RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos no supervisados de segmentación acústica, su aplicación en la identificación del idioma hablado

Ana Montalvo Bereau y
José R. Calvo de Lara

RT_062 **julio 2014**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos no supervisados de segmentación acústica, su aplicación en la identificación del idioma hablado

Ana Montalvo Bereau y
José R. Calvo de Lara

RT_062 **julio 2014**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

1 Introducción 1
2 Métodos de segmentación acústica 2
2.1 Segmentación basada en el retraso de grupo 4
2.2 Segmentación basada en un filtro Gaussiano 5
2.3 Segmentación basada en una medida de transición o salto 6
2.4 Segmentación basada en el algoritmo de agrupamiento de Máximo Margen 7
3 Segmentación Gaussiana 7
3.1 Segmentación sobre el dominio cepstral 8
3.2 Representación sobre el dominio Gaussiano 10
3.3 Experimentos y discusión 10
4 Conclusiones 11
Referencias 14

Anexos 15
1 Modelación estadística del idioma 15
2 Definición del cepstrum 15

Métodos no supervisados de segmentación acústica, su aplicación en la identificación del idioma hablado

Ana Montalvo Bereau y José R. Calvo de Lara

Equipo de Investigaciones de Imágenes y Señales, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), La Habana, Cuba
{amontalvo, jcalvo}@cenatav.co.cu

RT_062, Serie Azul, CENATAV
Aceptado: 17 de julio de 2014

Resumen. Por segmentación acústica entendamos dividir las locuciones en segmentos de tiempo. El siguiente reporte constituye un estudio de los más recientes métodos de segmentación acústica no supervisada, de la dependencia de los mismos con el posterior uso de los segmentos y de las formas de evaluación de su eficacia. También se exponen las propuestas de los autores para abordar dicha temática dirigida a la identificación del idioma hablado. Se reportan los resultados, que aunque no positivos permitieron llegar a conclusiones.

Palabras clave: segmentación acústica, segmentación no supervisada del habla.

Abstract. By acoustic segmentation understand: utterances divided into time segments. The following report is a study of the latest methods of unsupervised acoustic segmentation, of their dependence with the subsequent use of the segments and of the methods for evaluating their effectiveness. Proposals of the authors are also discussed to address this issue particularly directed to the spoken language identification. Results are reported, which although not positives, permitted to reach conclusions.

Keywords: acoustic segmentation, unsupervised speech segmentation.

1 Introducción

Es conocido que el cerebro humano, la más perfecta computadora respecto a tareas cognitivas, segmenta la señal de habla antes de reconocer sonidos o palabras [1]. Las células pulpo [1], en el núcleo coclear, detectan los inicios de voz [2] mediante la activación simultánea de las fibras del nervio auditivo [3], y el reconocimiento de fonemas[2] o palabras no tiene lugar hasta que la señal no llega a la corteza auditiva. La segmentación es una de las primeras etapas en el proceso de percepción, de ahí su importancia estratégica para las tareas relacionadas con el procesamiento automático de voz hablada.

Resulta igualmente interesante el hecho demostrado en experimentos de percepción, de que el ser humano luego de determinado tiempo expuesto a un idioma del cual no tiene ningún conocimiento lingüístico, es capaz de reconocerlo de otros igualmente desconocidos para él [5]. O sea, se percata de cierta

[1] Estas células (*octopus cells*) se encuentran en la parte posterior del núcleo coclear ventral, producen una respuesta a estímulos tonales y se consideran fundamentales en la extracción de información temporal.

[2] El aparato fonador humano es capaz de producir un amplia variedad de sonidos. Los sonidos del habla, como eventos acústicos concretos, son llamados fonos; mientras que vistos como entidades de un sistema lingüístico, son referidos como fonemas[4]

información contenida en el habla y no precisamente en las palabras, que le permite identificar el idioma escuchado.

Entonces, ¿cuán necesarios son realmente los fonemas, o cualquier otra unidad lingüística definida, para modelar e identificar un idioma?

Los enfoques que abordan la segmentación acústica pueden ser divididos en dos clases. La primera requiere información lingüística y modelos acústicos de los fonemas o entidad elegida. Esta segmentación es comúnmente seguida de una alineación de dichos segmentos al texto dado y uno de los métodos más empleados de esta clase es la alineación forzada, basada en los Modelos Ocultos de Markov (HMM) [6]. Por su parte, los métodos de la otra clase intentan realizar la segmentación sin ningún conocimiento previo, lo que es conocido como segmentación no supervisada, el enfoque explotado por nosotros pertenece a esta segunda clase. La mayoría de los acercamientos a estos métodos se centran en la detección de puntos de cambio en la señal de habla y se toman estos puntos como frontera.

Nuestro dominio de acción radica por tanto en la segmentación no supervisada de la señal de habla, en unidades acústicas como alternativa a los fonemas, para el reconocimiento del idioma hablado. En particular el problema será enfocado para poca disponibilidad de datos, lo cual implica una limitación práctica grande.

El reconocimiento del idioma hablado es un problema atacado fundamentalmente con dos enfoques: uno explota la información acústica de los rasgos con que se describe la señal (enfoque acústico) y el otro la estadística de ocurrencia de los fonemas en cada idioma [5] (enfoque fonotáctico). Ambos se complementan en términos de información.

Se supone que las características globales de sonido de todas las lenguas habladas pueden ser cubiertos por una amplia colección de unidades acústicas, que puede caracterizarse por segmentos acústicos [7]. La segmentación presentada en el presente reporte es seguida de un enfoque estadístico del problema, explotando la información que brinda la co-ocurrencia de los *tokens*[3] e imponiendo a priori que los mismos sean una unidad lingüística distinta de los fonemas.

Este reporte está estructurado de la siguiente forma: en la Sección 2 se presenta un estudio crítico y explicativo de los principales y más actuales métodos de segmentación acústica no supervisada. La Sección 3 muestra los métodos de segmentación acústica basados en Gaussianas, desarrollados y explorados experimentalmente por los autores. Finalmente en Sección 4 se presentan las conclusiones del reporte y las futuras líneas de trabajo.

2 Métodos de segmentación acústica

La segmentación automática de señales de habla continúa siendo un reto para la comunidad de investigadores del tema. Incluso luego del avance en técnicas supervisadas y no supervisadas, persiste el reto de igualar la segmentación manual que es capaz de realizar un ser humano. Las técnicas de segmentación basadas en HMM con correcciones y modificaciones, han sido el estado del arte, sin embargo las mismas demandan grandes volúmenes de señal etiquetada o transcrita fonéticamente. Las técnicas no supervisadas, por otro lado, exploran cambios o variaciones en propiedades espectrales y temporales de la señal de habla, lo que se conoce como información de bajo nivel.

