XBIC: nueva medida para segmentación de locutor hacia el indexado automático de la señal de voz

Xavier Anguera, Javier Hernando, Jan Anguita

Departamento de Teoría de Señal y Comunicaciones, Centro de Investigación TALP
Universidad Politécnica de Cataluña (UPC)
{xanguera,javier,jan}@gps.tsc.upc.es

Resumen

La evolución de la sociedad de la información ha traído consigo un incesante incremento de contenidos audiovisuales que normalmente se archivan en bases de datos multimedia por tal de poder ser consultadas posteriormente. Debido a la ingente cantidad de datos almacenados resulta difícil y muy costoso poder mantener un indexado fiable de estos datos.

En la actualidad empiezan a aparecer sistemas de indexado automático de material sonoro que podrían pronto ser usados en esta tarea. En esta publicación dirigimos nuestra atención a la segmentación de locutor, que es una parte de estos sistemas y que pretende extraer información de donde y cuando habla cada locutor.

Presentamos una nueva medida llamada XBIC para detectar cambios de locutor, desarrollada en nuestro grupo. Se describen pruebas sobre una base de datos de telenoticias en catalán y en dos bases de datos de Broadcast News en Inglés, sobre las cuales obtenemos resultados prometedores.

1. Introducción

En la actual sociedad de la información en que nos encontramos inmersos son muchas las emisoras de radio y de televisión que emiten sus contenidos 24 horas al día, todos los días del año. Estos contenidos se acostumbran a archivar para poder ser utilizados con posterioridad en futuros programas o contenidos. Para poder acceder de manera eficiente a la información contenida en estos archivos se han ido desarrollando en el pasado sistemas de indexado que permitan una busca rápida de los fragmentos deseados. El problema principal de estos métodos de indexado es que se tienen que hacer de manera manual (etiquetas en las cintas, entradas a mano a una base de datos ...) de manera que resulta muy costoso mantener un registro exhaustivo de los contenidos del archivo.

Con el uso de sistemas de indexado automático se mejora la cantidad y calidad del indexado, requiriendo mucha menos actuación manual. Los métodos de indexado automático comprenden técnicas de tratamiento de audio y vídeo. En esta publicación nos centramos en la parte de audio, común en los contenidos de emisoras de radio y de televisión.

El indexado de la señal de voz se puede lograr concatenando diferentes sistemas de tratamiento de la señal de voz especializados en la extracción de características particulares. Algunas de las informaciones indexables automáticamente son la identidad y el lugar dónde aparecen cada una de las personas que hablan, lo que dicen los locutores y en qué ambiente de ruido y en qué condiciones (lectura, habla espontánea) lo dicen.

Para poder detectar los lugares donde se producen cambios de locutor nosotros proponemos la medida XBIC, similar al Criterio de Información Bayesiana (BIC) [2], la cual mide la similitud entre dos segmentos adyacentes de voz para decidir si puede existir entre ellos un punto de cambio de locutor. La similitud se mide calculando las probabilidades cruzadas entre dos modelos de Markov entrenados con cada uno de los dos segmentos y evaluados en el segmento contrario.

Para aplicar el método de XBIC se construye un sistema basado en dos ventanas deslizantes sobre la señal a evaluar que realiza dos pasadas para encontrar los puntos de cambio. Un primer paso hace un barrido rápido sobre la señal hasta encontrar un posible punto. Un segundo paso se centra alrededor de este punto para fijar exactamente donde existe el cambio de locutor.

2. Segmentación de locutor

Entendemos por segmentación de locutor de la señal de voz la separación de esta en los diferentes locutores que aparecen, definiendo los puntos en qué empieza y acaba cada uno de ellos. Si el sistema, además de encontrar los puntos de cambio, determina qué locutores aparecen más de una vez, lo llamamos sistema de aglomerado de locutores (speaker clustering). Si, además, este determina la identidad de dichos locutores, se llama sistema de identificación de locutores.

