Segmentación Automática de Voz basada en Modelos Ocultos de Markov y Características Acústicas *

Laura Docío Fernández y **Carmen García Mateo**
E.T.S. E. Telecomunicación, Universidade de Vigo
Tlf. 986 812133
ldocio,carmen@gts.tsc.uvigo.es

Resumen Un aspecto muy importante en el ámbito de las tecnologías del habla, lo constituyen las bases de datos segmentadas y etiquetadas de forma precisa ya sea a nivel fonético, de sub-palabra o de palabra. Sin embargo, un etiquetado y segmentación manual es una tarea que consume mucho tiempo y muy propensa a errores. Este artículo describe un procedimiento automático para realizar la segmentación de voz en un conjunto de unidades acústicas: dado el contenido fonético o lingüístico de una locución, el sistema proporciona las fronteras temporales de las unidades. La técnica se basa en el uso de un reconocedor que utiliza modelos ocultos de Markov (HMMs) para modelar cada una de las unidades acústicas. Dicho reconocedor proporciona una segmentación burda inicial obtenida a través de un alineamiento de Viterbi, la cual será refinada posteriormente por medio de una "segmentación acústica" y un pequeño conjunto de reglas basadas en características acústicas. Estas reglas representan conocimiento fonético y su finalidad consiste en en la corrección de errores de segmentación inesperados, los cuales son un principal problema de los reconocedores basados en HMMs.

1 *Introducción*

La segmentación y etiquetado de bases de datos de voz, de acuerdo con la transcripción fonética, representa una tarea fundamental en muchas áreas de las Tecnologías del Habla. Disponer de un gran corpus de voz transcrita de forma precisa resulta esencial en muchos ámbitos incluyendo el entrenamiento de reconocedores de voz basados en fonemas o sub-palabras, la extracción de prototipos e información de duración para el diseño de un sintetizador de voz, y para análisis de voz. Dichas bases de datos son poco frecuentes y contienen muchas variaciones en cuanto a la elección de fonemas, el tipo de locuciones, la cobertura de trifonemas, y la consistencia y calidad de la transcripción proporcionada.

Cuando la meta final es la localización temporal de las fronteras de unidades de tipo fonema, se puede utilizar una segmentación manual, pero ésta tiene dos importantes inconvenientes:

- Se trata de un proceso muy laborioso, tedioso, y lento.
- Debido a la naturaleza subjetiva de la segmentación manual habrá inconsistencias de una segmentación a otra, incluso para segmentar la misma locución.

Con el objetivo de resolver los anteriores problemas se han propuesto, en la literatura, diversos procedimientos automáticos para segmentar voz en unidades de tipo sub-palabra. La mayoría de tales métodos han seguido una de las dos siguientes aproximaciones básicas al problema. La primera, consiste en utilizar la información que se conoce a priori, tal como la transcripción correcta de la frase [3] [4]. La segunda categoría no utiliza información relativa a la transcripción, sino sólo información acústica incluida en la señal de voz [5] [1] [2].

Los métodos de segmentación englobados en la primera de las categorías dividen la señal de entrada en segmentos definidos explícitamente por la transcripción fonética u ortográfica. En general, una desventaja de estas técnicas de segmentación es que precisan de *patrones de referencia* o de algún tipo de *modelos* antes de su utilización. Al utilizar información fonética y lingüística sobre

* Este trabajo ha sido parcialmente por la CICYT con del proyecto FEDER 1FD97-0077 "Sistemas de diálogo para el acceso telefónico a servicios telemáticos"

la señal de voz a segmentar, se podría pensar que estos métodos proporcionarían muy buenos resultados. Sin embargo, los patrones de referencia o modelos utilizados para describir cada segmento no siempre representan bien la señal a segmentar, y además, normalmente no tienen en cuenta toda la variabilidad de la señal de voz. Además, los errores inducidos por estos métodos de alineamiento temporal limitan su campo de aplicación.