La segmentación de datos de audio se ha convertido en un procedimiento muy importante en los sistemas que procesan audio. Es especialmente significativo en aplicaciones tales como reconocimiento

[3] Unidad genérica que pudiera ser trifonema, sílaba, fonema, fono, índice, PLUs (del inglés *phoneme like units*)[8],[9].

automático de habla (ASR[4]), donde es preciso analizar solo los intervalos de habla y separar del análisis los de no habla [10].

La segmentación a priori de audio es muy importante también para aplicaciones de transcripción de noticias [11], donde típicamente el habla está intercalada con música y ruido de fondo. Con el auge de internet, el indexado de acuerdo al contenido ha emergido con fuerza como tarea a desarrollar porque hay gran cantidad de audio en la web que no está indexado en los motores de búsqueda. La segmentación también es usada en sistemas para la diarización de audio y locutores, y cada vez con más fuerza para separar habla de música.

Los métodos de segmentación acústica dependen fuertemente del objetivo para el cual se diseñen. Un método de segmentación será más efectivo que otro, siempre que con sus unidades acústicas se logre un desempeño mejor del sistema como un todo. No obstante también hay formas de medir la eficacia de la segmentación cuando se conoce qué es lo que se quiere delimitar y se dispone del número de segmentos existentes para evaluar. Por ejemplo, en [12] se presentan los resultados usando la razón de detección correcta α definida por:

$$\alpha(\%) = \frac{N_{detectadas}}{N_{existentes}} * 100, \tag{1}$$

donde $N_{detectadas}$ se refiere al número total de fronteras detectadas para un determinado nivel de tolerancia y $N_{existentes}$ es el número real respaldado por la base de datos. El nivel de tolerancia determina el intervalo acordado para asumir como correcta una marca, su valor es resultado de una calibración a prueba y error, y generalmente se expresa en milisegundos.

Otro parámetro empleado es la razón de sobre-segmentación β que se calcula:

$$\beta(\%) = \left(\frac{N_{detectadas}}{N_{existentes}} - 1\right) * 100. \tag{2}$$

Estos dos métodos de evaluación son aplicables solo en casos en los que se disponga de una cantidad, estadísticamente significativa, de señales con las fronteras marcadas, lo cual resulta una desventaja en situaciones desprovistas de etiquetado. De igual forma no son funcionales estas métricas cuando la unidad acústica es abstracta y no está asociada directamente con una representación fonética o lingüística, como es nuestro caso.

Por su parte la segmentación para sintetizar habla es mucho más efectiva cuando las unidades en las que se segmenta, tienen información de contextos mayores que difonos (segmentos de señal de más de dos sonidos). Mientras más información dinámica tenga la unidad, mayor capacidad de modelar la co-articulación tendrá, por tanto más cercano a lo real será la voz sintetizada [13].

La tarea de detección de eventos acústicos es también muy dependiente de la calidad con que se segmente la señal, y hay trabajos que proponen una segmentación no supervisada[14] y ciertamente arbitraria, en intervalos de 5 segundos. Sobre dichos segmentos extraen los *i-vectors*[5] y entrenan una Máquina de Vectores Soporte (SVM) con ellos, para determinar si el evento acústico buscado está en dicho segmento o no.

Actualmente la dicotomía del "huevo o la gallina" existe entre la segmentación automática de habla y el reconocimiento de habla. Por un lado los sistemas de reconocimiento de habla requieren transcripciones de entrenamiento bien delimitadas para poder reconocer la secuencia de fonos y por el otro, los sistemas de segmentación demandan un conocimiento preciso de la secuencia de fonemas para poder segmentar en sus componentes fonéticas.

[4] *Automatic Speech Recognition*

[5] Representación vectorial muy reciente y de probada eficacia en el procesamiento de voz [15].

Es así como se presenta la disyuntiva a la hora de resolver esta tarea, por lo que sigue siendo un reto importante segmentar la señal para reconocimiento de habla, de forma automática y no supervisada.

De forma general, los métodos de segmentación acústica pueden clasificarse básicamente en enfoques supervisados y no supervisados. Para los supervisados el conjunto de entrenamiento tiene que estar transcrito en términos de unidades sonoras del habla (fonemas, sílabas, palabras). En el caso de los métodos no supervisados, no demandan etiquetado o transcripciones fonéticas y se basan en la comprensión y reproducción del proceso humano de aprendizaje. Estos métodos utilizan algoritmos para colectar picos que marquen fronteras acústico-fonéticas y posteriormente agrupan los segmentos resultantes para modelarlos. Entre los más empleados se encuentran los basados en la función de retraso de grupo (GDF), en algoritmos de agrupación de margen máximo (MMC) y en discontinuidades en la serie temporal (*jump functions*).

A continuación exponemos algunos de los más importantes métodos de segmentación acústica.

2.1 Segmentación basada en el retraso de grupo

De acuerdo al modelo Fuente/Filtro, la voz es el resultado del paso de una excitación glotal a través de un filtro lineal invariante en el tiempo (LTI), responsable de modelar las características de resonancia del tracto vocal (3). La voz por tanto ($s(t)$) puede ser considerada como la convolución de la respuesta al impulso del filtro LTI ($h(t)$) con la excitación ($e(t)$):

$$s(t) = \int_0^t h(t-\tau)e(\tau)d\tau. \tag{3}$$

Muchos intentos han habido por mejorar el modelado de la señal fuente (excitación) en el contexto lineal, sin embargo la representación más compacta y flexible es la referida a los modelos sinusoidales [16]. En los modelos sinusoidales, la señal de excitación es representada con una suma de sinusoides:

$$e(t) = \sum_{k=0}^{K(t)} a_k(t)e^{i\Phi_k(t)}, \tag{4}$$

donde $a_k(t)$ y $\Phi_k(t)$ son la amplitud de la excitación y la fase de la k-ésima sinusoide respectivamente, y $K(t)$ es el número de sinusoides que podría variar en el tiempo. Sin perder generalidad, convenientemente se asume que la amplitud de la excitación es constante en el tiempo e igual a la unidad: $a_k(t) = 1$.

Por su lado la función de transferencia del filtro que modela el tracto vocal viene dada por:

$$H(\omega,t) = G(\omega,t)e^{i\Psi[(\omega,t)]}; \tag{5}$$

con $G(\omega,t)$ y $\Psi(\omega,t)$ referidas a la amplitud y la fase del sistema respectivamente.