En la presente publicación presentamos un algoritmo de segmentación de locutores. En esta misma área podemos encontrar otros métodos comúnmente usados para afrontar el mismo problema. Las técnicas basadas

en métricas ([1]) definen distancias acústicas para evaluar la similitud entre dos ventanas adyacentes de voz. Estas ventanas se deslizan uniformemente por toda la señal y la curva de medidas resultante es usada para encontrar cambios de locutor. Otro método comúnmente usado es el de BIC ([2]). Dada una ventana de trabajo y un punto de posible cambio, los segmentos de voz a ambos lados de ese punto se modelan con uno y dos modelos gausianos. La diferencia entre las dos alternativas comparada a la complejidad de entrenar más parámetros en el caso de dos modelos decide si el punto evaluado es un buen punto de cambio. La ventana de evaluación de va moviendo a través de toda la señal para encontrar todos los puntos de cambio existentes. Otro método es el basado en segmentación iterativa ([3],[4]), donde se empieza por un número inicial de locutores y de van añadiendo o eliminando locutores hasta llegar al punto óptimo y por tanto encontrando la segmentación.

Regularmente se han ido desarrollando evaluaciones llevadas a cabo por el National Institute for Standards and Technology (NIST) [5] para evaluar los avances en segmentación de locutor y en general en transcripción de señal de audio para indexado automático.

3. Criterio de segmentación BIC

El "Bayesian Information Criterion"(BIC) es un método bien conocido para hacer segmentación de locutor ([2]). Este permite la creación de sistemas de segmentación en tiempo real. Presentamos aquí brevemente la teoría ya que sienta la base para la medida que presentamos.

Dado $\Theta = \{\theta_j \epsilon \Re^d | j \epsilon 1 \dots N\}$, una secuencia de N vectores de características de dimensión d, que han sido extraídas de la señal que se desea segmentar.

El ´´Bayesian information criterion´´ en Θ es el logaritmo de probabilidad penalizado, según se puede ver:

$$BIC_\Theta = \mathcal{L} - \lambda P \tag{1}$$

Donde P es la penalización y λ es un parámetro de diseño libre dependiente de los datos que se modelen. Por defecto se pone a 1.

Dado un instante $\theta_i \epsilon \Theta$, podemos definir dos particiones de Θ: $\Theta_1 = \{\theta_1 \dots \theta_i\}$ and $\Theta_2 = \{\theta_{i+1} \dots \theta_N\}$ con longitudes N_1 y N_2.

Para tomar una decision de si existe un cambio de locutor en el instante θ_i consideraremos dos hipótesis: dos modelos independientes modelan mejor los datos en los dos lados del instante de cambio θ_i o bien un solo modelo es mejor para todos los datos. La mejor hipótesis se elige evaluando la siguiente expresión:

$$\Delta BIC = BIC_{H_0} - BIC_{H_1} = BIC_\Theta - (BIC_{\Theta_1} + BIC_{\Theta_2}) \tag{2}$$

En muchos de los sistemas presentes en la bibliografía ([2], [6], [7], [8]), los datos se modelan mediante una gausiana simple, con matriz de covarianzas plena de dimensión d. En estos casos la probabilidad $\mathcal{L}$ queda de la forma:

$$\mathcal{L} = -\frac{1}{2} N log|\Sigma| + NC \tag{3}$$

Donde $|\Sigma|$ es el determinante de la matriz de covarianza y C es una constante, $-\frac{1}{2}d(1 + log(2\pi))$.

Para algunas aplicaciones nos interesaría poder tener más flexibilidad al elegir el tipo de modelos a usar. Como se explica en [7], la probabilidad en la ecuación 1 también puede escribirse como:

$$\mathcal{L} = \mathcal{P}(\Theta_i | \lambda i) = \sum_{k=1}^{N} logp(\theta i(k) | \lambda_i) \tag{4}$$

Como se puede ver en [4], ΔBIC se puede expresar como la relación entre los logaritmos de las probabilidades de las dos hipótesis de la siguiente manera:

$$\Delta BIC(i) = \mathcal{P}(\Theta|\lambda) - \mathcal{P}(\Theta_1|\lambda_1) - \mathcal{P}(\Theta_2|\lambda_2) - \frac{1}{2}\Lambda K log N \tag{5}$$

Donde λ_1, λ_2 y λ representa los modelos entrenados mediante las particiones Θ_1, Θ_2 y Θ respectivamente. K es la diferencia en el número de parámetros entre λ_1 y λ_2, y Λ es una constante de diseño.