Los métodos de segmentación que encajan dentro de la segunda categoría no asumen ningún tipo de categorización lingüística de los "segmentos" a obtener, y se basan únicamente en las características acústicas de la señal de voz. Por lo tanto, los "segmentos" se obtendrán a partir de información acústica que es *local* a los mismos. Estas técnicas están siempre más o menos asociadas a una cierta *medida de distorsión* entre observaciones acústicas.

Las características generales de ambos tipos de segmentación son casi complementarias, lo que indica que una combinación de dos de tales métodos puede proporcionar mejores prestaciones que cada método por separado. De este modo, teniendo en mente lo anterior, en este artículo se propone una aproximación en la que, con el objetivo de mejorar el resultado final obtenido, la segmentación se realiza utilizando un sistema *híbrido* que hace uso de ambos tipos de segmentación automática.

Como la aplicación de dicha segmentación automática está destinada tanto al entrenamiento y evaluación de un reconocedor como al diseño de un sistema de conversión texto-a-voz (TTS), se asumirá en todo momento que se conoce la transcripción fonética de la locución, y que sólo se ha de obtener de forma automática la localización temporal de las fronteras de las unidades.

2 Descripción del sistema

El sistema de segmentación automática propuesto se deriva de un reconocedor de voz basado en un conjunto específico de HMMs que modelan unidades acústico-fonéticas.

La figura 1 muestra un diagrama de bloques del sistema de segmentación propuesto. La entrada al sistema consiste en una señal de voz y su correspondiente transcripción fonética u ortográfica. La salida la constituyen un conjunto de instantes temporales que acotan las unidades incluidas en la transcripción fonética proporcionada.

En primer lugar se realiza un análisis acústico de la señal de voz con el objetivo de definir el espacio sobre el que se realizarán todos los cálculos. A continuación, se realiza el cálculo de una primera aproximación a la localización de las transiciones entre las unidades empleadas en el modelado de las palabras del vocabulario, utilizando para ello un alineamiento de Viterbi. Paralelamente, se obtiene un conjunto de fronteras ("segmentación acústica") que, con elevada probabilidad, representarán a las unidades acústicas buscando zonas estables en la señal de voz, aplicando para ello criterios de estacionariedad espectral sobre una función de un conjunto de funciones de variación espectral. Finalmente, utilizando dichas fronteras se realiza un **reajuste** de las fronteras proporcionadas por el alineamiento de Viterbi. El módulo que realiza esta tarea se puede considerar como un "combinador o mezclador" de la información de las dos segmentaciones.

2.1 Análisis acústico

Este módulo proporciona los vectores de observación, constituidos por parámetros acústicos, a los subsiguientes módulos los cuales constituyen el bloque fundamental del sistema de segmentación automática desarrollado.

El módulo que implementa el alineamiento de Viterbi utiliza los mismos parámetros que los empleados en el diseño de los HMMs utilizados. Realiza el análisis acústico cada 10ms utilizando una ventana Hamming de 20ms, para obtener los siguientes parámetros que definen a los vectores de observación:

1. 12 coeficientes mel-cepstrum calculados a partir de un banco de 24 filtros paso banda triangulares.
2. Las correspondientes 12 derivadas de primer orden.
3. La log-energía normalizada y su derivada temporal de primer orden.

Para las etapas de refinamiento de la segmentación este módulo genera además otro conjunto de parámetros acústicos. En concreto, proporciona un vector de observación obtenido a través de un análisis acústico realizado cada 1ms, y consistente de un conjunto de parámetros de ambos dominios: tiempo y frecuencia