De lo anterior obtenemos el siguiente modelo del habla:

$$s(t) = \sum_{k=0}^{K(t)} G[\omega_k(t)]e^{i\{\Phi_k(t)+\Psi[\omega_k(t)]\}}, \tag{6}$$

$$s(t) = \sum_{k=0}^{K(t)} A_k(t)e^{i\theta_k(\omega,t)}, \tag{7}$$

siendo $A_k(t)$ la magnitud y $e^{i\theta_k(\omega,t)}$ el espectro de fase de $s(t)$.

El retardo de grupo es una forma de medir la no linealidad de la fase de un sistema. En un sistema en el que la fase es lineal, todas las frecuencias experimentan el mismo retardo al atravesarlo. Una fase no

lineal lleva a una dispersión en tiempo de las diferentes frecuencias, lo que se conoce como distorsión de fase. En los sistemas no lineales, como es el caso del sistema de producción del habla, este concepto se conoce por retardo de grupo.

El retardo de grupo es el desplazamiento de una banda o grupo de frecuencias determinada y se define como la pendiente de la fase de dicha frecuencia, es decir, como la variación negativa de la fase.

La función de retraso de grupo es utilizada para la segmentación silábica [17]. Este método considera la energía a corto término de la señal de fase mínima y el retraso de grupo de dicha energía, para determinar los límites de las sílabas.

Ha sido observado que en el espectro de retraso de grupo de la energía a corto término, para los sistemas de fase mínima [18], tanto los picos como los valles son resueltos correctamente (donde los picos corresponden a polos y los valles corresponden a ceros)[19]. La ubicación de los picos en la función de retraso de grupo, se corresponde con las fronteras de las sílabas.

La metodología empleada es la siguiente:

- se obtiene la energía a corto término de la señal (STE[6]) $s(n)$, sobre una ventana pequeña de N muestras,

$$STE(n) = \sum_{l=n-N+1}^{n} |s(l)|^2, \tag{8}$$

  donde $STE(n)$ representa la energía a corto término de la muestra n,
- se calcula el retraso de grupo de la señal STE

$$\tau(\omega) = -\frac{d}{d\omega} H(e^{j\omega}), \tag{9}$$

  siendo $H(e^{j\omega})$ la respuesta de frecuencia del sistema o señal analizada.

Los máximos de la función de retraso de grupo son considerados como frontera de sílaba.

Este método de segmentación es vulnerable frente a segmentos fricativos, ya que la energía de los mismos es bien alta. Con los nasales y trinos (*trills*), sucede que son sonidos altamente transitorios[7] y el algoritmo pierde esas fronteras.

2.2 Segmentación basada en un filtro Gaussiano

Recientemente ha sido demostrado [20] que si la STE de una señal es filtrada con un filtro Gaussiano paso-bajo, los mínimos de la señal filtrada se corresponden con los límites de las sílabas. Es sabido que el filtro Gaussiano funciona muy bien como filtro paso-bajo para señales unidimensionales como la voz y que tiene una excelente resolución tiempo-frecuencia porque la transformada de Fourier de una Gaussiana es también Gaussiana [21].

Entonces para realizar esta segmentación basta con calcular la STE de la señal y filtrarla con un filtro cuya respuesta al impulso sea $g(t) = e^{-a^2t^2}$, con a constante. Finalmente donde se localicen los mínimos se asumirá que existe una frontera silábica.

Las facilidades para la implementación de este método, sus probados resultados y su independencia del etiquetado, lo convierten en un candidato fuerte para fusionar con otros y mejorar el desempeño, sobre todo compensando métodos como la función de retraso de grupo.

[6] *Short Term Energy*

[7] Un transitorio es una forma de onda no cíclica de mucho mayor nivel o amplitud que los sonidos alrededor o el nivel promedio, puede también ser definido como un pico de corta duración, ej.: los instrumentos de percusión, el "pluck"generado al pulsar una cuerda con una púa, algunas consonantes en el habla generan transitorios marcados (como la T).

2.3 Segmentación basada en una medida de transición o salto

El principio detrás de este paradigma de segmentación es localizar los "saltos" en las secuencias temporales, marcando las tramas de habla en las que el valor de los rasgos acústicos cambia rápida y significativamente. En [22] esto se hace definiendo una función de salto (*jump function*) asociada a la serie temporal $x[n]$ dada por:

$$J^a[n] = \left| \sum_{m=n-a}^{n-1} \frac{x[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x[m]}{a} \right|, \tag{10}$$

donde $n \in [1,N]$ representa el intervalo de tramas correspondientes a la frase hablada.

El significado de esta función de salto es intuitivo: para cada trama n, representa la diferencia absoluta entre los valores de la media de $x[n]$ calculada en las a tramas anteriores y en las a posteriores. Si $x[n]$ tiene un salto en una trama en particular, $J^a[n]$ presentará un pico o máximo local en dicha trama, y la altura de dicho máximo será proporcional a la brecha.

El ejemplo anterior es una de las formas de detectar un cambio espectral a bajo nivel, y su relativa simpleza obliga a repensar las representaciones sobre las que trabajará. Este aspecto no es objetivo de profundización en este reporte, sin embargo no se puede obviar que rasgo y criterio de segmentación son un binomio a optimizar en conjunto.

Las variaciones espectrales resultan muy útiles para detectar los límites de las sílabas, teniendo en cuenta que el cerebro humano reacciona pobremente a estímulos estacionarios y mucho mejor a transiciones o cambios espectrales [23].

En el procesamiento de voz, la representación sobre el dominio cepstral tuvo un gran impacto, de hecho los rasgos cepstrales tiene la cuestionable ventaja de haber probado ser eficaces en casi todas las tareas de reconocimiento relacionadas con la voz. Sobre el cepstrum (Anexo 2) han sido desarrollados muchos métodos de segmentación.

Otra de las medidas de transición espectral (STM[8]) muy empleada, es la propuesta en [23] y retomada por [24][18]. En estos trabajos, la medida de transición puede ser interpretada como la magnitud de la razón del cambio espectral, y se calcula como un valor cuadrático medio. Los pasos para esta segmentación son:

- La señal, en ventanas móviles de 20 milisegundos, es multiplicada con la función de Hamming, buscando atenuar la señal en los bordes de la ventana y minimizar posteriormente frecuencias parásitas.
- Se calculan los coeficientes cepstrales en el escala Mel (MFCC[9]) de dimensión 10, excluyendo al coeficiente que representa la energía total de la señal.
- Se calcula la STM:

$$\xi(m) = \frac{\sum_{i=1}^{D} a_i^2(m)}{D}, \tag{11}$$

  donde $\xi(m)$ es la STM en la trama m, D es la dimensión de los rasgos (en este caso 10) y a_i es el coeficiente de regresión o razón de cambio de los rasgos espectrales:

$$a_i(m) = \frac{\sum_{n=-I}^{I} MFCC_i(n+m) * n}{\sum_{n=-I}^{I} n^2}, \tag{12}$$

  sean n el índice de la trama e I el número de tramas (a un lado y a otro de la trama en cuestión) utilizados para calcular los coeficientes de regresión .

