación.

Las grabaciones utilizadas se han obtenido de conversaciones espontáneas desarrolladas tanto sobre líneas telefónicas fijas como sobre líneas GSM. Cada fichero contiene únicamente voz de un solo interlocutor (un único lado de conversación). Los datos de voz correspondientes al locutor en cuestión están concatenados directamente, de forma que en un solo fichero está toda la información del locutor. Por lo general, la voz está tomada de una única línea telefónica (no necesariamente de un único terminal), aunque hay casos en que se utilizaron dos líneas.

Para homogeneizar las pruebas se ha limitado la duración de todos los segmentos de voz de los distintos locutores a 13 minutos y medio (el mínimo disponible), desechando el resto. Asimismo, para dotar de mayor flexibilidad al estudio, cada segmento de voz se dividió en fragmentos de 30 segundos. De este modo, la base de pruebas final queda compuesta por 27 segmentos de 30 segundos para cada uno de los 50 locutores.

Para realizar el entrenamiento de los distintos sistemas se agruparon 10 de estos segmentos de 30 segundos, por un total de 5 minutos. Los ensayos se llevaron a cabo enfrentando segmentos de 30 segundos con los modelos, de modo que en total se dispuso de 16 *target tests* por locutor. Esto supone un total de 800 *target tests*. Para construir modelos universales se utilizó un fichero de 30 segundos de cada uno de los locutores (un segmento no utilizado ni como fichero de prueba ni como fichero de entrenamiento), para obtener un total de 25 minutos de entrenamiento. De este modo, se logra maximizar el número de ensayos y a la vez crear un UBM de cierta entidad, partiendo de los 50 locutores disponibles.

Por último, cabe señalar la dificultad que presenta esta base de datos para tareas de identificación de locutor, lo que va parejo precisamente a la buena caracterización que hace de una posible situación real de identificación. La voz ha sido capturada a través de distintos terminales, a lo que hay que añadir la variabilidad de estados de ánimo y el alto grado de espontaneidad que presentan los locutores en las grabaciones.

3. TÉCNICAS DESARROLLADAS

En este trabajo se han implementado dos conjuntos de técnicas diferentes, ambas basadas en la extracción de información derivada de la evolución de f0 y energía, pero con dos enfoques distintos.

En primer lugar se describe una técnica que utiliza modelos de lenguaje para caracterizar la variación temporal de los contornos de información prosódica bajo estudio, esto es, frecuencia fundamental y energía. El objetivo principal de este enfoque será utilizar esta información de alto nivel para mejorar el rendimiento del GMM básico.

Después, se detalla un sistema basado en los estudios realizados en [4], pero con una partición del espacio acústico fundamentada en la prominencia silábica. En este apartado se analizará también la relación de este tipo de modelado con el que utiliza información fonética. Por tanto, en este caso el objetivo es mejorar directamente el sistema de GMMs básico.

Finalmente, una vez determinado qué sistema acústico es el que presenta mejores resultados, se combinará con la información prosódica proveniente de los modelos de lenguaje con el objeto de mejorar el rendimiento global del sistema.

3.1. Sistema basado en modelado de lenguaje.

Este sistema utiliza modelos de lenguaje para representar las trayectorias cuantificadas de energía y f0, previamente segmentadas en base a los puntos de inflexión de ambas señales.

La segmentación se lleva a cabo sobre las señales de energía y f0, obtenidas utilizando el método *get_f0* incluido en la librería de tratamiento de voz *Snack* [6]. Las señales se filtran paso bajo y a continuación se extraen sus derivadas. Los puntos de inflexión quedan determinados por los cortes de estas derivadas con cero, siempre que en ese punto tengan una pendiente mayor que cierto umbral. Este umbral, de valor 0.1, fue obtenido de forma experimental conforme a un criterio que garantice segmentos no excesivamente cortos.

Una vez segmentadas las señales de energía y f0, se cuantifican ambas trayectorias, utilizando solamente dos niveles para cada una de ellas. De este modo se tienen segmentos crecientes o decrecientes tanto para energía como para f0. Esta información se combina en un solo flujo utilizando las fronteras de ambas señales de forma conjunta, del modo indicado en la figura 1.