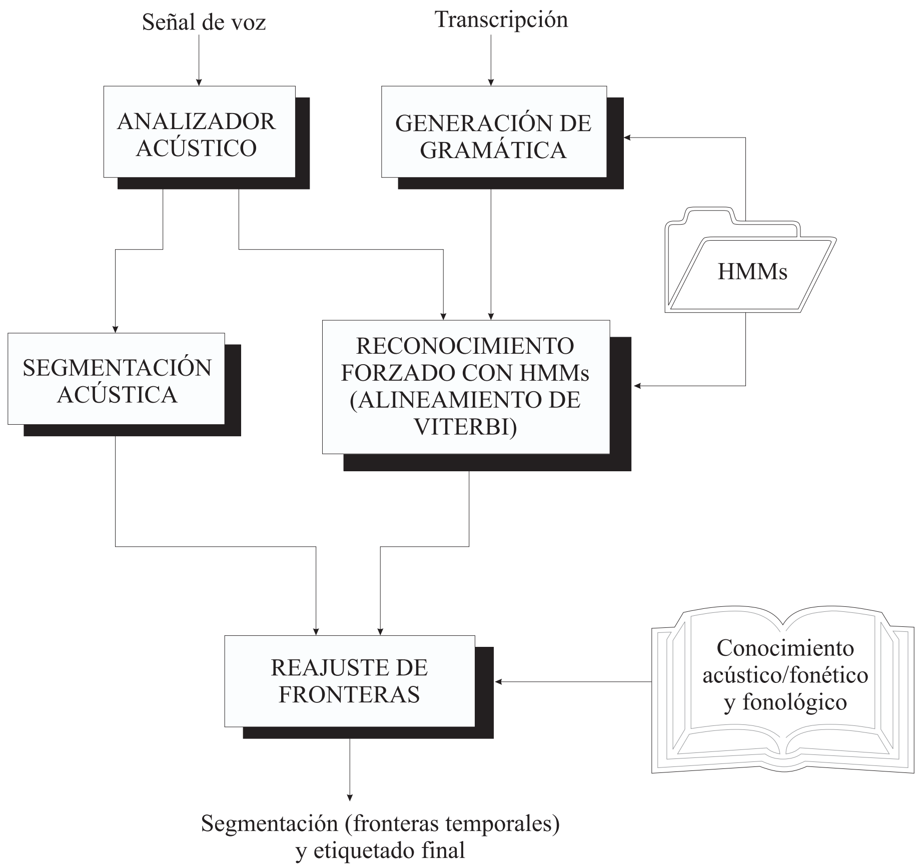

Figura 1: Arquitectura del sistema de segmentación automática

- La tasa de cruces por cero.
- El contorno de energía en el dominio logarítmico.
- La frecuencia fundamental o pitch.
- Un promedio de tres funciones de variación espectral.
- La energía promediada por un conjunto de filtros en la escala mel, que da una medida de la distribución de energía espectral.

Cada parámetro obtenido se normaliza al rango $[0, 1]$, con el objetivo de conseguir un margen dinámico similar para cada uno de ellos. También se calcula la *tasa de cambio* de cada uno de ellos para añadir al vector de características.

2.2 Alineamiento de Viterbi

Como se ha dicho anteriormente, el problema de segmentación consiste en la búsqueda las fronteras más probables correspondientes a las "unidades fonéticas" contenidas en la frase. En primer lugar, se obtienen las fronteras o marcas temporales de las transiciones entre las *unidades de modelado* presentes en la locución a segmentar. El método utilizado para llevar a cabo esta tarea utiliza un conjunto de HMMs que modelan a dichas unidades, y, básicamente, realiza un *reconocimiento forzado* entre la secuencia de vectores acústicos y los HMMs disponibles en base a la transcripción dada, a través de un alineamiento básico de Viterbi. Si los modelos están bien entrenados, es de esperar que las fronteras obtenidas están próximas a las transiciones verdaderas entre las unidades presentes.

2.3 Segmentación acústica

Los sistemas basados en un reconocimiento automático de habla pueden no ser la mejor ruta para la obtención de una segmentación automática de alta calidad, debido a que se construyen para identificar segmentos fonéticos y no para detectar fronteras entre segmentos fonéticos.

El objetivo de esta etapa consiste en obtener un conjunto de *posibles* fronteras que permitan fijar posteriormente, con ayuda de los resultados proporcionados por el alineamiento de Viterbi y un conjunto de reglas acústico-fonéticas, las fronteras óptimas entre unidades.