[8] *Spectral Transition Measure*
[9] *Mel-Frequency Cepstral Coefficients*

- La ubicación de los máximos de la secuencia de STM es una estimación de las fronteras de las sílabas.

Esta metodología depende mucho del ajuste de los parámetros para el cálculo de la regresión, intervalos mayores pudiera implicar la pérdida de algunas fronteras mientras que intervalos pequeños conllevarían a una sobre segmentación. No obstante este método siempre tendrá que hacer frente al problema de que no todos los máximos de STM delimitan fronteras entre fonemas, siendo el más ilustrativo contraejemplo los diptongos.

2.4 Segmentación basada en el algoritmo de agrupamiento de Máximo Margen

Los métodos que emplean funciones kernels han sido muy usados a partir del desarollo de las SVMs y su exitosa aplicación en diversos campos. Por mencionar algunos, las SVMs se han convertido en un componente esencial de los sistemas de reconocimiento de locutores del estado del arte, baste con analizar las competencias NIST anuales. Por el contrario, el uso de funciones kernels en el reconocimiento automático del habla es comparativamente impopular.

El agrupamiento de máximo margen (MMC[10]) es un método con kernels relativamente nuevo [25] y prometedor. Es una variante no supervisada de la SVM, ya que ambas buscan maximizar el margen o separación al hiperplano. La diferencia radica en que SVM maximiza el margen entre las clases etiquetadas y MMC por su parte, asigna etiquetas a los rasgos que maximicen el margen, es por esto que es un algoritmo no supervisado.

La propuesta de segmentación realizada en [26] emplea rasgos cepstrales y sus derivadas, luego define las dimensiones para una ventana móvil que se desplazará a lo largo de las tramas (duración de la señal), con un paso de 1 rasgo. A los rasgos en dicha ventana se le aplica un kernel de base radial (*RBF kernel*) o Gaussiano y en ese espacio se realiza el agrupamiento con MMC, asignando a cada uno de los rasgos de la ventana una etiqueta.

Del procedimiento descrito, se obtiene, por cada ventana analizada, un vector binario de dimensión N, donde N es el número de tramas de la ventana. Como las ventanas se desplazan con paso uno, los vectores tendrán información redundante con los vectores de su N-vecindad, y en una imagen binaria de esta representación, es de esperar patrones diagonales.

Resulta interesante la posibilidad de emplear técnicas de segmentación de imágenes para localizar estos fenómenos, de igual forma, la evaluación de este método en [27] deja abierta la puerta de una información sub-fonética observada, y con potencial, a nuestro criterio, para ser explotado en la identificación de idiomas hablados.

3 Segmentación Gaussiana

Como es conocido, la señal de habla es portadora de muchas y variadas fuentes de información. Para la tarea de reconocimiento del idioma hablado (LRE [11] de sus siglas en inglés.) surge la interrogante de si los fonemas u otra unidad lingüística similar son realmente necesarios, quizás pudiera prescindirse del concepto de palabra formada por fonemas.

Sobre esta idea nuevos enfoques de LRE han comenzado a emerger [28].

En [29] Torres-Carrasquillo utiliza una secuencia de índices Gaussianos para modelar la información del idioma. Adami en [30] emplea trayectorias temporales de la frecuencia fundamental y de la energía a

[10] *Maximum margin clustering*
[11] *Language Recognition Evaluation*

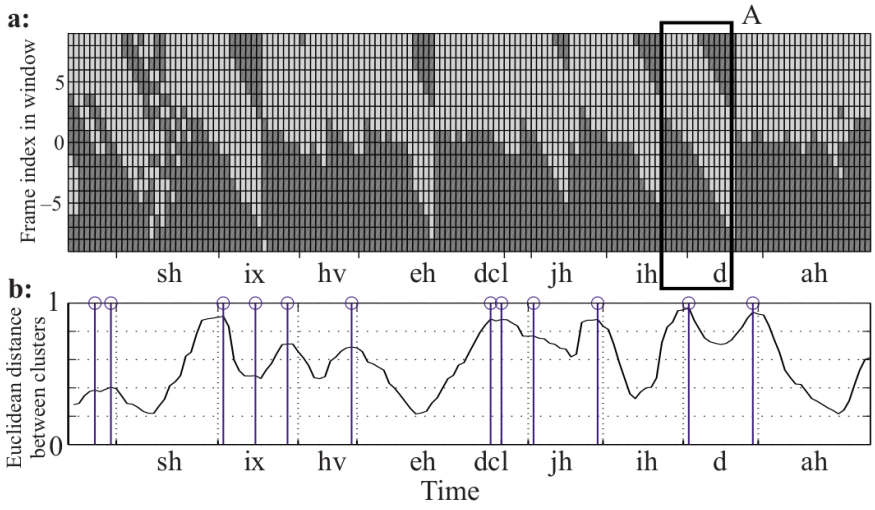

Fig. 1. Representación en una imagen binaria de las etiquetas asignadas por MMC. Las columnas tienen las etiquetas de los $N = 18$ rasgos de cada ventana, correspondiendo las etiquetas inferiores a rasgos anteriores a las superiores. (Imagen tomada de [27])

corto término para segmentar y etiquetar la señal de habla como pequeños conjuntos de unidades discretas que permiten caracterizar el idioma. Más recientemente Spada en [31] intenta aproximar una segmentación fonética usando las variaciones en el espectrograma de la señal.

Uno de los atributos atractivos del Modelo de Mezclas Gaussianas (GMM) es su capacidad para modelar distribuciones arbitrarias de datos, a partir de aproximarse a las clases subyacentes con las componentes Gaussianas individuales. En esta sección explicaremos el algoritmo de segmentación acústica basado en Gaussianas, presentado en [32]. La idea del mismo es condicionar la entrada al *tokenizador* Gaussiano propuesto en [29]. Se busca eliminar segmentos ruidosos y dar mayor peso a eventos de mayor duración. La información empleada parte, en todos los casos, del dominio cepstral[12].