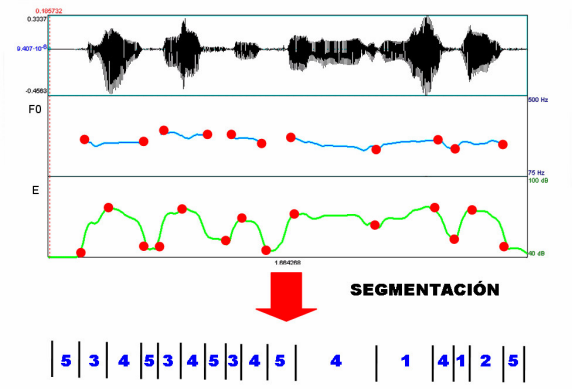

Figura 1. *Proceso de segmentación y cuantificación de las trayectorias de energía y f0.*

A cada segmento se le asigna un identificador de entre los 5 siguientes, en función de la combinación de pendientes de energía y f0.

Clase	Cuantificación de E y f0
1	F0 ascendente y Energía ascendente
2	F0 ascendente y Energía descendente
3	F0 descendente y Energía ascendente
4	F0 descendente y Energía descendente
5	Segmento no sonoro

Tabla 1. *Clasificación de segmentos*

Esta técnica se ha complementado ampliando las 5 clases anteriores con información sobre la duración del segmento (cuantificada en tres niveles: corto, medio o largo). De este modo se amplia el número de clases a 15.

Una vez cuantificados los segmentos, se procesa la secuencia resultante mediante modelos de lenguaje. Así, se crea un modelo de bigramas para cada locutor.

La puntuación utilizada para la evaluación del sistema se basa en la perplejidad de la secuencia de test frente al modelo de lenguaje. Esta puntuación se compara con la puntuación de la secuencia de test frente a un modelo de lenguaje universal, mediante una relación de verosimilitudes. Esta relación determina la puntuación final del segmento de test.

Esta técnica se ha fusionado a nivel de puntuaciones con un sistema fonético de reconocimiento similar al descrito en [7]. Éste emplea un transcriptor fonético que genera una secuencia de fonemas, modelados también a través de un sistema de modelos de lenguaje. En la sección 4 se presentan los resultados de la combinación de el sistema propuesto con el sistema acústico de referencia.

3.2. Sistema acústico por clases de sílabas según valor de prominencia

En este caso, intentaremos aprovechar la información de energía y frecuencia fundamental para segmentar el espacio acústico de características de un sistema de GMMs, de forma similar a la presentada en [4]. En este último trabajo, la señal de voz se divide en segmentos diferentes correspondientes a pseudo-sílabas con diferente estructura fonética. Posteriormente cada tipo de sílaba se modela utilizando un GMM específico de forma que el número total de gaussianas se distribuye uniformemente entre los GMMs de todas las clases fonéticas. En nuestro sistema la categorización se realizará a través de clases definidas por sílabas según el nivel de prominencia con el que se pronuncian. Con esto pretendemos estudiar si es posible conseguir un reparto más eficiente de las gaussianas que modelan toda la voz, controlando su distribución en funcion de los diferentes niveles de prominencia de las sílabas. En un sistema básico de GMMs no se puede asegurar de antemano un número determinado de gaussianas para cada clase, cosa que sí se puede garantizar con sistemas de GMMs categorizados por clases. Además, otro punto favorable para estos sistemas de clases, es que con este esquema de modelado se pueden eliminar clases que no contribuyen a diferenciar entre distintos locutores.

3.2.1 Segmentado en Sílabas

El primer paso para categorizar la voz según el nivel de prominencia consiste en el segmentado en sílabas de la señal de voz, para lo cual se han estudiado sistemas capaces de identificar automáticamente sílabas sin ayuda de ningún tipo de transcripción fonética previa. La implementación utilizada se basa en [8], sistema que trabaja con la envolvente de energía sobre la que buscan máximos y mínimos para localizar núcleos y límites de las sílabas. Este algoritmo es muy popular puesto que es particularmente sencillo, ya que trabaja directamente sobre la envolvente espectral sin la necesidad de utilizar ninguna herramienta estadística. El algoritmo implementado se puede resumir en los siguientes puntos [9].