Básicamente, la estrategia implementada se basa en un criterio de homogeneidad acústica que evalúa la "probabilidad" de que un segmento acústico corresponda a un único segmento fonético conforme a sus variaciones espectrales, i.e., se pretende determinar una medida de "probabilidad" o "fiabilidad" de que en un determinado instante haya una transición espectral, s_i, sujeto a que en s_{i-1} está la frontera precedente y a su evidencia acústica. Hay que destacar aquí que se producirán inserciones y omisiones de fronteras, por lo que un objetivo básico de los algoritmos utilizados será intentar que el número de fronteras insertadas y omitidas sea el mínimo posible.

Se asume que en una frontera o transición espectral un parámetro determinado de la voz puede aumentar, disminuir o permanecer constante, mientras que entre fronteras los parámetros son aproximadamente constantes. La tasa de cambio de cada parámetro exhibe por lo tanto cambios en una frontera. Estos cambios son, normalmente, muy robustos e independientes del locutor.

El método diseñado para realizar dicha segmentación acústica se basa en una *combinación o promedio* de un conjunto de funciones de variación espectral, utilizadas cada una de ellas como medida de similitud entre tramas consecutivas de la señal de voz.

2.4 Ajuste de las fronteras

Las dos segmentaciones disponibles presentan características casi complementarias. Este hecho hace viable la construcción de un sistema que utilice simultáneamente ambos tipos de técnicas con el objetivo de aumentar la fiabilidad en la segmentación. Se propone entonces una estrategia de *combinación o mezcla* de los resultados de ambas segmentaciones. Se va a analizar una aproximación que tiene en cuenta información contextual.

Como estrategia de combinación de los resultados obtenidos en la etapa anterior del sistema de segmentación automático, se ha diseñado un mecanismo basado en **reglas** con el propósito de transferir el conocimiento acústico/fonético en la segmentación. Cada regla se refiere a una clase de transición y representa *conocimiento* en una forma declarativa. Una regla genérica se expresa de la siguiente forma:

$$\text{Premisas y Contexto} \rightarrow$$
$$\rightarrow \text{Conclusion(es) y Acción(es)}$$

Las **premisas** evalúan parámetros acústicos así como conclusiones de otras reglas. Las conclusiones son secuencias de acciones en la forma de procedimientos que refinan la segmentación explícita.

Las premisas se expresan en términos de *"parámetros acústicos", "condiciones" y "operadores".* Analicemos con un poco de detalle cada uno de estos términos.

Una **condición** permite determinar una trama específica o un intervalo de valores para un parámetro dado en el que se puede aplicar una regla.

Los **operadores** son símbolos utilizados para combinar lógicamente las condiciones mencionadas sobre los parámetros acústicos.

Como un ejemplo, una regla se puede expresar como sigue:

#defregla regla_izqda
[tipo transición] regla_decha

donde "regla_izqda" y "regla_decha" pueden ser:

$$\mathbf{P1 == V1\ operador\ P2} < X1, X2 >$$

La primera condición (P1 == V1) busca la trama en la que el parámetro P1 tiene el valor más próximo a V1. La segunda condición (P2 $< X1, X2 >$) verifica si en una trama dada (e.g. la resultante de la condición previa) el valor del parámetro P2 está dentro del intervalo $[X1, X2]$; finalmente, el *operador* liga lógicamente las dos condiciones, de forma que la regla sólo se aplica si se satisfacen las limitaciones lógicas impuestas por el operador.

La clase de transición especifica los contextos derecho e izquierdo a los que se aplican la regla derecha y la regla izquierda, respectivamente. Por ejemplo, la siguiente regla de segmentación:

$SVF == 0.0$ **&** $\mathbf{E} < 0.8, 1.0 >$
[vocal fricativa] $E == 0.0$ & P1$< 0.3, 1.0 >$

se aplica a todas las vocales seguidas por una fricativa; la frontera izquierda se coloca en la trama en la que el parámetro SVF toma su valor mínimo, sólo si el parámetro energía está dentro del rango 80-100%; la frontera derecha se coloca en la trama de mínima energía, sólo si está alrededor de la parte central de la fricativa, en el rango 30-70% del pitch. Si se satisfacen todas las condiciones se aplica la regla, en otro caso se salta.