3.1 Segmentación sobre el dominio cepstral

Los vectores de rasgos pueden ser vistos como puntos en el espacio N-dimensional, donde N es la dimensión de los rasgos. Junto a la influencia de los efectos de la variabilidad de sesión [33], dichos vectores de rasgos representan también el estado de nuestros órganos articulatorios, por tanto, dado que el movimiento de los órganos articulatorios es lento, sin perder generalidad se considera que los rasgos consecutivos en el dominio temporal resultarán igualmente cercanos en el dominio cepstral. En otras palabras, intervalos sonoros acústicamente estables, corresponden a rasgos consecutivos en el dominio cepstral.

Estas ideas nos motivaron a pensar que una buena forma de definir unidades acústicas pudiera ser agrupando rasgos cercanos en el dominio cepstral. Spada en [31] obtuvo segmentos acústicos usando una función de variación espectral basada en la distancia Euclidiana entre los MFCC, a la izquierda y derecha de la trama en cuestión.

Nuestra propuesta incorpora la dinámica de los rasgos, utilizando los MFCC, sus derivadas (Δ) y su aceleración ($\Delta\Delta$).

[12] Entiéndase por dominio cepstral al espacio donde yacen los rasgos cepstrales (Anexo 2).

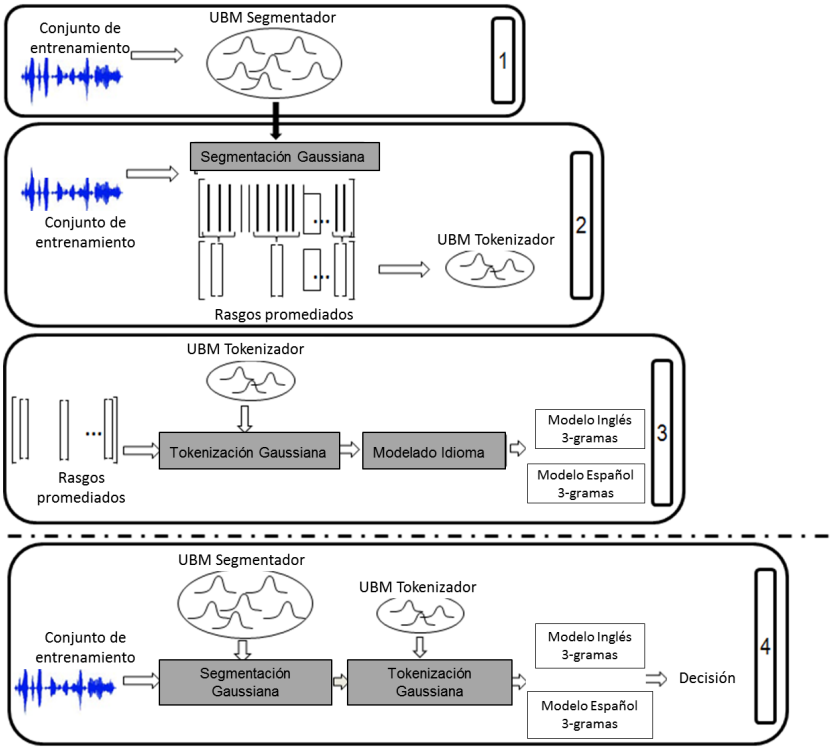

Fig. 2. Metodología propuesta: los tres primeros pasos pertenecen al entrenamiento, el 4^{to} a la prueba.

El primer paso es la creación de un GMM entrenado con los idiomas involucrados en el reconocimiento, este modelo lo llamaremos UBM[13] segmentador, pues una vez entrenado indicará las fronteras entre los segmentos (ver Figura 2). Los GMM permiten modelar de forma general la distribución acústico-fonética de los idiomas sobre los cuales el modelo fue entrenado.

Usando el UBM segmentador, se asigna a cada trama del conjunto de entrenamiento las dos Gaussianas más probables, significando que la información acústica del rasgo de la trama en cuestión se encuentra ubicado con mayor probabilidad debajo de esas dos campanas. Dos tramas permanecerán unidas si comparten una de las dos Gaussianas más probables, de otra forma serán separadas y cada una formará parte de una unidad acústica diferente. Si existe una trama cuya Gaussiana más probable no está presente en ninguna de sus tramas vecinas, dicha trama es eliminada. Los rasgos pertenecientes a la misma unidad acústica son promediados, quedando representado cada segmento con un vector promedio.

Una vez realizada la segmentación, otro clasificador generativo es entrenado, pero esta vez con muchas menos clases. A este modelo le llamamos UBM *tokenizador*, y su función es identificar con la Gaussiana más probable, las unidades acústicas previamente segmentadas y representadas con un vector promedio. Con este procedimiento, el vocabulario o número de *tokens* en el alfabeto, es el mismo que el número de mezclas Gaussianas del UBM *tokenizador*. Esto impone una cierta relación entre las fronteras de las unidades acústicas y la de los fonemas, acercando de forma indirecta la segmentación propuesta a una segmentación fonética.

[13] *Universal Backgroud Model*

3.2 Representación sobre el dominio Gaussiano

La asignación de un índice Gaussiano a un rasgo cepstral, es algo sumamente sensible a variaciones provocadas por ruido, es además el punto de partida sobre el cual se construye la segmentación, la modelación estadística de los idiomas (Anexo 1) y el reconocimiento con la metodología propuesta en [32]. Lo anterior lleva a valorar la posibilidad de diseñar una estrategia más robusta de representación.

La propuesta presentada a continuación es seguir segmentando los rasgos como en [32], o sea viendo el comportamiento de los índices de las Gaussianas más probables en cada rasgo cepstral. La diferencia radica en que el vector que representará a cada segmento no será el vector promedio de los correspondientes rasgos cepstrales, sino el promedio de los vectores de media de las Gaussianas (ver Figura 3).

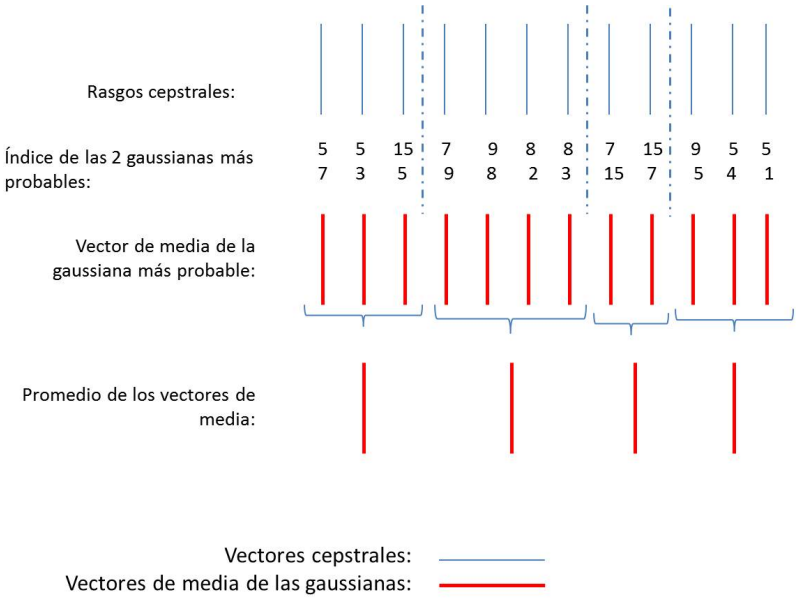

Fig. 3. Representación de las unidades acústicas en el dominio Gaussiano.