Decidimos que hay un cambio de locutor en un instante concreto si $\Delta BIC > 0$, significando que dos modelos se adaptan más a los datos que uno solo.

4. Medida de distancia probabilística para Modelos de Markov

Podemos ver como la expresión de BIC en términos de probabilidades se puede relacionar con la distancia introducida por L. Rabiner en [9],[10] como a una medida de distancia entre modelos ocultos de Markov. Rabiner definió la distancia entre dos modelos existentes de Markov como a combinación de las probabilidades de dos conjuntos de datos generados artificialmente mediante estos dos modelos.

Dados dos modelos HMM definidos por $\lambda_1 = (A_1, B_1, \pi_1)$ and $\lambda_2 = (A_2, B_2, \pi_2)$ consideramos que cada uno es capaz de generar un conjunto de datos $\Theta_1 = \{\theta_1^1, \dots, \theta_{N_1}^1\}, \Theta_2 = \{\theta_1^2, \dots, \theta_{N_2}^2\}$.

La distancia, escrita como $D(\lambda_i, \lambda_j)$, entre los dos modelos se define como:

$$D(\lambda_i, \lambda_j) = \frac{1}{N_j}\Big(\sum_{k=1}^{N_j} log\, p(\theta_j(k)|\lambda_i) - \sum_{k=1}^{N_j} log\, p(\theta_j(k)|\lambda_j)\Big) \tag{6}$$

Como la distancia $D(\lambda_i, \lambda_j)$ no es simétrica, necesitamos tomar en consideración el otro lado $D(\lambda_j, \lambda_i)$. La distancia se puede definir como:

$$D_{rab} = \frac{D(\lambda_i, \lambda_j) + D(\lambda_j, \lambda_i)}{2} \tag{7}$$

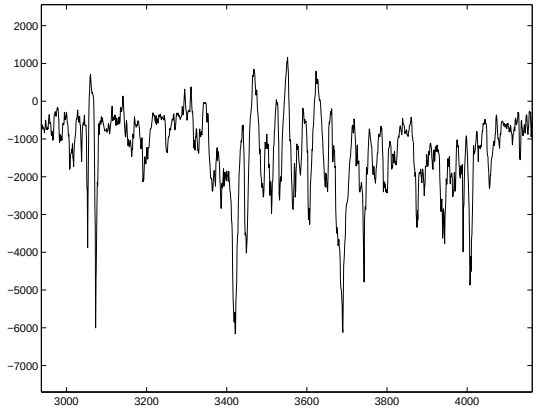


Figura 1: *Distancia XBIC para un segmento con diferentes locutores*

5. Método XBIC de probabilidades cruzadas

Así como la distancia en la ecuación 7 fue definida para comparar un par de modelos existentes usando datos generados artificialmente, nosotros proponemos comparar las dos secuencias de vectores de características calculando la distancia de dos modelos entrenados con estos.

Tomando como restricción que los dos segmentos tengan la misma longitud ($N_1 = N_2$) y reordenando los términos de la ecuación 7 podemos definir:

$$D'_{rab} = (\mathcal{P}(\Theta_1|\lambda_2)+\mathcal{P}(\Theta_2|\lambda_1))-(\mathcal{P}(\Theta_1|\lambda_1)+\mathcal{P}(\Theta_2|\lambda_2)) \quad (8)$$

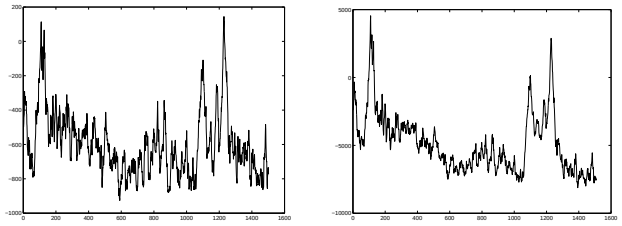
Donde en general:

$$\mathcal{P}(\Theta_i|\lambda_j) = \sum_{k=1}^{N_i} log\, p(\theta_i(k)|\lambda_j) \quad (9)$$