3 Base de Datos y Marco Experimental

El sistema de segmentación automática propuesto se ha aplicado a una base de datos prosódica en **gallego**.

Se ha considerado material diferente para entrenamiento y para test. Todas las locuciones han sido segmentadas y etiquetadas manualmente por un fonetista experto.

El entrenamiento de los HMMs se ha realizado partiendo de unos HMMs "semilla" los cuales se han reentrenado utilizando 160 sentencias de un mismo locutor.

Para el idioma gallego se ha considerado un subconjunto de 26 unidades acústico-fonéticas independientes del contexto definidas en la tabla 1. Dichas unidades se han modelado utilizando HMMs continuos con una topología de izquierda-a-derecha de tres estados. Además de los modelos de tales unidades fonéticas se han utilizado también un conjunto de 4 HMMs para modelar silencios así como otros tipos de sonidos: golpes, respiraciones, etc.

Clave	**SAMPA**
p	p
b	b, B
t	t
d	d, D
k	k
g	g, G
m	m
n	n
J	J
N	N
C	tS
f	f
T	T
s	s
S	S
l	l
Z	jj
R	rr
r	r
i	i
e	e, E
a	a
o	o, O
u	u
x	h

Tabla 1: Conjunto de unidades consideradas

Número de repeticiones	Alófono
301	B
313	D
84	E
112	G
41	J
232	N
90	O
116	S
243	T
48	Z
1310	a
56	b
117	d
1144	e
119	f
16	g
840	i
361	k
323	l
363	m
442	n
969	o
291	p
866	r
795	s
611	t
332	u
2	x

Tabla 2: Número de repeteciones de cada alófono en la base de datos de test.

La base de datos de test consiste en 509 sentencias, diferentes de las de entrenamiento, pronunciadas todas ellas por el mismo locutor. El número de repeticiones de cada unidad en dicha base de datos es el mostrado en la tabla 2.

4 Experimentos y Resultados

Una evaluación preliminar del procedimiento de segmentación automática para una base de datos de voz en Gallego ha mostrado que los principales problemas aparecen en transiciones de tipo vocal/vocal y consonante/consonante.

En la tabla 3 se presentan los resultados obtenidos utilizando cuatro contextos fonéticos generales: vocal seguida de vocal, vocal seguida de consonante, consonante seguida de vocal y consonante seguida de consonante. El porcentaje de unidades para los que la diferencia entre la segmentación manual y la automática es mayor que 70ms o menor que 30ms se indica en las dos últimas filas.

	Copus total	V/V	V/C	C/V	C/C
$t < 30ms$	85%	94%	94%	92%	83%
$t > 70ms$	3.5%	2%	1%	0.3%	4%

Tabla 3: Porcentaje de fronteras colocadas correctamente

Se ha comprobado que las transiciones de fonemas pertenecientes a la misma clase fonética son, en general, son menos exactas que las de fonemas pertenecientes a clases distintas.

Referencias

[1] V.R. Algazi and K.L. Brown. Automatic speech recognition using acoustic subwords and no time alignment. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, I:465–468, 1988.

[2] F. Brugnara, D. Falavigna, and M. Omologo. A HMM-based system for automatic segmentation and labeling of speech. *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, pages 803–806, 1992.

[3] L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon, and T.M. Zampini. A bootstrapping training technique for obtaining demisyllable reference patterns. *Journal of the Acoustic Society of America*, 71:1588–1595, 1982.

[4] T. Svendsen and K.F. Soong. On the automatic segmentation of speech signals. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 77–80, 1987.

[5] J.G. Wilpon, B.H. Juang, and L.R. Rabiner. An investigation on the use of acoustic sub-word units for automatic speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, II:821–824, 1987.