3.3 Experimentos y discusión

En esta sección se describen los detalles experimentales del método propuesto en [32] y de la variante del mismo presentada en 3.2, mostrando solo los resultados de este último.

Como parte del procesamiento estándar, la señal de habla es dividida en tramas superpuestas de 25ms, con un desplazamiento de 10ms, donde se asume estacionaridad o invarianza en el tiempo[14]. Los rasgos empleados para describir cada trama son los MFCC.

Los experimentos fueron llevados a cabo sobre parte de la base OGI[15] [35]. Los idiomas involucrados fueron inglés, alemán y español inicialmente, luego incluídos el francés, hindi y japonés.

[14] El habla nunca es invariante en el tiempo, los órganos articulatorios (sobre todo los pulmones, la lengua y los labios) están en constante movimiento en lo que se habla. La hipótesis de invarianza en el tiempo sostiene que tanto la fuente como el resonador son fijos, por tanto lo menos errado que pudiera asumirse es que el habla es aproximadamente estacionaria sobre intervalos cortos de tiempo, y los parámetros de cualquier modelo deberían constantemente ser actualizados. En la práctica esta es la aproximación más frecuente para lidiar con esta limitación[34].

[15] *Oregon Graduate Institute Multi-Language Telephone Speech*

Una evaluación inicial del método de segmentación en el dominio cepstral se realizó y publicó en [32], los mejores resultados pertenecen a la evaluación sobre señales de prueba de 30s de duración[16] y probaron ser superiores a los obtenidos en [31], sobre todo teniendo en cuenta el reducido tamaño del conjunto de entrenamiento (inferior a las 3 horas).

Teniendo en cuenta este positivo resultado, inicialmente con 2 idiomas, en un segundo grupo de experimentos se incluyó el alemán al UBM segmentador, y con dicho modelo se segmentaron señales de los 6 idiomas anteriormente mencionados, se *tokenizaron* y se construyeron sus respectivos LM. Obsérvese en la Tabla 1, en negrita, cómo es determinante el hecho de que en la segmentación no se haya entrenado un idioma en particular, para que luego no funcione el reconocimiento con las unidades acústicas.

Tabla 1. Segmentación Gaussiana. El UBM segmentador fue entrenado solo con Inglés, Español y Alemán, obteniendo en estos idiomas los mejores resultados.

Idioma	EER	DCF
Inglés	**25.22 %**	**0.16 %**
Español	**26.66 %**	**0.91 %**
Alemán	**22.19 %**	**0.15 %**
Francés	76.35 %	0.50 %
Hindi	68.99 %	0.49 %
Japonés	63.47 %	0.48 %

Buscando obtener resultados competitivos para los 6 idiomas, fue entrenado un UBM segmentador con todos los idiomas, sin embargo aquí la explicación a los resultados es que al incluir más idiomas, aumenta la variabilidad del espacio acústico a modelar y se impone elevar el número de mezclas del GMM. Esto obligaba a aumentar los conjuntos de entrenamiento, rompiendo de cierta forma con la premisa de obtener resultados competitivos con un mínimo de datos. Esta situación no fue prevista porque encontrar el número óptimo de mezclas Gaussianas por lo general es empírico, y había una posibilidad de que los fonos incorporados por los nuevos idiomas no se esparcieran tanto en el espacio de las clases acústicas, sin embargo no fue el caso.

También fueron llevados a cabo experimentos aplicando la representación Gaussiana a la segmentación propuesta en [32], buscando respaldar la hipótesis de que sería una representación más robusta ante la variabilidad de sesión o ruido.

Se reprodujeron las condiciones de los experimentos presentados en [32], o sea se entrenó un UBM segmentador con 3 idiomas (inglés, español y alemán). Los resultados no son alentadores, pues asignar a cada trama el vector de medias Gaussiano ya es un proceso de "suavizado" muy fuerte (todos los rasgos cepstrales modelados con la misma mezcla, pasarán a ser representados por el vector de medias de dicha mezcla), si además son promediados estos vectores de media, la descripción pierde aun más detalle.

Por lo que una conclusión importante de estos resultados es que para delimitar fronteras entre objetos, es fundamental que los mismos representen características puntuales, que la información que exhiban asuma independencia entre los objetos vecinos.

4 Conclusiones

Del estudio y análisis de los métodos de segmentación acústica no supervisados expuestos en la Sección 2, cabe señalar que todos a priori proponen acercarse a la segmentación fonética, pues la mayoría se realizan

[16] Las pruebas de reconocimiento de idioma de manera estándar se realizan sobre conjuntos de señales de duración 3, 10 y/o 30 segundos.

con vistas a preceder el reconocimiento de habla. Por lo anterior muchas veces su diseño no se ajusta del todo al fin de segmentar en unidades subfonéticas o abstractas, como las buscadas por nosotros para el reconocimiento del idioma hablado.

La unidad acústica perseguida por nosotros con la segmentación en el dominio Gaussiano (3.2), pudiera coincidir en algunos casos con fonemas, pero no parten de ese supuesto inicial, de hecho las semejanzas se buscan a un nivel en el que es muy difícil definir algo como fonemas.

Otra característica observada en los métodos analizados es que la forma en la que evalúan la segmentación, exige disponer de una base con las fronteras marcadas de las unidades acústicas buscadas. Calculan el desempeño del método hallando razones entre los resultados obtenidos y los reales. Estas métricas no son funcionales cuando la unidad acústica es abstracta y no está asociada directamente con una representación fonética o lingüística, como es nuestro caso.

Esta es precisamente una de la cuestiones aún sin resolver en nuestra metodología, ya que actualmente estamos considerando que un método de segmentación será más efectivo que otro, siempre que con sus unidades acústicas se logre un desempeño mejor del sistema como un todo, y aquí se impone una utilización bien pensada y elegida, del resto de las herramientas que intervienen en un proceso de reconocimiento de idioma hablado, o sea más grados de libertad.