La ecuación 8 es similar a la expresión 5 sin embargo esta no tiene un término de penalización, el cual solamente desplaza a 0 el umbral de decisión del test de decisión. En ambos métodos estamos definiendo un test entre dos términos. El segundo término ($\mathcal{P}(\Theta_1|\lambda_1) + \mathcal{P}(\Theta_2|\lambda_2)$) es común a ambas ecuaciones y muestra lo bien que dos modelos independientes pueden representar los datos con los que han sido entrenados.

El primer término de las ecuaciones 8 y 5 mide lo bien que se pueden modelar los dos segmentos juntos. En la formulación del BIC esto se hace evaluando como un solo modelo que contenga ambos segmentos puede representar a estos. En la ecuación 5, se mide lo bien que los segmentos entrenados con cada modelo pueden representar al otro segmento. En ambos casos, como más similar sean los dos segmentos, mayor va a ser la probabilidad resultante.

Dado un segmento de voz, la distancia propuesta en la ecuación 8 tiene un valor cercano a 0 para instantes



BIC *-XBIC*

Figura 2: Puntuaciones para un segmento que contiene diferentes locutores

contenidos dentro de una región acústica similar y se vuelve negativo cuando estas son diferentes. Mínimos en esta distancia muestran los instantes donde es más probable encontrar cambios de locutor. En estos instantes la la medida calculada decrece de manera abrupta mayormente debido al término de probabilidades cruzadas, siendo el segundo término residual. Como nuestro interés está únicamente en encontrar cambios acústicos, podemos simplificar la ecuación y definir la medida XBIC como:

$$XBIC(i) = (P(\Theta_1|\lambda_2) + P(\Theta_2|\lambda_1)) \quad (10)$$

En la figura 1 podemos ver representada la medida XBIC(i) para un segmento de voz con varios locutores diferentes donde los segmentos Θ_1 y Θ_2 han sido desplazados a lo largo de la señal de entrada y la medida ha sido calculada en cada punto. La existencia de regiones de valor muy bajo en la gráfica indican los puntos de cambio de locutor. Definiendo un umbral apropiado podemos determinar que hay un cambio de locutor si $XBIC(i) < Thr_{XBIC}$

Hay algunas ventajas de usar la medida XBIC en vez de BIC, independientemente de la implementación usada. Primero, tal como podemos ver en la figura 5, si representamos las dos medidas dado un mismo segmento de voz que contenga varios cambios de locutor, el valor en los puntos de posible cambio se ve magnificado usando XBIC. De esta manera podemos reducir la tasa de falsa detección de cambios, la cual es un problema bien conocido de los sistemas BIC.

Por otra parte, la medida XBIC tiene en general menos requisitos computacionales que el BIC. XBIC solamente necesita entrenar dos modelos, mientras que BIC necesita entrenar tres (los dos modelos independientes y el modelo conjunto).

6. Algoritmo XBIC de segmentación de locutor

Para la aplicación de la medida XBIC presentada, se ha desarrollado un sistema se segmentación de locutor con dos características en mente. La primera es que el sistema pueda segmentar la señal secuencialmente, de man-

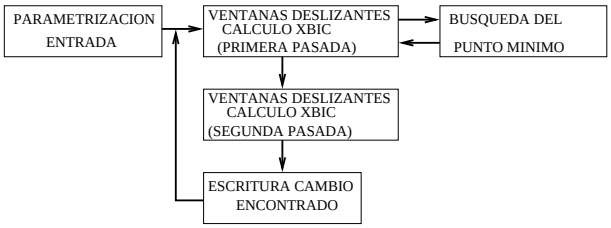


Figura 3: *Sistema de segmentación de locutor usando XBIC*

era que no sea necesario tener toda la señal antes de empezar a encontrar cambios de locutor. La segunda característica es poder segmentar señales que contengan locutores que hablen en intervalos muy cortos.

Para aplicar esta medida se utilizan dos ventanas deslizantes, de longitud fija e igual, conectadas por un punto donde realizamos la medida. Esta implementación guarda bastante similitud con los sistemas de segmentación basados en métricas.

Como vemos en la figura 3,el proceso se hace en dos pasadas de manera similar a [8], una rápida y otra más refinada. En la primera pasada se define la posición aproximada de cambio de locutor, que se resuelve exactamente en la segunda pasada. De esta manera el sistema puede analizar la señal de manera rápida sin perder resolución en los posibles puntos de cambio. El módulo de búsqueda del punto mínimo se encarga de evitar que mínimos locales de la función de medidas sean tomados como a posibles cambios de locutor y que por tanto puedan existir múltiples cambios muy cercanos unos de otros. En nuestro sistema definimos que como mínimo debe haber 1 segundo entre cambios de locutor. En ambas pasadas se define que un punto puede ser un cambio de locutor si es un mínimo y su valor XBIC es menor que un valor límite predefinido.Una vez se decide que un punto es un cambio de locutor, esta información se escribe a la salida y se continua con el análisis a partir del punto encontrado.

Podemos ver más en detalle el funcionamiento del algoritmo de doble pasada en la figura 4. Los dos segmentos de longitud T se encuentran en el punto de análisis, que es dónde se comprueba si hay un cambio de locutor. En un primero paso, se calcula la medida XBIC y se va desplazando todo el conjunto hacia adelante en intervalos de T_2 segundos. Cuándo el valor XBIC cumple la condición de cambio de locutor se hace un segundo paso dentro de un intervalo alrededor del punto detectado y de longitud T_2 y con un paso de desplazamiento mucho menor que antes (T_3), para encontrar el punto exacto de cambio. Esto se repite hasta el final de la señal buscando todos los cambios de locutor existentes.

El uso de este tipo de sistema en vez de los sistemas típicamente usados en implementaciones de BIC tiene dos ventajas principales. Por una parte la simplicidad y rapidez del sistema. Esta implementación avanza sobre

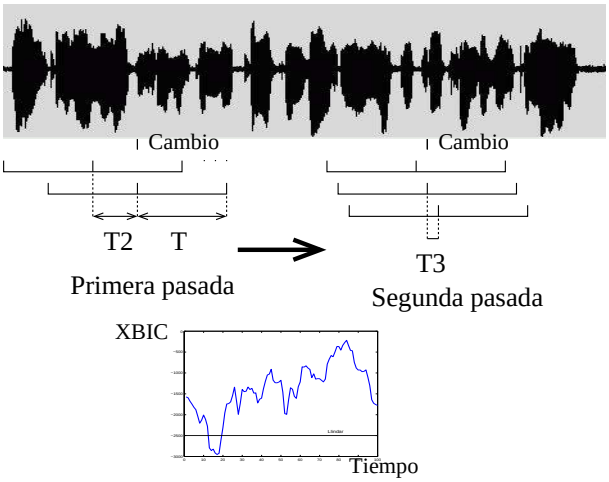


Figura 4: *Segmentación de locutor en dos pasadas usando la medida XBIC*

los datos con un paso prefijado y solamente retrocede cuando encuentra un punto potencial de cambio de locutor. Los sistemas típicos de BIC están basados en dos ventanas variables que aplican el criterio BIC de manera repetitiva en un segmento de señal semejante hasta que un cambio es encontrado.

Por otra parte, una vez se encuentra un cambio de locutor, el sistema propuesto reinicia el cálculo de la medida XBIC justo después del cambio, de manera que el tiempo de enmascarado (tiempo después de un cambio de locutor dentro del cual no se pueden detectar más locutores) es virtualmente 0. El diseñador del sistema puede imponer un tiempo mínimo si lo desea (en nuestro sistema este es de 1 segundo). En los sistemas típicos de BIC, después de un cambio de locutor existe un tiempo de enmascarado mínimo e igual al tiempo mínimo de ventana. Cualquier cambio de locutor en ese intervalo no se detectará y afectará al cálculo de los puntos de cambio posteriores.

7. Pruebas de segmentación de locutor

Para la evaluación del algoritmo de segmentación XBIC se han recogido y segmentado a mano 2,5 horas de señal de audio procedentes de diferentes transmisiones del Telenoticias de las cadenas de TV3 (3 telenoticias mediodía) y 3/24 (1 Telenoticias mañana). En todos los casos se han eliminado los anuncios emitidos en medio de estas transmisiones. Por otra parte también se presentan resultados comparativos de XBIC y BIC usando dos bases de datos en inglés, la 1996 HUB-4 y 1997 HUB-4 Evaluation Test Material.

Las señales están almacenadas en formato PCM, con 16 bits/muestra y a una frecuencia de muestreo de 16 KHz. Para ser tratados por el sistema de segmentación se han parametrizado utilizando parámetros MFCC de 16

muestras estáticas + 16 dinámicas de primer orden. El algoritmo de segmentación se ha aplicado en toda la señal pero para la evaluación sólo se ha tenido en cuenta la segmentación resultante en las áreas dónde aparecen locutores. Los modelos de Markov utilizados se forman con un estado compuesto por una gaussiana de matriz de covariances llena (todos los valores son diferentes de 0). Para cada iteración los modelos se entrenan mediante entrenamiento EM (Expectation Maximization).

Para implementar el algoritmo de BIC se ha implementado un sistema muy similar al propuesto en [8].

La métrica usada para evaluar los resultados tiene en cuenta dos tipo de error. Errores del tipo 1 debidos al hecho de encontrar más cambios de la cuenta, llamados precisión (PRC):

$$PRC = \frac{Num.\_cambios\_encontrados\_correct.}{numero\_total\_de\_cambios\_encontrados} \quad (11)$$

Por otra parte, los errores de tipo 2 debidos al hecho de no encontrar los cambios existentes, llamados ´´recall´´ (RCL):

$$RCL = \frac{Num.\_cambios\_encontrados\_correct.}{numero\_total\_de\_cambios\_existentes} \quad (12)$$

En el caso de funcionamiento óptimo las dos medidas llegarán al 100 %, y en el peor caso bajarán hasta el 0 %. Si invertimos estas medidas podemos definir también la False Rejection (FR) = 1 - RCL y la False Acceptance (FA) = 1 - PRC;

En la tabla 1 podemos ver los resultados para la base de datos de telenoticias en catalán:

Telenoticias	Medida RCL	Medida PRC
TV3 migdia 6-7-2004	50	67
TV3 migdia 7-7-2004	43.01	63.4
TV3 migdia 8-7-2004	53.04	64.89
3/24 TN matí 5-7-2004	46.47	41.27

Tabla 1: Resultados usando XBIC en diferentes programas en catalán

El análisis de los segmentos resultantes permite ver que hay tres tipos principales de problemas que causan que el sistema de segmentación produzca errores. Por una parte durante los periodos de transmisión dónde hay locutores hablando con música de fondo (cómo en la lectura de los titulares) el sistema tiende a confundir los locutores al no distinguir entre la música y la voz. Por otro lado, cuándo un locutor está hablando se producen falsas alarmas (se detectan cambios de locutor erróneos) si la noticia consta de varias piezas del mismo locutor pero grabadas en ambientes de ruido diferentes, o bien a través de equipos de grabación diferentes. Finalmente, en muchos casos en un telenoticias aparecen personas que al hablar en otra lengua son dobladas al catalán. En estos

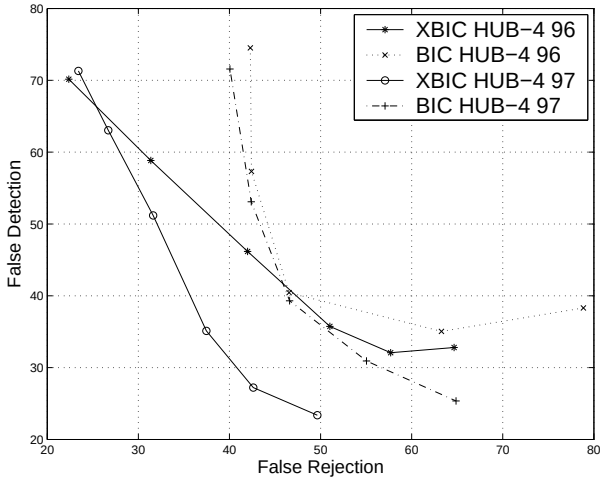


Figura 5: *Curva DET para BIC y XBIC para Broadcast News en Inglés*

casos tenemos más de una persona hablando al mismo tiempo que hace que el sistema se confunda.

Podemos ver como estos errores y otros causan que haya una diferencia importante entre los telenoticias de TV3 y el telenoticias del 3/24, en el cual la PRC decrece más de un 20 % respecto a los otros.

Por otra banda podemos ver en 7 la evolución de la curva de DET de BIC e XBIC para las bases de datos de Broadcast news en Inglés para diferentes valores del límite de segmentación y del valor λ de penalización.

Podemos ver como en HUB-4 97 hay una mejora substancial entre XBIC y BIC, y en HUB-4 96 esta mejora se mantiene para casi todos los valores. Por otra parte, podemos ver como las curvas de XBIC presentan en general un comportamiento más lineal, que permite que en casos reales se pueda elegir puntos de funcionamiento fuera del de EER sin que el acierto del sistema disminuya mucho.

8. Conclusiones

En esta publicación planteamos el problema del indexado de material sonoro y proponemos la indexado automática como una posible alternativa a los costosos sistemas manuales.

Como unos de los bloques dentro del indexado automático, presentamos la segmentación automática de locutor y presentamos una medida nueva (XBIC) que permite realizar segmentación en tiempo real y sin tener ninguna información previa sobre en número de locutores o su disposición. De manera similar a la medida BIC, la medida XBIC decide si en un punto concreto es mejor representar los datos provenientes de dos segmentos adyacentes mediante uno o dos modelos ocultos de Markov.

Esta decisión se toma calculando las probabilidades cruzadas entre cada unos de los dos segmentos y el modelo entrenado mediante el segmento contrario. Cuando en

el punto de cálculo existe un cambio de locutor el valor de XBIC decrece de manera abrupta, siendo fácil detectarlo mediante la selección de un valor de corte apropiado.

Presentamos también el sistema implementado para usar XBIC, el cual se basa en dos segmentos de señal de longitud constante que se trasladan por toda la señal calculando los valores y determinando los puntos de cambio en dos pasadas. Este sistema resulta ser muy simple y dar buenos resultados al mismo tiempo.

Aplicamos el sistema de segmentación a una base de datos grabada y segmentada manualmente, proveniente de telenoticias en lengua catalana y en dos bases de datos de Broadcast News en inglés, en las que comparamos XBIC con BIC. Los resultados obtenidos mejoran en general los obtenidos con BIC y dan a indicar la posible aplicación de este sistema como bloque para un sistema de indexado automático de locutor.

9. Referencias

[1] J.W. Hung, H.M. Wang, and L.S. Lee, "Automatic metric based speech segmentation for broadcast news via principal component analysis," in *ICSLP'00*, Beijing, China, 2004.

[2] S. Shaobing Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.

[3] X. Anguera and J. Hernando, "Evolutive speaker segmentation using a repository system," in *ICSLP'04*, Jeju Island, Korea, Oct. 2004.

[4] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.

[5] National Institute for Standards and Technology, ," `www.nist.gov/speech`.

[6] L. Perez-Freire and C. Garcia-Mateo, "A multimedia approach for audio segmentation in tv broadcast news," in *ICASSP'04*, Montreal, Canada, May 2004, pp. 369–372.

[7] S.E. Tranter and D.A Reynolds, "Speaker diarization for broadcast news," in *ODISSEY'04*, Toledo, Spain, May 2004.

[8] P. Sivakumaran, J. Fortuna, and A.M. Ariyaeeinia, "On the use of the bayesian information criterion in multiple speaker detection," in *Eurospeech'01*, Scandinavia, Sept. 2001.

[9] B.H. Juang and L.R. Rabiner, "A probabilistic distance measure for hidden markov models," AT&T Technical Journal 64, AT&T, Feb. 1985.

[10] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb. 1989.