UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

SEGMENTACION NO SUPERVISADA DE SEÑALES DE AUDIO Y VOZ

Guillermo Suárez Pedrero
Tutor: Javier Franco Pedroso

JUNIO 2018

SEGMENTACION NO SUPERVISADA DE SEÑALES DE AUDIO Y VOZ

AUTOR: Guillermo Suárez Pedrero
TUTOR: Javier Franco Pedroso

AUDIAS – Audio, Data Intelligence and Speech
Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2018

Resumen (castellano)

Este Trabajo de Fin de Grado se encuentra dentro del campo de la segmentación de audio, en concreto se comparan varias técnicas de segmentación no supervisada sobre aplicaciones tanto de segmentación de audio como de segmentación de locutores, de manera que pueda detectarse cualquier tipo de cambios, sean producidos por fragmentos de audio de diferente naturaleza o por distintos locutores. Esta etapa del procesado del audio resulta fundamental, ya que la correcta segmentación inicial del audio permitirá a etapas posteriores en diferentes aplicaciones clasificar y catalogar los diferentes tramos del audio con mayor facilidad y precisión.

El proyecto se ha realizado sobre la base de datos de Albayzín 2014, que contiene 24 ficheros de en torno a 4 horas de duración cada uno (aproximadamente 100 horas de audio en total), realizando un nuevo etiquetado mediante la combinación de un etiquetado según los cambios de clase acústica (entre música, voz y ruido) y otro etiquetado según los cambios de locutor, generando un etiquetado tanto de segmentación como de diarización.

Se ha partido de las características tímbricas MFCC-SDC extraídas con tres detectores GMM-UBM diseñados para detectar música, voz y ruido respectivamente [1]. Sobre estas características, se han aplicado diferentes métodos de segmentación (el Criterio de Información Bayesiano y la Razón de Verosimilitud Generalizada), implementados con una ventana temporal de tamaño fijo.

La medición de los resultados obtenidos se ha realizado tanto mediante la evaluación habitual de los errores de inserción y borrado, tanto como con una medida experimental basada en las distancias entre los puntos de cambio reales y los puntos de cambio detectados por los distintos algoritmos llamada Diarization Error Rate (DER), originalmente pensada para evaluar únicamente la diarización de locutores en ficheros de audio, pero aquí empleada sobre la segmentación en general.

Palabras clave (castellano)

Audio, música, voz, ruido, segmentación, diarización, detector, multimedia.

Abstract (English)

This Bachelor Thesis is framed within the area of audio segmentation, it specifically compares several techniques of unsupervised segmentation on both audio segmentation and speaker segmentation applications, so every type of change can be detected, either if they have been produced by audio fragments of different nature or by different speakers. That phase of the audio processing it's essential because the correct initial segmentation of the audio will be key in allowing the subsequent phases in different applications to classify and catalog the different audio sections with easier and more accurately.

The project has been tested with the database of Albayzín 2014, which contains 24 files of approximately 4 hours each (near to 100 hours of audio). A new labeling has been created combining the existing labeling based on the acoustic class (music, voice, noise) with the existing labeling based on the speaker changes, generating a labeling for audio segmentation and speaker segmentation jointly.

It starts off the MFCC-SDC timbre characteristics extracted with three GMM-UBM detectors designed to detect music, voice, and noise respectively. With these characteristics, different segmentation methods have been applied i.e.: tthe Bayesian Information Criterion and the Generalized Likelihood Ratio. These segmentation methods have been implemented with a time window of fixed size.

The measurement of the results obtained has been done through the usual evaluations of the errors on insertion/delete, as well as with an experimental metric based on the distances between the real change points and the points of change detected by the different algorithms called Diarization Error Rate (DER), originally designed to evaluate the speakers diarization in audio files, but here used with segmentation in general.

Keywords (English)

Audio, music, voice, noise, segmentation, diarization, detector, multimedia.

Agradecimientos

A Javier Franco por darme la oportunidad de realizar este Trabajo de Fin de Grado, así como toda la ayuda y facilidades que me ha prestado durante su desarrollo.

A Carlos Mora y Esther Jiménez, por enseñarme la cara divertida de las Matemáticas.

A Diego, por su amistad, que ha hecho de esta carrera un camino más llano y transitable.

A mis padres y hermanos, por ser unos fantásticos padres y excelentes hermanos.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

Motivación

Debido al acercamiento general a las tecnologías y al uso de las redes sociales por prácticamente la totalidad de la población, se da lugar a una gran cantidad de archivos multimedia, generados en una gran diversidad de escenarios: un vídeo grabado con el móvil durante un concierto, la reproducción de un programa de radio yendo en coche, la producción de un trabajo musical en un entorno libre de ruidos ni interferencias… Esto conlleva la coexistencia de señales de voz, ruido, y música, en un mismo fichero de audio.

La segmentación tanto de audio como de locutores se hace necesaria para que las aplicaciones que necesitan aislar el audio de una naturaleza concreta, como las de indexación de contenidos, funcionen correctamente sobre este tipo de archivos.

Para poder procesar correctamente cada clase acústica (identificar los fragmentos de audio en los que aparecen distintos locutores, extraer la armonía y el ritmo de una obra musical) es importante realizar una segmentación previa, para que el procesado de una clase acústica se realice sobre los tramos de audio que se corresponden con ella, y no sobre otras clases acústicas donde los resultados no tendrían sentido.

1.1 Objetivos

El propósito de este Trabajo de Fin de Grado es el desarrollo de diferentes métodos de segmentación de audio, abordando de forma conjunta la segmentación de audio por clases acústicas junto con la diarización de locutores, separando el audio en función tanto de los cambios de clase acústica como de los cambios de locutor presentes en los ficheros de audio.

También se ha creado una nueva base de datos a partir de los etiquetados existentes para segmentación y diarización, para poder evaluar esta segmentación tanto de clases acústicas como de locutores.

Por último, se busca estudiar la eficacia y fiabilidad de los métodos anteriores, empleando para ello la base de datos de Albayzín 2014, de unas 100 horas de audio en total, mediante diferentes medidas de error.

1.2 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1: Introducción**

  Este capítulo expone las causas que motivan el desarrollo de este Trabajo de Fin de Grado, así como el desarrollo y objetivos de este. También se especifica la estructura de esta memoria en diferentes capítulos.

- **Capítulo 2: Estado del arte**

  En este segundo capítulo se detalla qué es el audio, y se introduce la segmentación de audio, así como las técnicas empleadas para ello, tanto las utilizadas en este TFG como otras a destacar. Además, se explican las diferentes métricas para evaluar los resultados obtenidos.

- **Capítulo 3: Diseño**

  Aquí se detallan los diferentes métodos de segmentación empleados y su implementación, además de los problemas encontrados y su resolución.

- **Capítulo 4: Desarrollo**

  En este apartado se detalla la base de datos empleada para la realización de los diferentes experimentos, junto con las métricas de error para evaluar los resultados obtenidos.

- **Capítulo 5: Integración, pruebas y resultados**

  En este capítulo se especifican las pruebas que se han realizado, así como los resultados obtenidos en ellas.

- **Capítulo 6: Conclusiones y trabajo futuro**

  Como término de la memoria, se explican las conclusiones obtenidas en la realización de este Trabajo de Fin de Grado, además de plantear posibles mejoras aplicables a este proyecto.

2 Estado del arte

2.1 Segmentación de audio

La segmentación de audio consiste en la división del contenido de un fichero de audio en intervalos temporales. En este caso, la segmentación se genera según se detecten tanto cambios de clase acústica, como cambios de locutor dentro de un mismo tramo.
Los sistemas que determinan la clase acústica concreta de un tramo de audio acertarán con mayor facilidad si se ha segmentado previamente en fragmentos en los que no sucede ningún cambio de clase.

Según se hayan entrenado o no los algoritmos de segmentación con bases de datos de manera previa a la segmentación de los ficheros de audio de interés, se distinguen la clasificación supervisada y la clasificación no supervisada, respectivamente.