De la revisión de métodos basados en el retraso de grupo, su aproximación a fonemas es muy sencilla de implementar, sin embargo tiene limitaciones para delimitar sonidos altamente transitorios y fricativos, porque no resuelve variaciones energéticas tan rápidas, por lo que su empleo pudiera pensarse complementado por otro método o criterio de segmentación. Ha demostrado su positivo impacto el filtro Gaussiano, mejorando el desempeño de la segmentación con un post procesamiento.

Los métodos basados en funciones de transición o salto, buscan variaciones espectrales y enfrentan el problema contrario al que tiene la función de retraso de grupo. Sucede que no todos los máximos de las funciones de transición son fronteras fonéticas y la sobresegmentación es frecuentemente obtenida con estos métodos.

Por su parte, el empleo del MMC, para segmentar analizando de forma binaria cada decisión y sin requerir un previo etiquetado para su entrenamiento, resulta la arista más interesante para nosotros, sobre todo porque es una técnica recientemente utilizada en la segmentación acústica, con probados resultados, y porque ha permitido representar información subfonética.

El presente reporte resume también los primeros resultados de los autores, sobre una línea investigativa que busca segmentar audio, en unidades acústicas no definidas como fonemas y que sean portadoras de información útil para el posterior reconocimiento de idiomas.

Todos los experimentos realizados arrojan luz sobre la idea central de investigación, pero aún continúan muchas preguntas sin responder. Un primer resultado alentador [32] permitió superar los *scores* de identificación de un método que buscaba regiones acústicamente homogéneas minimizando la distancia Euclidiana entre los rasgos cepstrales. Vale la pena resaltar el pequeño volumen de datos empleado para obtener dichos resultados. No obstante los intentos por incorporar más idiomas e igualar o disminuir los *scores* no fueron positivos ya que intentamos reproducir los experimentos realizados para 3 idiomas, con 6; sin tener en cuenta que el nuevo espacio acústico era portador de una mayor variabilidad acústica, por tanto resultaría saludable modelar su distribución con GMMs de mayor número de componentes, esto pudo definir la baja eficacia.

Delineamos como futuro trabajo la adopción o ajuste de una métrica para evaluar los resultados de la segmentación, sin tener que llegar a la clasificación final del idioma. Continúa siendo nuestro principal objetivo proponer nuevos métodos de segmentación y *tokenización* a partir del análisis de los ya existentes y de la experiencia acumulada con las experimentaciones iniciales.

Referencias bibliográficas

1. King, S., Hasegawa-Johnson, M.: Accurate speech segmentation by mimicking human auditory processing. In: ICASSP. (2013) 8096–8100
2. Hemmert, W., Holmberg, M., Ramacher, U.: Temporal sound processing by cochlear nucleus octopus neurons. In Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S., eds.: ICANN (1). Volume 3696 of Lecture Notes in Computer Science., Springer (2005) 583–588
3. Oertel, D., Bal, R., Gardner, S.M., Smith, P.H., Joris, P.X.: Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. Proc Natl Acad Sci U S A **97**(22) (2000) 11773–9
4. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing. Elsevier Science (2006)
5. Li, H., Ma, B., Lee, K.A.: Spoken language recognition: From fundamentals to practice. Proceedings of the IEEE **101**(5) (2013) 1136–1159
6. Brugnara, F., Falavigna, D., Omologo, M.: Automatic segmentation and labeling of speech based on hidden markov models. Speech Communication **12**(4) (1993) 357–370
7. Montalvo, A., Calvo, J.R.: Reconocimiento del idioma hablado: tendencias actuales. Technical Report 057, Centro de Aplicaciones de Tecnologías de Avanzada (Octubre 2013)
8. Oropeza Rodríguez, J.L., Suárez Guerra, S.: Algoritmos y métodos para el reconocimiento de voz en español mediante silabas. Computación y Sistemas **9**(3) (2006) 270–286
9. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1993)
10. Kos, M., Kacic, Z., Vlaj, D.: Acoustic classification and segmentation using modified spectral roll-off and variance-based features. Digital Signal Processing **23**(2) (2013) 659–674
11. Kos, M., Grasic, M., Kacic, Z.: Online speech/music segmentation based on the variance mean of filter bank energy. EURASIP J. Adv. Sig. Proc. **2009** (2009)
12. Prasad, R., Yegnanarayana, B.: Acoustic segmentation of speech using zero time liftering (ztl). In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., eds.: INTERSPEECH, ISCA (2013) 2292–2296
13. Torres, H.M., Gurlekian, J.A.: Acoustic speech unit segmentation for concatenative synthesis. Computer Speech & Language **22**(2) (2008) 196–206
14. Huang, Z., Cheng, Y.C., Li, K., Hautamäki, V., Lee, C.H.: A blind segmentation approach to acoustic event detection based on i-vector. In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., eds.: INTERSPEECH, ISCA (2013) 2282–2286
15. Dehak, N.: Discriminative and Generative Approaches for Long-and Short-term Speaker Characteristics Modeling: Application to Speaker Verification. Thèse de doctorat en génie. École de technologie supérieure (2009)
16. McAulay, R., Quatieri, T.F.: Speech analysis/Synthesis based on a sinusoidal representation. Acoustics, Speech and Signal Processing, IEEE Transactions on **34**(4) (aug 1986) 744–754
17. Ganesh, A.A., Ravichandran, C.: Syllable based continuous speech recognizer with varied length maximum likelihood character segmentation. In: ICACCI, IEEE (2013) 935–940
18. Patil, H., Patel, T., Talesara, S., Shah, N., Sailor, H., Vachhani, B., Akhani, J., Kanakiya, B., Gaur, Y., Prajapati, V.: Algorithms for speech segmentation at syllable-level for text-to-speech synthesis system in gujarati. In: Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference. (Nov 2013) 1–7
19. Nagarajan, T., Murthy, H.A., Hegde, R.M.: Segmentation of speech into syllable-like units. In: INTERSPEECH, ISCA (2003)
20. Talesara, S., Patil, H.A., Patel, T.B., Sailor, H., Shah, N.: A novel gaussian filter-based automatic labeling of speech data for tts system in gujarati language. In: IALP, IEEE (2013) 139–142
21. Mallat, S.: A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way. 3 edn. Academic Press (2008)
22. Esposito, A., Aversano, G.: Text independent methods for speech segmentation. In: Summer School on Neural Networks. (2004) 261–290
23. Furui, S.: On the role of spectral transition for speech perception. The Journal of the Acoustical Society of America **80**(4) (1986) 1016–1025
24. Dusan, S., Rabiner, L.R.: On the relation between maximum spectral transition positions and phone boundaries. In: INTERSPEECH, ISCA (2006)
25. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: Advances in Neural Information Processing Systems 17, MIT Press (2005) 1537–1544
26. Estevan, Y., Wan, V., Scharenborg, O.: Finding maximum margin segments in speech. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Volume 4. (April 2007) IV–937–IV–940
27. Scharenborg, O., Wan, V., Ernestus, M.: Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. The Journal of the Acoustical Society of America **127**(2) (2010) 1084–1095