1. Filtramos paso banda la señal de audio en el rango de 500 a 4000 Hz utilizando para ello un filtro de segundo orden de Butterworth.
2. Filtramos paso bajo el cuadrado de la señal resultante a 12 Hz obteniendo de esta forma la envolvente de energía. Para asegurar que no haya desfase se utilizará un filtro bidireccional.
3. Calculamos el *convex-hull* de la envolvente de energía.
4. Restamos la envolvente de energía al *convex-hull* de la envolvente. La señal resultante tiene picos que se corresponden a los valles de la envolvente.
5. Sujeto a ciertas condiciones resumidas abajo el máximo de la señal diferencia es seleccionado como límite de sílaba y el algoritmo (desde el paso 3) es ejecutado recursivamente sobre los subintervalos delimitados por el límite.
6. El algoritmo finaliza cuando no se encuentran límites que cumplan las condiciones dentro del subintervalo.

Las condiciones comentadas más arriba que hacen que el máximo de la señal diferencia sea considerado o no límite de sílaba se pueden ver a continuación:

- La máxima diferencia entre la envolvente de energía y el *convex-hull* debe ser mayor de 2 dB
- Los subintervalos a la izquierda y derecha del límite de sílaba deben ser mayores de 80 ms.
- La diferencia entre el pico de intensidad de cada subintervalo y el pico de intensidad de la señal entera no debe ser mayor de 25 dB

Todos estos parámetros han sido ajustados a las singulares condiciones de la BDRA. El algoritmo de Mermelstein generalmente funciona bien con sílabas claramente articuladas pero no detecta correctamente sílabas cortas no acentuadas. Las condiciones anteriores que impiden la inserción de nuevas sílabas son las responsables de este problema. Y es que la segmentación precisa de la señal de voz en sílabas es una tarea compleja y exigente, problema que se acentúa en mayor medida en bases de datos de ámbito forense como la BDRA.

3.2.2. Cuantificación de las sílabas por prominencia

Una vez dividida la señal de voz en sílabas, se procede a la cuantificación de los segmentos resultantes según su nivel de prominencia. Una sílaba se considera prominente si es percibida por encima de las de su entorno, lo que supone la aparición de al menos uno de los dos fenómenos prosódicos siguientes [10]:

- ***Pitch accent*** (acento de pitch): que está acústicamente relacionado con la evolución de la frecuencia fundamental y la energía total de la sílaba.

- ***Stress accent*** (acento tónico): Correlado con la duración del núcleo de la sílaba y la energía de la banda de 300-2200 Hz (*mid-to-high-frequency emphasis).*

A continuación examinaremos cada uno de los parámetros acústicos anteriormente citados y que nos permitirán cuantificar el nivel de prominencia de cada sílaba. Hay que resaltar aquí que todos ellos deben ser normalizados para evitar las variaciones naturales entre los diferentes locutores:

a. **Duración**: como comentamos anteriormente, la segmentación precisa de la señal de voz en sílabas es una tarea costosa que requiere de un esfuerzo minucioso al no poder automatizarse fácilmente. Es por ello, por lo que es necesario introducir una medida, que pueda ser automáticamente obtenida y que sea alternativa en la medición de la prominencia a la duración total de la sílaba. Este parámetro es la duración del núcleo de la sílaba que puede ser obtenida con facilidad mediante una variación del algoritmo de Mermelstein y que aporta la misma información que la duración de la sílaba en el cálculo de la prominencia.
b. **Energía de la banda de frecuencias media-alta**: como se puede ver en trabajos anteriores [11], la energía en la banda media del espectro (de 300 a 2200 Hz.) es un parámetro útil para determinar sílabas acentuadas.
c. **Energía**: en concreto el valor RMS de energía del núcleo de la sílaba.
d. **Contorno de f0**: siguiendo el modelo propuesto por Taylor, el contorno de f0 se convierte primero en un modelo RFC. La curva de f0 se divide en tramas de 0.02 seg. que son clasificadas, dependiendo de su gradiente y tras una interpolación lineal de los mismos, en subida (*rise*), bajada (*fall*) y contorno plano (*connection*). Contornos seguidos clasificados de la misma forma son fusionados en un solo intervalo y se mide la duración y amplitud de la sección de subida o bajada. Una vez obtenida esta representación RFC, es posible identificar en la curva eventos prosódicos tales como los acentos de *pitch* (una perfil de subida seguido de otro de bajada) que serán asignados al núcleo de la sílabas más cercana con un valor marcado por los parámetros del modelo TILT que están definidos a continuación:

$$A_{\text{event}} = |A_{\text{rise}}| + |A_{\text{fall}}|$$
$$D_{\text{event}} = |D_{\text{rise}}| + |D_{\text{fall}}|$$

donde A_{rise}, A_{fall}, D_{rise} y D_{fall} son respectivamente la amplitud y duración de los segmentos de subida y bajada del acento de *pitch*. En concreto se puede cuantificar la importancia del evento del acento del *pitch* multiplicando el evento amplitud (A_{event}) por su duración (D_{event}) y normalizar el producto por un factor de peso que exprese la relevancia del evento a lo largo de la locución (R_{event}). Este factor es calculado dividiendo el evento amplitud por la máxima variación de f0 y el máximo valor absoluto de f0 a lo largo de locución.

Como se muestra más arriba en esta misma sección, la prominencia puede ser definida desde un punto de vista teórico, como la combinación de dos eventos prosódicos (*pitch accent* y *stress accent*) que a su vez están relacionados cualitativamente con los parámetros acústicos vistos anteriormente. Es por esto, por lo que teniendo en cuenta estas relaciones, podemos construir una función continua de prominencia para cada una de la sílabas, directamente obtenida de los valores de los parámetros acústicos [10]:

$$Prom^i = \max \{ en^i_{300\text{-}2200} \cdot dur^i , en^i_{ov} \cdot (A^i_{\text{event}} \cdot D^i_{\text{event}} \cdot R^i_{\text{event}})\}$$

Este valor continuo se cuantificará en 4 niveles con una cuantificación de perfil exponencial con el fin de garantizar un reparto lo más uniforme posible de la señal de voz entre las clases de prominencia.

4. RESULTADOS

En esta sección se presentan los resultados obtenidos con la aplicación de las distintas técnicas desarrolladas en los puntos anteriores sobre la B.D.R.A.

Las pruebas llevadas a cabo se centran en la verificación de locutor, es decir, dada una frase dubitada, el sistema debe decidir si se corresponde con un modelo de locutor determinado o no.

En sistemas de ámbito forense se suelen utilizar como herramientas de evaluación las curvas Tippet [12]. Sin embargo, en nuestro estudio preliminar los resultados se compararán a través del *Equal Error Rate* (EER), o Tasa de Error Equivalente. Este indicador representa el error correspondiente al umbral que hace que los dos tipos de errores posibles en un sistema de verificación de locutor, Falsa Aceptación y Falso Rechazo, sean iguales. A menor EER, mejor es el rendimiento del sistema. Asimismo se proporcionan las curvas DET (*detection error tradeoff*), en cuyo punto central se sitúa el citado EER.

4.1. Sistema acústico básico de referencia.

El sistema básico de referencia que se ha empleado es un GMM de 256 gaussianas. Los modelos de locutor se han adaptado a partir de un modelo universal. Este sistema produce un EER de 4.5 %.

4.2. Sistema basado en modelado de lenguaje.

En la figura 2 se presentan los resultados correspondientes al sistema prosódico basado en trayectorias de f0 y energía. Las cuatro curvas representan el sistema prosódico básico, el sistema prosódico con adición de duraciones, el sistema fonético de referencia y la fusión de los dos últimos. La fusión se ha realizado a nivel de puntuaciones.

Los valores de EER figuran a continuación:

Sistema	EER
f0_e0	18.38 %
F0_e0_duración	15.75 %
Fonético	12.38 %
Fusión fonético + f0_e0_dur	8,37%

Tabla 2. *EER para el sistema basado en el modelado de lenguaje de segmentos prosódicos*

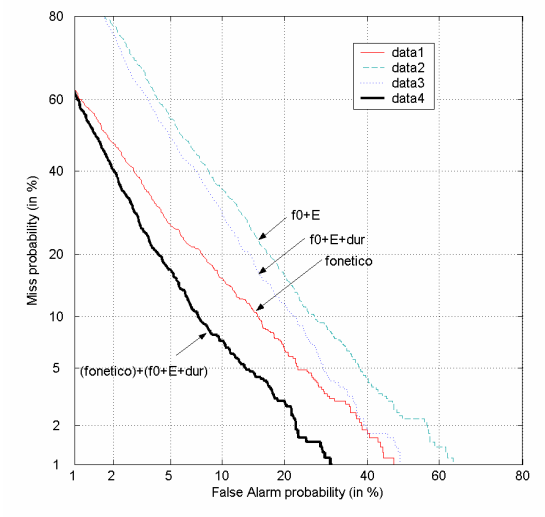

Figura 2. *Resultados para el sistema basado en el modelado de lenguaje de segmentos prosódicos*

A la vista de los resultados se puede concluir que la adición de nueva información prosódica a un sistema fonético de modelos de lenguaje supone una mejora en el rendimiento del sistema.

4.3. Sistema acústico por clases de sílabas según el valor de prominencia.

A continuación se detallan los resultados derivados del sistema acústico-lingüístico descrito en el apartado 3.2. La clasificación de GMMs por prominencias produce los siguientes resultados:

Clase	EER
Prominencia 1	7.25 %
Prominencia 2	7 %
Prominencia 3	7 %
Prominencia 4	6.125 %
Fusión de las 4 Clases	**5.875 %**

Tabla 3. *EER para clases de prominencias.*

Estos resultados muestran que las sílabas prominentes son las que poseen más capacidad discriminativa. No obstante, se puede observar que ninguna de las clases aisladas ni la fusión de ellas mejora el sistema acústico de referencia. La posible causa de este comportamiento puede ser que cada una de las clases resultantes tiene un contenido fonético heterogéneo. Esto hace que el espacio acústico de cada clase sea similar al de un GMM básico pero modelado con un menor número de gaussianas.

Para comprobar que una distribución adecuada del espacio acústico mejora el rendimiento del sistema básico de GMMs, se proporcionan además resultados (Tabla 4) obtenidos de un sistema en el que la partición del espacio acústico se obtiene a partir de una segmentación de la señal de voz en clases fonéticas, de modo similar al descrito en [5]:

Clase	EER
Clase 1 (vocales)	5.75 %
Clase 2 (nasales + vibrantes)	7.375 %
Clase 3 (fricativas + líquidas)	6.25 %
Clase 4 (oclusivas)	5.625 %
Fusión de las 4 Clases	**5.75 %**

Tabla 4. *EER para clases fonéticas*

Estos resultados nos muestran que este sistema tampoco mejora el GMM básico. La posible causa de estos resultados puede ser que la transcripción fonética no es precisa con las deficientes condiciones de articulación del habla de los locutores que se dan en la BDRA. Esto no suponía un problema con el sistema fonético descrito en la sección 3.1 puesto que son precisamente los errores en la transcripción los que contribuyen en mayor medida al proceso de reconocimiento de locutor [13]. Sin embargo, en nuestro sistema, para asegurar una adecuada distribución del espacio acústico sí es necesaria una segmentación precisa.

Por otro lado, puede resultar llamativo que las nasales aporten el peor de los resultados de entre todas las clases, cuando son junto a las vocales las que tienen mayor poder discriminativo [14]. Analizando el resultado de la transcripción fonética observamos que hay una carencia de datos correspondientes a la clase 2, lo que justifica su peor funcionamiento.

Se presentan además resultados de nuevas propuestas de fusión, mezclando las clases fonéticas y clases de prominencia más discriminativas, de modo que se construya un GMM que aproveche de forma más eficiente las gaussianas disponibles

Sistema	EER
Fus. C.Foneticas + C.Prominencia	5.375 %
Fus. CF1 + CF4 + CP3 + CP4	5.125 %

Tabla 5. *EER para la fusión de sistemas.*

En primer lugar, figura la fusión de las cuatro clases fonéticas y las cuatro clases de prominencia. En segundo lugar se presenta el EER de la fusión de las dos mejores clases fonéticas y las dos mejores clases de prominencia. Se puede ver como se mejora la tasa de error equivalente respecto a los sistemas aislados, aunque no mejora en ningún caso el rendimiento del GMM básico.

4.4. Fusión del sistema acústico básico con el sistema prosódico de modelos de lenguaje.

A partir de los resultados de los apartados anteriores, el modelado acústico que ha proporcionado mejores resultados es el uso de un GMM único. Por lo tanto se ha estudiado la fusión de este reconocedor con el de información prosódica y fonética (Sección 3.1). Así, fusionando a nivel de puntuaciones el sistema acústico con el sistema de alto nivel de modelos de lenguaje para secuencias de fonemas y *tokens* prosódicos, se obtiene un ERR del 4.05 % que supone una mejora del 10%. Esta mejora, demuestra que la inclusión de datos de alto nivel aporta información incorrelada al sistema acústico básico.

5. CONCLUSIONES Y LINEAS FUTURAS

A lo largo del artículo se ha podido ver como una misma información, en este caso, la energía y la frecuencia fundamental, se puede estudiar y emplear de dos formas muy distintas, con el objetivo de mejorar el rendimiento del sistema acústico básico. Por una parte, se ha implementado un sistema de modelos de lenguaje sobre los contornos de f0 y energía que, junto con un sistema fonético, produce una mejora de un 10% respecto al GMM básico.

Por otro lado, se ha propuesto un nuevo esquema en el uso de la información de f0 y energía, basado en [4], que permite segmentar el espacio acústico mediante la prominencia de las sílabas presentes en la locución. Sin embargo, se ha comprobado que los resultados de este sistema no superan al sistema básico de referencia. Por otro lado, el sistema basado en la segmentación fonética del espacio acústico, a pesar de resultar efectivo en [4], tampoco supone una mejora respecto al sistema de referencia sobre la base de datos utilizada. Una explicación para este comportamiento podría estar en la dificultad de realizar de forma precisa una segmentación fonética de la base de datos. Debido a las condiciones que presenta la BDRA, se pueden producir clases poco uniformes que incluyan segmentos pertenecientes a otras clases. No obstante, el marco presentado permite desarrollar líneas futuras de trabajo centradas en una mejor distribución del espacio acústico.

6. AGRADECIMIENTOS

Este trabajo ha sido posible gracias a la subvención aportada por el Proyecto TIC2003-09068-C02-02 del Plan Nacional de I+D.

7. REFERENCIAS

[1] D.A. Reynolds, T.F.Quatieri, y R.B. Dunn "Speaker verification using adapted gaussian mixture models", *Digital Signal Processing*, 10 (1-3), pp 19-41, 2000.

[2] SuperSID. Supersid: "Exploiting high-level information for high-performance speaker recognition", http://www.clsp.jhu.edu/ws2002/groups/supersid/. 2002.

[3] A. Adami, R. Mihaescu, D.A. Reynolds, y J. Godfrey, "Modelling prosodic dynamics for speaker Recognition", en *International Conference on Acoustics, Speech and Signal Processing, ICASSP* 2003.

[4] B. Baker, R. Vogt, y S. Sridharan, "Gaussian Mixture-Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification", en *International Conference on Speech and Language Processing* 2005

[5] M. Hébert, y L.P. Heck, "Phonetic class-based speaker verification", en *INTERSPEECH* 2003.

[6] K. Sjölander, "The Snack Sound Toolkit v.2.2.8", http://www.speech.kth.se/snack/ 2004

[7] W.D. Andrews, M.A. Kholer, y J.P. Campbell, "Phonetic speaker recognition" en *INTERSPEECH* 2001

[8] P. Mermelstein, "Automatic segmentation of speech into syllabic units", *Journal Acoustical Society of America*, 58 (4), pp 880-3, Octubre 1975

[9] R. Villing, J. Timoney, T. Ward, y J. Costello, "Automatic Blind Syllable Segmentation for Continuous Speech", en *Irish Signal and System Conference* Julio 2004

[10] F. Tamburini, y C. Caini, "An automatic system for detecting prosodic prominence in American English continuous speech", *International Journal of Speech Technology*, 8, pp 33-44, 2005

[11] A. Sluijter, y V. van Heuven, "Acoustic correlates of linguistic stress and accent in Dutch and American English", en *International Conference on Speech and Language Processing* 1996.

[12] J. Gonzalez, J. Fierrez, y J. Ortega, "Forensic identification reporting using automatic speaker recognition systems", en *International Conference on Acoustics, Speech and Signal Processing, ICASSP,* Abril 2003.

[13] D. Torre, C. Fombella, J. González, y L. Hernández, "On the Relationship between Phonetic Modeling Precision and Phonetic Speaker Recognition Accuracy", en *INTERSPEECH* 2005

[14] M. Antal, y G. Toderean, "Broad Phonetic Classes Expressing Speaker Individuality", *Studia Informatica*, Vol. LI, No. 1, pp. 49-58, 2006.