Los métodos analizados en este Trabajo de Fin de Grado son métodos de segmentación no supervisada, ya que se busca segmentar audio sin ningún tipo de información previa sobre este, a diferencia de sistemas de extracción de armonía/ritmo o similares, donde sí resultaría útil un entrenamiento con ficheros concretos (de música, para este ejemplo).

2.1.1 Audio

Las señales de audio son señales analógicas que representan señales sonoras. Estas señales sonoras son escuchadas por el ser humano, que las distingue entre sí debido a su nivel de intensidad acústica, su tono, y su timbre. Estos atributos se corresponden con las propiedades físicas de intensidad, frecuencia fundamental y forma de onda respectivamente. Estas cualidades se definen como [2]:

- **Volumen (nivel de intensidad sonora).** Depende de la amplitud de la señal. Se define como la potencia transmitida por la onda por unidad de superficie, $I = P/A$ (en W/m^2). Se mide mediante el nivel de presión sonora (SPL por sus siglas en inglés), generalmente expresado en decibelios (dB), según $I(dB) = 10\log(I/I_0)$, donde I_0 es el nivel de intensidad mínimo que es capaz de escuchar el ser humano (umbral de audición), siendo para 0 dB de nivel igual a 10^{-12} W/m^2.

- **Tono o frecuencia (pitch).** Es la propiedad que nos permite diferenciar un sonido grave de otro agudo. Mide la cantidad de vibraciones por segundo de una onda en Herzios (Hz). Cuanto más agudo sea el sonido, más Herzios tendrá su frecuencia. El ser humano es capaz de percibir sonidos con frecuencias desde 20Hz hasta 20kHz.

- **Timbre.** A igualdad de volumen y tono, el timbre permite diferenciar diferentes sonidos. Cada fuente de emisión sonora tiene un timbre particular, caracterizado por la composición de su espectro armónico.

2.1.2 Clases acústicas

En la base de datos empleada en este proyecto, el etiquetado contempla tres tipos de señales de audio, pudiendo distinguir tres clases acústicas diferentes, como son la música, la voz hablada, y el ruido, y combinaciones entre ellas. Se ha respetado esta clasificación por compatibilidad con la base de datos, además de ser una clasificación bastante común y coherente.

Este Trabajo de Fin de Grado busca segmentar el audio según la aparición o desaparición de nuevos tramos pertenecientes a una de esas clases, además de segmentar también los tramos de voz en función de los locutores que haya (diarización). Se pueden clasificar y definir de la siguiente forma [3]:

- **Voz.** El espectro de frecuencias de las señales de voz se extienden hasta los 8kHz (aunque con los primeros 4kHz la señal ya es suficientemente inteligible). Como no solo se quiere segmentar en tramos de voz, sino fraccionarlos según la aparición de diferentes locutores, se deben tener en cuenta también características propias de cada locutor, que generalmente se extraen mediante los Coeficientes Cepstrales en Frecuencias de Mel (MFCCs).

- **Música.** La música tiene un contenido frecuencial mayor, y se extiende hasta los 20kHz. La presencia de armonía, patrones rítmicos, y en general una mayor cantidad de energía entre otras cosas la diferencian como clase acústica de la voz.

- **Ruido.** El ruido se extiende por todo el espectro de frecuencias, y se caracteriza por lo aleatorio de su naturaleza. En la base de datos empleada se considera dentro de la clase acústica “ruido” a cualquier elemento que no sea ni voz ni música, como pudiera ser una conversación de fondo o el ruido de una muchedumbre hablando.

2.1.3 Segmentación supervisada

La segmentación supervisada es aquella segmentación que emplea una base de datos etiquetada para la construcción del modelo de segmentación.

2.1.3.1 *Métodos de segmentación basados en modelos estadísticos*

Para cada clase acústica se define un modelo estadístico, generalmente Modelos de Mezclas Gaussianas (GMM). Estos modelos combinan diferentes distribuciones gaussianas, donde los parámetros de cada distribución gaussiana G_i son su media (μ_i) y su matriz de covarianzas (Σ_i), y se determinan a partir de pruebas sobre ficheros de entrenamiento.

La segmentación de audio se realiza comparando cada ventana con los diferentes modelos estadísticos, y asignándole la clase correspondiente al modelo que más se le asemeje, de manera que si dos ventanas consecutivas se corresponden con diferentes modelos, se propondrá un cambio de clase acústica entre ambas. Este sistema aumenta el coste computacional con respecto a los métodos basados en energía, pero supone mucha más precisión en la segmentación frente a los métodos basados en energía [4].

2.1.4 Segmentación no supervisada

La segmentación no supervisada no construye el modelo de segmentación a partir del etiquetado de la base de datos, sino que únicamente utiliza las etiquetas para evaluar el rendimiento del sistema.

2.1.4.1 *Métodos de segmentación basados en la energía*

Estos métodos detectan periodos de silencio en los ficheros, midiendo y umbralizando la energía del audio. El método propone los cambios de segmento en estos periodos de silencio. Una aproximación muy simple a este proceso sería una puerta de ruido.

Pese a que resulta muy sencillo de implementar, los cambios propuestos no tienen una relación directa con los cambios acústicos, ya que sería incapaz de detectar cuando una persona empieza a hablar si hay música de fondo, o podría detectar como cambio de segmento, un silencio perteneciente a alguna parte de una obra musical.

2.1.4.2 *Métodos de segmentación basados en la distancia*

Los métodos de segmentación basados en la distancia se pueden considerar como una variante de los métodos basados en modelos estadísticos, ya que hacen uso también de los modelos estadísticos para cada ventana de audio, pero emplean un criterio más refinado para decidir si ha sucedido un cambio de clase acústica.

Este tipo de métodos deciden si hay un nuevo segmento a partir de la similitud entre los contenidos de dos ventanas de audio consecutivas. Estas ventanas se modelan cada una mediante una distribución gaussiana, y se calcula la similitud entre los contenidos de ambas sobre todo el audio, creando una curva de distancias entre ventanas para todo el audio, siendo los puntos de cambio de clase estos donde la curva de similitudes se haga máxima, o donde estos máximos superen un cierto umbral.

A la hora de implementar estos métodos, se debe tener en cuenta el tamaño de las ventanas de que se va a emplear, el tamaño de la distancia recorrida por ellas al desplazarse por el fichero de audio (pudiendo solapar diferentes tramos, o no), además de la función que se escoja para calcular la distancia. Algunas de las funciones de distancia más frecuentes son:

Distancia Kullback-Leibler

La Distancia Kullback-Leibler, o Entropía Cruzada Relativa, propone que partiendo de las funciones de densidad de probabilidad (FDP) P_A y P_B de dos ventanas de audio adyacentes, puede calcularse la métrica como la tasa de bits acumulada al codificar *B* con un código diseñado para codificar de manera óptima la ventana *A* [5]. Se formula como:

$$KL(A,B) = \int_X P_A(x)\ \log\left(\frac{P_A(x)}{P_B(x)}\right)dx$$

Como no es una expresión simétrica, y por tanto no puede considerarse estrictamente una distancia, se define la métrica *KL2* como:

$$KL2 = KL(A,B) + KL(B,A)$$

Si introducimos en esta definición la igualdad para la distancia Kullback-Leibler descrita más arriba, obtenemos:

$$KL2(A,B) = \frac{1}{2}\int_X \left(P_A(x)\log\frac{P_A(x)}{P_B(x)} + P_B(x)log\frac{P_B(x)}{P_A(x)}\right)dx$$

Siendo *KL2* ya una función simétrica, y por tanto puede hablarse de distancia propiamente dicha. Si en lugar de las funciones de densidad de probabilidad P_A y P_B empleamos distribuciones gaussianas para representar las ventanas, se obtiene una forma más compacta para esta métrica, en función de los parámetros de cada distribución gaussiana:

$$KL2(A,B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (\mu_A - \mu_B)^2 \cdot \left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right)$$

Donde μ y σ son la media y la desviación típica de cada una de las dos distribuciones gaussianas. Cuando *A* y *B* tienen la misma función de densidad de probabilidad, la métrica *KL2* es igual a cero.

Distancia euclídea

El sistema más simple para obtener una métrica de distancia obtiene la métrica para la distancia euclídea comparando las medias de las distribuciones gaussianas de dos ventanas de audio adyacentes, G1(μ_1,σ_1) y G2(μ_2,σ_2).
Esta distancia se calcula como:

$$DistEUCL = (\mu_1 - \mu_2)^T \cdot (\mu_1 - \mu_2)$$

Criterio de Información Bayesiano

El Criterio de Información Bayesiano (BIC en adelante) [6], emplea para cada par de ventanas consecutivas, tres modelos gaussianos, siendo estos uno por cada ventana G1(μ_1,σ_1) y G2(μ_2,σ_2) respectivamente, y un tercer modelo, G(μ,σ) para ambas ventanas conjuntamente.
La métrica ahora se obtiene como:

$$distBIC = BIC\{G1\} + BIC\{G2\} - BIC\{G\}$$

Donde:

$$BIC\{G\} = -\frac{N\,log(|\Sigma|)}{2} - \frac{\lambda\left(d + \frac{d(d+1)log(N)}{2}\right)}{2} - \frac{dNlog(2\pi)}{2} - \frac{N}{2}$$

Por tanto, *distBIC* resulta:

$$distBIC = \frac{N\,log(|\Sigma|)}{2} - \frac{N_1\,log(|\Sigma_1|)}{2} - \frac{N_2\,log(|\Sigma_2|)}{2} - \frac{\lambda d}{2} - \frac{\lambda d(d+1)}{4}(\log(N_1) + \log(N_2) - \log(N))$$

Siendo *N*, *N1*, y *N2* el número de muestras de cada ventana, λ un factor de ajuste y *d* el número de características en el vector de características de entrada.

Razón de Verosimilitud Generalizada

La Razón de Verosimilitud Generalizada (GLR por sus siglas en inglés), es en esencia un método muy parecido al método BIC, pero modificado para tener un menor coste computacional. Emplea de igual manera los tres modelos gaussianos para dos ventanas consecutivas de audio, pero la distancia se obtiene como:

$$DistGLR = w(2\log(|\Sigma|) - \log(|\Sigma_1|) - \log(|\Sigma_2|))$$

Donde *w* es el tamaño de la ventana en muestras.

Distancia Gish

Esta distancia es una variación de la Razón de Verosimilitud Generalizada. Para calcularla, la función GLR se divide en dos partes, λ_{cov} y λ_{mean}, y la parte que depende del fondo se ignora, mediante la ecuación [7]:

$$Dist_{GISH}(i,j) = -\frac{N}{2}\log\left(\frac{|S_i|^{\alpha}|S_j|^{1-\alpha}}{|W|}\right)$$

Donde:
S_i y S_j son las matrices de covarianza para cada segmento.
La variable α se calcula como $\alpha = \frac{N_1}{N_1+N_2}$
Y W es el promedio ponderado de muestras, calculado como $W = \frac{N_1}{N_1+N_2}S_1 + \frac{N_2}{N_1+N_2}S_2$

3 Diseño

3.1 Diseño

En este Trabajo de Fin de Grado se han estudiado dos de los métodos de segmentación no supervisada descritos en el estado del arte, concretamente el Criterio de Información Bayesiano y la Razón de Verosimilitud Generalizada (ambos funciones para el cálculo de distancias).

Aunque el desarrollo matemático de ambos criterios ya ha sido descrito en el Estado del Arte, se van a exponer a continuación las implementaciones usadas en este Trabajo de Fin de Grado.

3.1.1 Criterio de Información Bayesiano (BIC)

A partir de las características de dos ventanas de audio contiguas, X_1 y X_2, y la ventana formada por ambas, X, se calculan las variables Σ, Σ_1 y Σ_2. Hay dos formas de calcularlas:

Si la matriz de covarianza se ha definido como *'diag'* se calculan las variables anteriores como matrices de valores nulos excepto la diagonal principal, cuyos valores son los de la diagonal principal de la matriz de varianzas de los vectores de características.

```
sigma_i = diag(var(X_i, 1));
```

Si la matriz se ha definido como *'full'*, el cálculo se realiza como la covarianza de los vectores de caracteristicas.

```
sigma_i = cov(X_i, 1);
```

El resto de la implementación consiste en las operaciones ya descritas en el Estado del Arte. Se calcula el valor de P según la siguiente igualdad:

$$P = \frac{\log(N)}{2}\left(d + \frac{d(d+1)}{2}\right)$$

Y se obtiene finalmente la distancia entre las dos ventanas mediante:

$$distBIC = \frac{Nlog(|\Sigma|)}{2} - \frac{N_1 \log(|\Sigma_1|)}{2} - \frac{N_2 \log(|\Sigma_2|)}{2} - \lambda P$$

Donde N, N_1 y N_2 son los tamaños de las ventanas, d el número de características, y λ un factor de ajuste.

Una vez se han obtenido todas las distancias entre ventanas, en la secuencia de valores de distancias, se calcula una altura mínima y una distancia mínima, de manera que dos puntos consecutivos de esta secuencia que se encuentren lo suficientemente cerca como para no superar ambos umbrales, no se considerarán como dos puntos de cambio.

El umbral de altura mínima se ha definido como cero, ya que así únicamente se podrán tener en cuenta los puntos de cambio cuyo valor en el vector de distancias sea mayor que cero:

$$min_{HEIGHT} = 0$$

Al variar el parámetro λ, se consigue que todos los valores del vector de distancias aumenten o disminuyan por igual, de manera que se aumenta o se disminuye también el número de cambios de clase acústica que el algoritmo propone.

La distancia mínima se ha definido como el mínimo entre la mitad del tamaño de la ventana *w* y la longitud del vector de valores de distancias.

$$min_{DISTANCE} = \min(w/2 \ , \ vectDistances)$$

Los máximos de este vector de distancias que superen ambos umbrales, se considerarán como puntos en los que hay un cambio de clase acústica.

3.1.2 Razón de Verosimilitud Generalizada (GLR)

Al igual que en el Criterio de Información Bayesiano, se parte de los vectores de características X_1 y X_2, y X. Se calculan las medias μ_1, μ_2 y μ de cada vector, y se calculan las matrices de covarianzas de la misma forma que en el apartado anterior, teniendo también las opciones de matriz de covarianzas completa, o únicamente con la diagonal no nula.

La implementación se completa con el uso de la función de matlab *ecmnobj*, que calcula el valor de la función de verosimilitud logarítmica negativa observada sobre los datos, dadas las estimaciones para la media y la covarianza de los datos.

Esto se realiza con las tres ventanas y sus parámetros, obteniendo la distancia entre las dos ventanas como:

$$distGLR = \text{ecmnobj}(X, \mu, \Sigma) - \text{ecmnobj}(X_1, \mu_1, \Sigma_1) - \text{ecmnobj}(X_2, \mu_2, \Sigma_2)$$

La altura mínima se ha definido como la media del vector de distancias, más tres veces su desviación típica:

$$min_{HEIGHT} = \text{mean}(vectDistances) + 3\text{std}(vectDistances)$$

La distancia minima se ha definido de igual manera que en el apartado anterior, como el mínimo entre la mitad del tamaño de la ventana *w* y la longitud del vector *vectDistances*.

$$min_{DISTANCE} = \min(w/2 \ , \ vectDistances)$$

Los máximos de la secuencia que contiene los valores de las distancias que superen ambos umbrales, se considerarán como puntos en los que hay un cambio de clase acústica.

4 Desarrollo

4.1 Base de datos

La base de datos de la que ha partido este proyecto es la de Albayzín 2014, con un contenido dividido en 24 ficheros de aproximadamente 4 horas de duración cada uno (unas 100 horas de audio en total), de la que se disponía de dos etiquetados: un primer etiquetado según la intervención de diferentes locutores en los segmentos en los que hay voz, y otro etiquetado según la clase acústica (música, voz, o ruido y combinaciones de las tres) que contenga cada segmento. Esta base de datos a su vez está formada por la combinación de tres bases de datos [8]:

- Una base de datos del canal de televisión catalán 3/24 TV de noticias, propuesto ya en Albayzín 2010, y grabada por la Universidad de Cataluña en 2009. Contiene 87h de grabaciones en las que se puede encontrar voz hablada el 92% del tiempo, música el 20% del tiempo, y ruido de fondo el 40% del tiempo.
- El segundo conjunto de datos es la base de datos de Radio Aragón, de la Corporación Aragonesa de Radio y Televisión (CARTV), que fue empleada en Albayzín 2012.
- El último conjunto de datos está compuesto por sonidos ambientales pertenecientes a Freesound.org y a HuCorpus, mezclados con segmentos pertenecientes a las dos bases de datos anteriores.

Todos los ficheros de audio empleados se encontraban en formato PCM, mono, con resolución de 16 bits y una frecuencia de muestreo de 16kHz.

Cada etiqueta para el fichero de voz indica el comienzo y fin de cada aparición de un locutor, de manera que las zonas de silencio no aparecen en etiquetas específicas, y en el caso de que varios locutores hablaran al mismo tiempo, habría una o más etiquetas que compartirían parte o la totalidad de la duración de otra etiqueta.

En los ficheros de etiquetado de clases acústicas esto no sucede, ya que cada cambio de clase por haber solapamiento entre dos clases acústicas (por ejemplo voz y música) se indica mediante una nueva etiqueta, de manera que las etiquetas son disjuntas. En el ejemplo, habría una etiqueta de voz, otra etiqueta de voz + música en la zona donde se produce el solapamiento, y otra etiqueta de solo música, en la zona desde donde termina el solape, hasta donde termina la música.

Se ha realizado nuevos ficheros de etiquetas combinando estos dos etiquetados, obteniendo una base para evaluar tanto la segmentación por clases acústicas como la diarización de locutores, de manera simultánea. El etiquetado contempla cualquier solape (ya sea entre locutores, entre clases acústicas, o entre locutores y clases acústicas) como segmentos diferentes, de manera que cualquier comienzo o final de cualquier clase acústica considerada en los dos tipos de etiquetado de los que se partía, se considera un nuevo segmento disjunto, en el etiquetado creado.

4.2 Medidas de error

La medición de los resultados obtenidos se ha realizado tanto mediante la evaluación habitual de los errores de detección (errores de inserción y borrado), como con una medida experimental mediante el empleo de la Diarization Error Rate (DER), originalmente pensada para evaluar únicamente la diarización de locutores, pero aquí empleada sobre la segmentación en general en ficheros de audio. Esta medida se basa en las distancias entre los puntos de cambio reales y los puntos de cambio detectados por los distintos algoritmos empleados.

4.2.1 DER

Esta métrica de error, pensada para evaluar la diarización de locutores, calcula la fracción del tiempo que no se ha asignado correctamente a un locutor, o a la clase sin locutor. En este proyecto se ha utilizado esta métrica para evaluar sobre audio en general, y no únicamente ficheros de voz, de manera que lo que mide es si los cambios de clase acústica o locutor, coinciden con los cambios reales correspondientes en el etiquetado. Para medir esto, se ha etiquetado cada segmento como una clase distinta, tanto en el fichero de etiquetas reales de la base de datos, como en los ficheros de etiquetas generadas, ya que no se evalúa la etiqueta de clase, sino que únicamente se comprueba que los segmentos generados y los segmentos reales estén alineados.

Para ello se ha usado la implementación de NIST (National Institute of Standards and Technology) empleada en las evaluaciones de 2006 (concretamente se ha empleado el script de perl *md-eval-v21*) [9] del que se ha extraído para la totalidad de los ficheros empleados en la base de datos únicamente el Error de Diarización Global (Overall Speaker Diarization Error) mediante otro script en perl diseñado a tal fin.

La Tasa de Error de Diarización (DER) se calcula como:

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot (\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}}$$

Donde *S* es el número total de segmentos. $N_{ref}(s)$ es el número de locutores que hablan en el segmento *s*, y $N_{correct}(s)$ indica el número de locutores que coinciden entre los ficheros de hipótesis y de referencia. Como en este proyecto no se identifica a los locutores en sí, y son segmentos disjuntos, el número máximo posible será 1 en cada segmento, de manera que el error para cada segmento puede ser 0, si la detección propuesta coincide con el etiquetado real, ó 1, si la detección es errónea.

Para generar la puntuación resultante de esta métrica, se normaliza entre el tiempo total analizado, obteniendo una medida porcentual para el DER.

4.2.2 Errores de detección

Otra forma de evaluar los resultados obtenidos, es mediante los errores de detección. Se evalúan todos los puntos en los que hay cambios reales. Cada cambio real que el sistema haya detectado correctamente, se considera un acierto, y para los casos en los que se detecte un cambio que no ha sucedido en realidad, o que no se detecte un cambio que sí ha sucedido, se definen los siguientes dos tipos de error:

Error de Falsa Alarma (FA, o falsa aceptación). Tiene lugar cuando en un instante temporal se considera cierto que hay un cambio de clase acústica, siendo esta decisión incorrecta. Los errores de Falsa Alarma sumados a los aciertos, hacen el total de cambios detectados por el sistema.

Error de Falso Rechazo (FR, o falso negativo). Sucede al no detectar un cambio de clase acústica, que sí que ha tenido lugar en el audio. Los errores de Falso Rechazo, junto con los aciertos, suman el total de cambios reales en el audio a analizar.

Estos errores nos sirven como indicadores de la segmentación llevada a cabo, ya que una tasa de error de falsa alarma demasiado elevada, nos señala que el sistema está sobre-segmentando el fichero de audio.

De la misma manera, una tasa de error de falso rechazo demasiado elevada nos indicará que el audio está siendo sub-segmentado. Además, por ser el error de falso rechazo los cambios no detectados dentro del total, es complementario de la tasa de aciertos, ya que la suma de ambos resulta igual al número total de cambios reales. Por tanto, cuanto menor sea la tasa de falso rechazo, mayor porcentaje de aciertos tendrá el sistema.

En general es preferible la sobre-segmentación, enmendable con una etapa posterior al sistema que reagrupe estos segmentos, que la sub-segmentación, que pasaría inadvertida y se clasificarían como un mismo tramo fragmentos de audio con diferentes clases acústicas.

En las siguientes gráficas se puede observar un caso de sobre-segmentación (alto porcentaje de falsas alarmas) y otro de sub-segmentación (alto porcentaje de falsos rechazos).

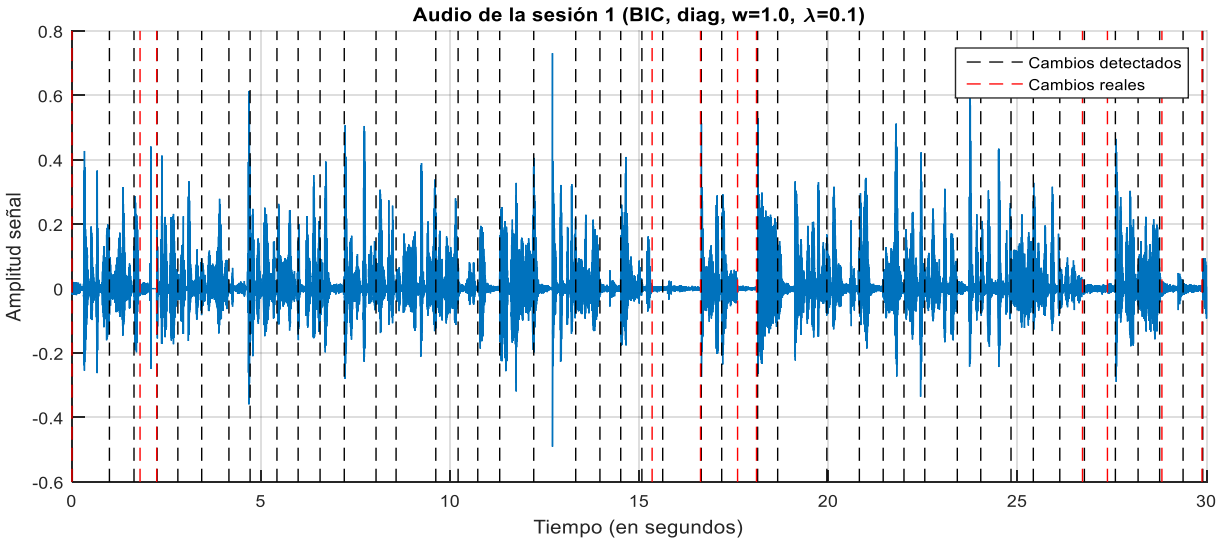

Figura 4-1: Tramo de audio sobre-segmentado

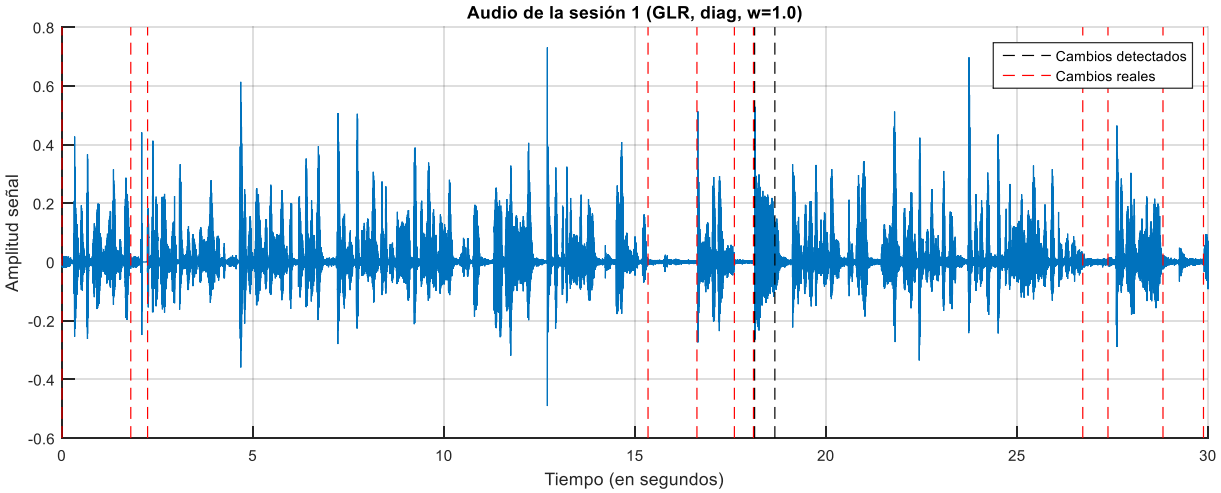

Figura 4-2: Tramo de audio sub-segmentado

4.2.3 Medida FA + FR

Esta medida de error compuesta, suma el porcentaje de falsas alarmas obtenido, con el de falsos rechazos. Como se comentaba en el apartado anterior, la tasa de falsos rechazos es complementaria al número de aciertos, por lo que, aunque esta métrica no evalúe de forma directa el número de aciertos, una cantidad elevada de falsos rechazos implica un menor porcentaje de aciertos, y se verá reflejado al tener un error mayor.

Para evitar la sobre-segmentación, situación que se daría si se penalizaran únicamente los falsos rechazos (ya que se favorecería así la detección de los cambios de clase a cualquier costa), se suma a esta tasa, la de falsas alarmas, de manera que un número elevado de detecciones erróneas, dará lugar a un mayor error con esta métrica.

4.2.4 Medida (FA + FR)*FR

Esta otra medida de error compuesta, multiplica la anterior por el porcentaje de falsos rechazos, para penalizar la sub-segmentación, al aumentar el error medido si aumenta el número de cambios no detectados entre segmentos.

Pese a que penaliza la sub-segmentación, cuando FR es cero el error se considera nulo, por lo que los mejores resultados con esta métrica serán aquellos en los que FR sea cero, independientemente del número de falsas alarmas. Esto favorece la sobre-segmentación, ya que al proponer el sistema gran cantidad de cambios, se minimiza la cantidad de falsos rechazos, aunque la gran mayoría de estos cambios no sean ciertos (falsas alarmas).

5 Integración, pruebas y resultados

Las pruebas llevadas a cabo con el sistema de segmentación se pueden dividir en tres etapas. En la primera fase (fase de entrenamiento), se han evaluado los dos métodos del sistema sobre un rango de valores para cada uno de sus parámetros. Para ello se han empleado 12 ficheros de audio de 15 minutos de duración cada uno (3 horas de duración en total).

En la segunda etapa (validación), se han seleccionado la mejor combinación de parámetros para cada método, tanto en versión *'full'* como en versión *'diag'*, y estas cuatro combinaciones se han evaluado sobre otros 12 ficheros diferentes de 15 minutos de duración.

Ya en la última fase (test), si estos parámetros se han considerado válidos según los resultados obtenidos en el entrenamiento y la validación, se evalúan los resultados para el resto de audio de la base de datos, compuesta por 24 ficheros de algo menos de 4 horas cada uno, tras descartar las 6 horas (3 horas para entrenamiento y otras 3 horas para validación) ya empleadas.

5.1 Fase de entrenamiento

En esta primera fase, se han evaluado los 15 primeros minutos de los 12 primeros ficheros de la base de datos de Albayzín 2014, generando un archivo por cada combinación de parámetros usada con el etiquetado asociado a todos los cambios de segmento propuestos. Los rangos de valores evaluados se muestran en las siguientes tablas:

BIC	Valores λ (paso de 0,1)		Tamaño ventana (s) (paso de 0,5 s)	
	min	max	min	max
full	0	2	1	10
Diag	0	1	1	10

Tabla 5.1-1: Valores de BIC evaluados en el entrenamiento

GLR	Tamaño ventana (s) (paso de 0,5 s)	
	min	max
full	0,5	5
diag	0,5	5

Tabla 5.1-2: Valores de GLR evaluados en el entrenamiento

Al ser 0,5s la distancia máxima entre un cambio real y un cambio detectado a partir de la cual este último deja de considerarse un acierto, el tamaño de la ventana se ha evaluado inicialmente entre 1s y 5s, con 0,5s como paso.

Como los resultados obtenidos con la métrica DER parecían apuntar a que, reduciendo el tamaño de ventana para el método GLR, y aumentando el tamaño de ventana y el de λ para el método BIC, se reducía el error (ver Figuras 5-1, 5-2 y 5-3), se ampliaron los rangos de evaluación, según lo descrito en las tablas anteriores (resultados completos en Anexo A.1).

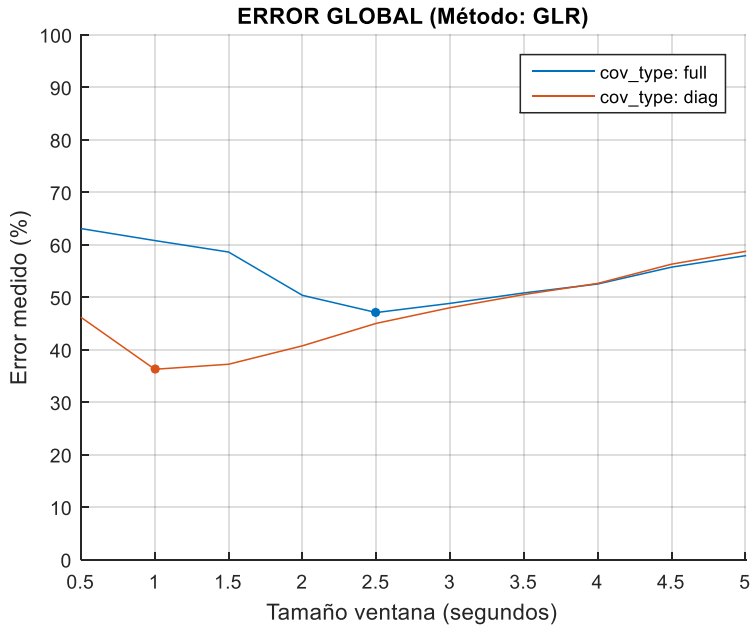

Figura 5-1: Error obtenido para los valores de ventana evaluados para GLR

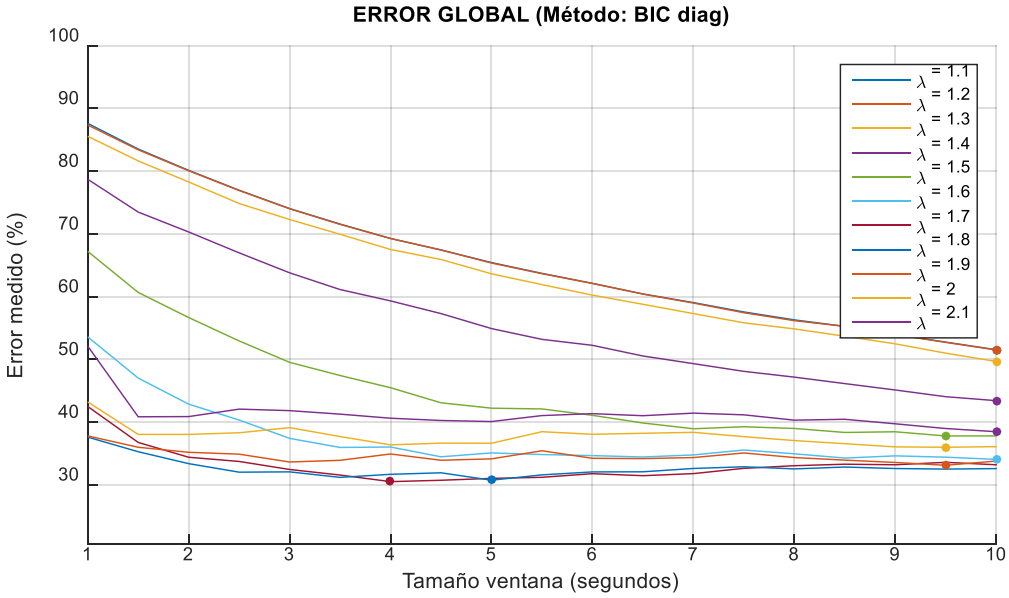

Figura 5-2: Error obtenido para los valores de ventana y λ evaluados para BIC diag

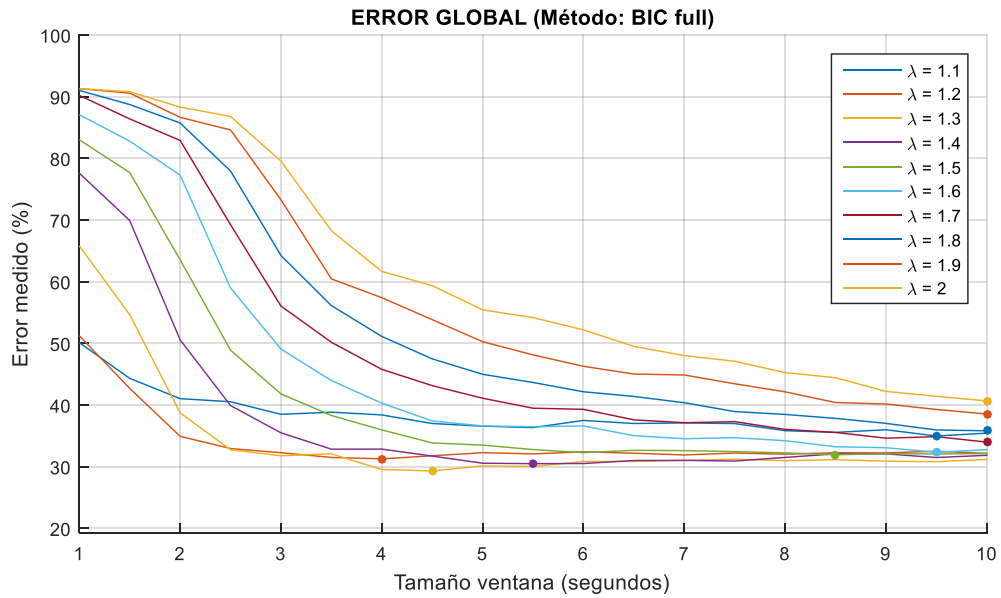

Figura 5-3: Error obtenido para los valores de ventana y λ evaluados para BIC full

Aunque en el método BIC, tanto para matrices de covarianza de tipo completo o diagonal, la tendencia sigue siendo de disminuir el error al aumentar el tamaño de ventana, emplear ventanas de un tamaño tan grande podría implicar que dos cambios de clase lo suficiente cercanos entre sí, no fueran detectados al encontrarse dentro de un mismo enventanado. Además, las parejas de parámetros que minimizan el error se encuentran en torno a 4 o 5 segundos para el método BIC con matriz de covarianza de tipo diagonal.

Como también se apreciaba una tendencia a disminuir el error a medida que aumenta λ en el método BIC con ventana de covarianzas de tipo completo, se evaluó este parámetro hasta λ=2, obteniendo un error mucho más bajo, para ventanas de menor tamaño que las de la gráfica anterior (ver Figura 5-4).

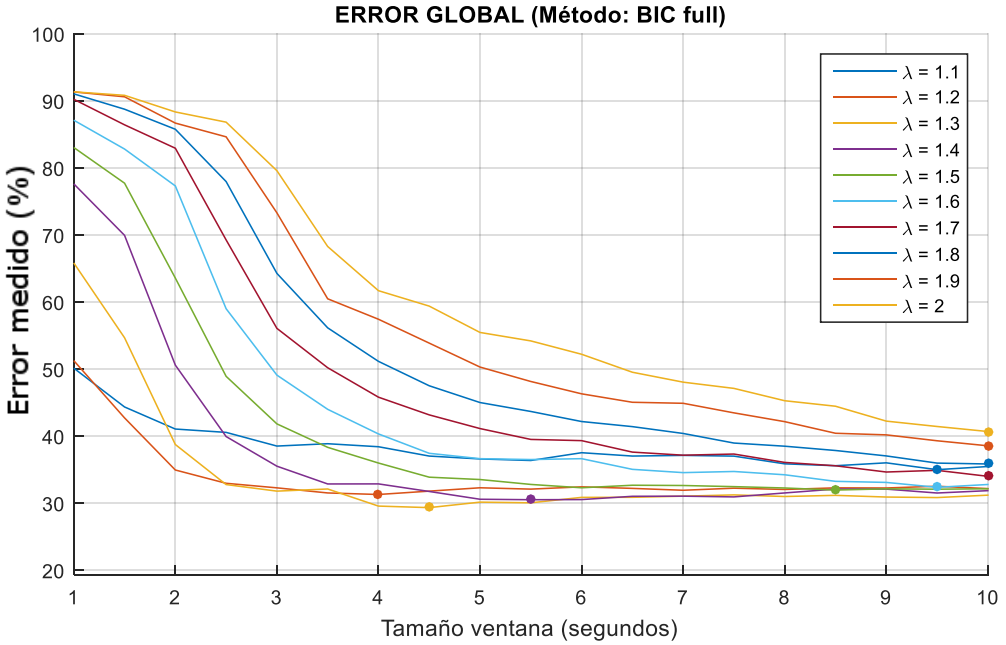

Figura 5-4: Error obtenido para los nuevos valores de λ evaluados para BIC full

Ahora el DER se minimiza para tamaños de ventana entre 4 y 6 segundos para el método BIC con matriz de covarianzas completa. Las combinaciones que mejores resultados devolvieron en la fase de entrenamiento fueron:

		ventana	λ
GLR	*full*	2,5	
	diag	1,0	
BIC	*full*	4,5	1,3
	diag	4,0	0,6

Tabla 5.1-3: Mejores resultados en la fase de entrenamiento

Al realizar una inspección visual de ciertos tramos de ficheros de la base de datos, y evaluar las tasas de aciertos / falsas alarmas / falsos rechazos obtenidas, se observó que en general había una sub-segmentación de los ficheros (elevado número de falsos rechazos), por lo que se consideró la métrica (FA+FR)*FR descrita en el apartado anterior, que penaliza la tasa de falsos negativos. Sin embargo, al hacerse FR=0, el error se considera como nulo independientemente del número de falsas alarmas que se tenga, por lo que esta métrica favorecía demasiado la sobre-segmentación.

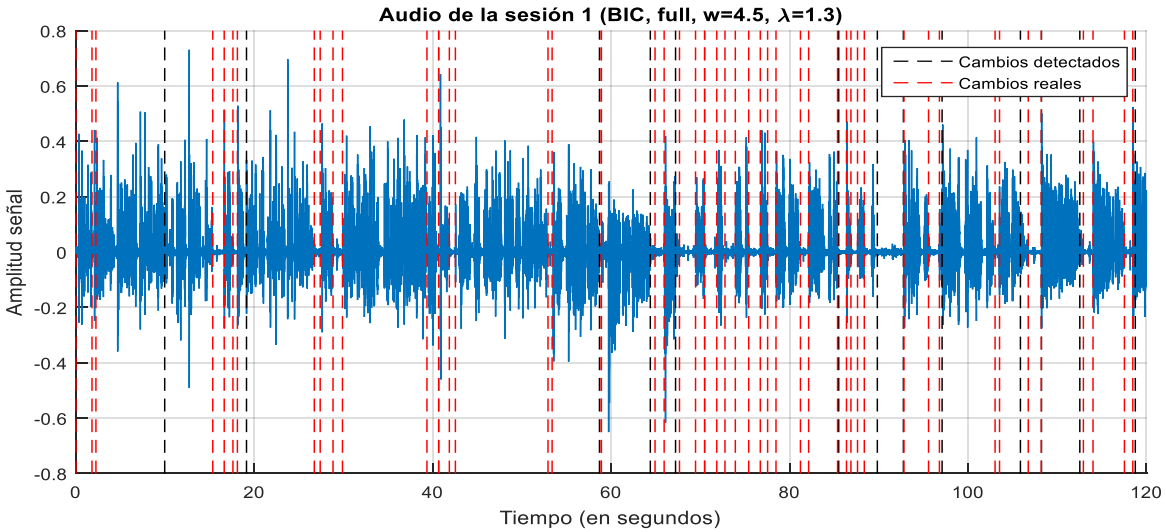

Figura 5-5: Ejemplo de la subsegmentación obtenida con la métrica DER

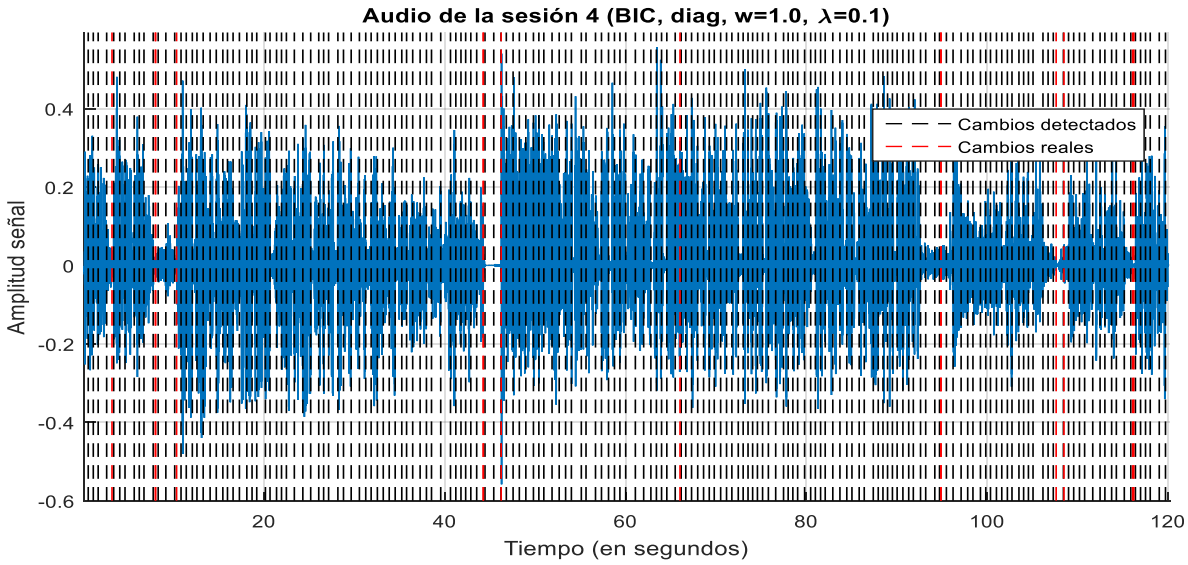

Figura 5-6: Ejemplo de la sobresegmentación obtenida con la métrica (FA+FR)*FR

Finalmente, se optó por la métrica FA+FR, que penalizaba por igual los falsos rechazos y los falsos positivos. Las mejores combinaciones de parámetros obtenidas tras la fase de entrenamiento, con las tres métricas, son las mostradas en la siguiente tabla:

Comparativa resultados		DER		(FA + FR)*FR		FA + FR	
		ventana	λ	ventana	λ	ventana	λ
GLR	*full*	2,5		2,5		2,5	
	diag	1,0		2,0		2,0	
BIC	*full*	4,5	1,3	1,0	0,3	2,5	1,3
	diag	4,0	0,6	1,0	0,1	1,5	0,8

Tabla 5.1-4: Comparativa de resultados obtenidos en el entrenamiento

5.2 Fase de validación

En la etapa de validación, se emplearon los 15 primeros minutos de los otros 12 ficheros restantes de la base de datos de Albayzín 2014. En esta fase ya sólo se han evaluado los resultados sobre la métrica FA+FR, ya que se descartaron las métricas DER y (FA+FR)*FR por sub-segmentar y sobre-segmentar demasiado el audio, respectivamente.

No obstante, aunque a partir de ahora solo se emplea la métrica FA+FR, se han evaluado todas las combinaciones de parámetros escogidas como mejores por los tres resultados.

En general el mejor resultado lo da la métrica FA+FR, pero en el caso del método BIC con ventana de covarianzas completa, los parámetros resultan mejores para la estimación de DER: se obtiene un menor error con la métrica FA+FR, hay un mayor porcentaje de

aciertos, y menor número de falsos rechazos, pero un mayor número de falsos positivos (es preferible que se sobre-segmente ligeramente, a que se sub-segmente, ya que se puede solucionar en con etapa posterior de reagrupación de segmentos).

Mejores resultados en en entrenamiento para:					ventana	λ	Resultados en validación ACIERTOS	FA	FR	FA+FR
DER	(FA+FR)*FR	FA+FR	GLR	*full*	2,5		13,12%	5,47%	86,88%	92,35%
DER				*diag*	1,0		26,89%	39,11%	73,11%	112,22%
	(FA+FR)*FR	FA+FR			2,0		21,63%	10,26%	78,37%	88,63%
DER			BIC	*full*	4,5	1,3	32,09%	25,66%	67,91%	93,57%
	(FA+FR)*FR				1,0	0,3	81,83%	770,67%	18,17%	788,84%
		FA+FR			2,5	1,3	28,53%	22,16%	71,47%	93,62%
DER				*diag*	4,0	0,6	34,75%	32,09%	65,25%	97,34%
	(FA+FR)*FR				1,0	0,1	80,87%	736,88%	19,13%	756,00%
		FA+FR			1,5	0,8	32,47%	29,76%	67,53%	97,29%

Tabla 5.2-1: Resultados obtenidos en la fase de validación

5.3 Fase de test

Finalmente, en la fase de test se ha evaluado la combinación de parámetros que mejor funcionamiento tuvo en la fase de validación, para cada método tanto con matriz de covarianzas completa como diagonal. Para ello se ha empleado la totalidad de la base de datos, exceptuando las partes de esta ya empleadas en las fases de entrenamiento y validación. Los resultados obtenidos se pueden ver en la siguiente tabla:

		ACIERTOS	FA	FR	FA+FR
GLR	*full*	14,88%	6,14%	85,12%	91,25%
	diag	22,81%	12,05%	77,19%	89,24%
BIC	*full*	32,20%	25,98%	67,80%	93,78%
	diag	34,63%	27,33%	65,37%	92,69%

Tabla 5.3-1: Resultados obtenidos en la fase de test

6 Conclusiones y trabajo futuro

6.1 Conclusiones

Este Trabajo de Fin de Grado ha buscado emplear métodos de segmentación de audio y diarización de locutores, así como las métricas de error características de cada uno, de manera conjunta.

Se ha podido observar la dificultad esto que conlleva: la mala elección de algún parámetro da lugar a una segmentación insuficiente, o por el contrario a una segmentación excesiva.

Otra conclusión obtenida es que la elección de una métrica de error suficientemente representativa es de gran importancia, ya que inicialmente se empleó la métrica DER, que favorecía una sub-segmentación solo advertida al emplear métricas relacionadas con los errores de detección, que mejoraron los resultados obtenidos.

Además, de manera paralela a la realización de las pruebas sobre el sistema de segmentación, se han creado herramientas en MATLAB y en perl para la lectura y manipulación de las etiquetas y la comprobación de los errores de segmentación.

6.2 Trabajo futuro

De cara al futuro, se podría profundizar en este Trabajo de Fin de Grado de diversas formas:

Se podrían implementar otras métricas de distancia como las descritas en el Estado del Arte, que pudieran resultar más idóneas en esta aplicación. Además, en este Trabajo de Fin de Grado se ha trabajado con los resultados de cada método por separado, pero evaluar el funcionamiento de la combinación de diferentes métodos en un mismo sistema también pudiera ser objeto de estudio.

También se podrían volver a implementar los métodos en este proyecto empleados, pero utilizando una ventana temporal de tamaño variable en lugar de una ventana deslizante de tamaño fijo.

Por último, también puede profundizarse más en la búsqueda de una métrica de error que represente de mejor manera los resultados obtenidos que las aquí expuestas. Estudiar el añadir una etapa de reagrupación de segmentos al final del sistema también podría resultar de utilidad, ya que se podría ser más permisivo con tasas elevadas de falsas alarmas (sobre-segmentado) que podrían reducirse luego en el reagrupado.

Referencias

[1] Elena Gómez Rincón, “Segmentación de Audio Mediante Características Cromáticas en Ficheros de Noticias”, TFM Junio 2015, EPS UAM

[2] Daniel Ramos, “Tecnologías de Audio Tema 1: Procesado de Audio 1.1.: Efectos de Audio (AFx)”, págs. 13 a 17

[3] Javier Ortega, “Tratamiento de Señales de Voz y Audio Tema 4: Codificación de Voz y Audio”, págs. 10 y 11

[4] T. Kemp, M. Schmidt, M. Westphal, A. Waibel (July 2000). "Strategies for Automatic Segmentation of Audio Data" in "Acoustics, Speech, and Signal Processing", 1988. ICASSP-88., 1988 International Conference on. 3. 10.1109/ICASSP.2000.861862.

[5] Matthew A. Siegler, Uday Jain, Bhiksha Raj, Richard M. Stern, “Automatic Segmentation, Classification and Clustering of Broadcast News Audio”, ECE Department - Speech Group Carnegie Mellon University

[6] Theodoros Theodorou, Iosif Mporas, Nikos Fakotakis, “An Overview of Automatic Audio Segmentation”, University of Patras, October 2014

[7] Gish, H., Siu, M.-H., Rohlicek, R., “Segregation of speakers for speech recognition and speaker identification”, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, 1991, Toronto, Canada, págs. 873 a 876

[8] Castán, D., Tavarez, D., Lopez-Otero, P. et al. “Journal on Audio, Speech, and Music Processing”, December 2015, ISSN 1687-4722

[9] Xavier Anguera, “Robust Speaker Diarization For Meetings”, Phd thesis, Speech Processing Group, Universidad Politécnica de Cataluña, October 2006

Anexos

A. Resultados obtenidos en la fase de entrenamiento

En este anexo se muestran las tablas con los resultados obtenidos para todos los parámetros evaluados en la fase de entrenamiento mediante la métrica FA+FR, y se resaltan aquellos que minimizan el error.

En las tablas relativas al método BIC, se resalta además la zona en la que el error no supera en más de un 10% al valor mínimo de la tabla, facilitando así la visualización de la tendencia que sigue el error al variarse los parámetros. Debido a su tamaño, en estas tablas los valores no van seguidos del símbolo de porcentaje, pero el cálculo se ha realizado de la manera explicada anteriormente.

GLR

window	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
full	117,86%	106,63%	102,04%	91,84%	88,27%	89,29%	93,37%	93,88%	94,90%	93,37%
diag	144,90%	99,49%	91,33%	89,29%	94,39%	95,41%	96,43%	95,92%	101,53%	99,49%

Tabla A-1: Resultados obtenidos en el entrenamiento para GLR

BIC *diag*

	window																		
λ	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5	10
0,0	961	627	475	393	335	291	281	238	232	226	211	199	188	180	178	165	173	162	160
0,1	954	625	475	392	334	290	281	238	232	226	211	199	188	180	178	165	173	162	160
0,2	876	585	446	367	319	278	273	234	228	221	203	193	184	180	174	164	173	161	159
0,3	593	418	337	278	242	217	217	195	193	193	180	170	165	156	158	144	153	145	146
0,4	346	251	215	180	169	155	157	150	152	155	147	138	139	138	139	127	139	131	135
0,5	196	148	136	128	128	121	133	128	129	134	133	125	129	126	128	119	128	120	126
0,6	124	101	109	104	108	105	115	110	114	119	118	116	119	115	116	114	122	116	118
0,7	105	92	94	96	98	101	108	103	110	116	113	110	113	112	113	108	114	111	114
0,8	93	88	89	94	95	102	108	106	106	111	111	107	110	109	111	107	110	107	109
0,9	93	93	88	94	98	102	103	106	104	108	109	105	108	107	107	105	105	105	105
1,0	97	91	90	98	101	105	109	108	104	107	109	105	109	106	105	104	105	101	99

Tabla A-2: Resultados obtenidos en el entrenamiento para BIC diag

BIC *full*

	window																		
λ	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**5,5**	**6,0**	**6,5**	**7,0**	**7,5**	**8,0**	**8,5**	**9,0**	**9,5**	**10**
0,0	1002	649	481	401	337	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,1	1002	649	481	401	337	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,2	1000	649	481	401	336	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,3	996	649	481	401	336	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,4	992	646	480	400	336	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,5	968	635	470	397	335	291	260	236	224	223	202	197	188	179	178	174	167	163	161
0,6	924	610	458	388	329	287	257	234	223	222	202	196	187	178	178	174	167	163	161
0,7	829	562	422	362	308	276	248	228	219	214	198	191	183	175	176	174	165	161	160
0,8	631	451	353	301	267	244	226	206	195	198	186	183	175	167	168	165	158	153	154
0,9	424	291	245	230	210	196	182	178	169	177	164	168	163	149	153	154	151	146	144
1,0	240	180	154	168	160	159	152	148	146	154	146	148	143	136	140	145	142	138	134
1,1	150	127	96	107	110	116	120	119	122	138	128	133	129	127	132	134	132	126	126
1,2	116	101	84,2	85,2	85,7	98	102	106	108	117	120	124	123	116	121	121	118	117	117
1,3	107	101	85,7	78,1	81,1	86,2	94	96	99	108	112	114	113	111	113	111	110	108	111
1,4	105	103	92	86,7	84,2	82,1	90	91	96	103	105	104	104	101	106	105	106	103	104
1,5	106	105	97	88,3	85,2	87,8	87,8	86,7	91	96	98	99	99	97	101	100	100	98	101
1,6	106	106	103	92	88,3	88,3	86,7	87,8	85,2	91	92	97	96	94	98	98	97	95	97
1,7	106	106	105	97	94	92	91	88,8	88,3	85,2	88,8	89	96	92	95	96	96	95	94
1,8	106	106	107	104	95	94	92	92	87,8	88,8	86,7	90	92	90	93	96	94	93	94
1,9	106	106	106	105	99	95	95	94	91	89	89	88	93	91	94	93	92	92	93
2,0	106	106	106	105	103	98	97	94	93	94	94	91	91	91	94	94	91	91	90

Tabla A-3: Resultados obtenidos en el entrenamiento para BIC full