28. Kłosowski, P., Dustor, A.: Automatic speech segmentation for automatic speech translation. In Kwiecień, A., Gaj, P., Stera, P., eds.: Computer Networks. Volume 370 of Communications in Computer and Information Science. Springer Berlin Heidelberg (2013) 466–475
29. Torres-Carrasquillo, P.A., Reynolds, D.A., Deller, J.R.: Language identification using gaussian mixture model tokenization. In: ICASSP. (2002) 757–760
30. Adami, A.G., Hermansky, H.: Segmentation of speech for speaker and language recognition. In: INTERSPEECH. (2003)
31. Spada, D., Lopez, I., Toledano, D., González-Rodríguez, J.: Acoustic event recognition for low cost language identification. In: V Jornadas en en Tecnologías del Habla 2007. UAM. (2007)
32. Montalvo, A., Calvo, J.R., Hernández, G.: Gaussian segmentation and tokenization for low cost language identification. In Ruiz-Shulcloper, J., Sanniti-di Baja, G., eds.: CIARP (1). Volume 8258 of Lecture Notes in Computer Science., Springer (2013) 551–558
33. Montalvo, A., Calvo, J.R.: Métodos para reducir la variabilidad de sesión en el reconocimiento del locutor. Technical Report 051, Centro de Aplicaciones de Tecnologías de Avanzada (Agosto 2012)
34. Little, M.A.: Mathematical foundations of nonlinear, non-gaussian, and time-varying digital speech signal processing. In Travieso-González, C.M., Hernández, J.B.A., eds.: NOLISP. Volume 7015 of Lecture Notes in Computer Science., Springer (2011) 9–16
35. Muthusamy, Y.K., Cole, R.A., Oshika, B.T.: The ogi multi-language telephone speech corpus. In: ICSLP. (1992)
36. Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H.: Universal attribute characterization of spoken languages for automatic spoken language recognition. Computer Speech & Language **27**(1) (2013) 209–227
37. Rosenfeld, R.: The cmu statistical language modeling toolkit and its use in the 1994 arpa csr evaluation. ARPA SLT **95** (1995)
38. Benesty, J., Sondhi, M.M., Huang, Y., eds.: Springer Handbook of Speech Processing. Springer, Berlin (2008)
39. Oppenheim, A., Schafer, R.: From frequency to quefrency: a history of the cepstrum. Signal Processing Magazine, IEEE **21**(5) (Sept 2004) 95–106

Anexos

1 Modelación estadística del idioma

El enfoque de las aproximaciones propuestas y estudiadas en el presente reporte, es el fonotáctico, ya que este ha probado ser el de mejor desempeño en las tareas de identificación de idioma hablado [36]. El objetivo de los modelos del lenguaje o del idioma (LM), es brindar una sintaxis que defina posibles secuencias de *tokens* y permita el cálculo de la probabilidad ($P(W|L)$) de una cadena de *tokens* $W = (w_1, w_2, ..., w_Q)$ dado el LM L.

El LM de un idioma particular es creado usando el conjunto de entrenamiento de dicho idioma, el cual fue previamente secuenciado con el Modelo Universal de *Background* (UBM) *tokenizador*, el cual no es más que una modelo de mezclas Gaussianas que contiene información acústica de todos los idiomas involucrados en el experimento y busca una representación genérica de los mismos. Haciendo uso de la herramienta CMU-SLM [37] (*Carnegie Melon University Statistical Language Modeling*) se obtuvieron modelos de tri-gramas para cada idioma. Posteriormente en la fase de prueba, la probabilidad de una tripleta de índices de Gaussianas (tri-grama) se calcula:

$$P(w_i|w_{i-1}, w_{i-2}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}; \tag{13}$$

donde $C(w_{i-2}, w_{i-1}, w_i)$ y $C(w_{i-2}, w_{i-1})$ son el conteo de ocurrencia del tri-grama (w_{i-2}, w_{i-1}, w_i) y del bi-grama $(w_{i-2}, wi-1)$ respectivamente, en el conjunto de entrenamiento.

Por lo que para cada secuencia de índices, el logaritmo de su probabilidad es calculada:

$$\log P(W|L) = \sum_{i=1}^{Q} logP_L(w_i|w_{i-1}, w_{i-2}), \tag{14}$$

siendo $P_L(w_i|w_{i-1}, w_{i-2}$ es la probabilidad de la tripleta w_{i-2}, w_{i-1}, w_i para el idioma L y Q la cantidad de tokens de la frase analizada. El idioma correspondiente al LM que maximice $P(W|L)$, es seleccionado como el idioma buscado.

2 Definición del cepstrum

El cepstrum de una señal [38] es el resultado de calcular la transformada discreta de Fourier (DTFT[17]) del espectro de la señal estudiada en escala logarítmica (dB). El nombre cepstrum deriva de invertir las cuatro primeras letras de *spectrum*. El cepstrum es un número complejo, por tanto, tiene su parte real y su parte imaginaria.

Para señales discretizadas temporalmente ($x[n]$), el cepstrum es definido por:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(e^{i\omega})| e^{i\omega n}, \tag{15}$$

donde la DTFT es:

$$x(e^{i\omega}) = \sum_{n=-\text{ínf}}^{\text{ínf}} x[n] e^{-i\omega n}. \tag{16}$$

[17] *Discret Time Fourier Transform*

Nótese que $c[n]$, siendo la transformada inversa de la DTFT, es una función de un índice discreto (n). Si la secuencia de entrada es obtenida muestreando una señal analógica ($x[n] = x_a(n/fs)$), entonces es natural asociar el tiempo con el índice n del cepstrum. Sin embargo, elaborando la "identidad del cepstrum" [39] se introduce el término *quefrency* para identificar a la variable independiente del cepstrum. Este nuevo término es útil describiendo las propiedades fundamentales del cepstrum. Por ejemplo, baja *quefrency* se corresponde con componentes lentas del espectro del logaritmo de la magnitud, mientras que valores elevados se asocian a componentes de rápida variación.

RT_062, julio 2014
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba