Universidad Zaragoza

1542

Proyecto Fin de Carrera

Estudio de métodos de Diarización en un entorno de Broadcast

Autor

Ignacio Viñals Bailo

Director:

Alfonso Ortega Giménez

Departamento de Ingeniería Electrónica y Comunicaciones
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

Diciembre 2013

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

instituto de investigación en ingeniería de Aragón

A mis padres

Agradecimientos

Estas son las líneas más difíciles de rellenar creo que por todo el mundo, dado que siempre tenemos el temor a olvidarnos a alguien. Por lo tanto, si alguien se me olvida, espero que me lo perdone.

En primer lugar querría agradecer a mis padres todo ese apoyo incondicional que me han brindado durante toda mi vida, y en especial durante estos años de estudio, tanto en los buenos como en los malos momentos.

También debo agradecer a mi director, Alfonso Ortega, toda la confianza que ha depositado en mí, así como su franqueza y cercanía en todo momento. Ha sido un gran placer trabajar contigo. Tampoco debo olvidar a Eduardo Lleida, el cual, si bien no estaba ligado directamente a la realización directa de este proyecto, en todo momento se ha portado para conmigo de una manera encomiable, dándome la oportunidad de vivir experiencias únicas.

Otra persona a la que tengo que agradecer mucho es Pedro Carro, mi profesor asignado para Programa Tutor. Durante toda la carrera has estado ahí para contestar esas preguntas, y aconsejarme cuando tenía alguna duda. Te animo a seguir con esta labor.

En lo referente a los compañeros de laboratorio, si bien tampoco he llegado a compartir grandes períodos de tiempo con ninguno de vosotros, me gustaría recordar esas charlas con Susana, que tan bien me recibiste cuando llegué a tu laboratorio con cara de novato. Otra mención que no debo olvidar es Diego Castán. Desde el primer día en el que llegué como el nuevo a los laboratorios hasta al último has sido una ayuda valiosísima, en todo momento con una sonrisa en la boca y con espíritu de ayudar, sin la cual, este trabajo no habría llegado a ver la luz. Sigue siendo así.

Finalmente están todos mis compañeros de estudio. Lo siento chicos, pero aquí si que no me atrevo a enumeraros, que sois muchos, y seguro que me olvido de alguien. Sin embargo quiero agradeceros estos fantásticos años que hemos vivido juntos, con nuestras penas, pero compensadas por la gran cantidad de alegrías.

Resumen

El actual estado del arte dentro de las tecnologías del habla permite una gran variedad de soluciones a los diferentes problemas existentes en esta rama, ya sean desde el reconocimiento de locutor al reconocimiento de discurso, pasando por la indexación de contenidos multimedia.

En muchos casos, para hacer posibles estas tareas, es necesario aislar unos locutores de otros, permitiendo procesar su información por separado. La rama de las tecnologías del habla que tiene esta misión es la Diarización.

Este Proyecto Fin de Carrera recogerá el testigo de otros trabajos para aplicarlos a un entorno poco trabajado, pero a su vez, poseedor de grandes dificultades. Me estoy refiriendo a los sistemas de radiodifusión o Broadcast. Este entorno se caracteriza principalmente por la existencia de un número desconocido de locutores, por la superposición de otras fuentes sonoras sobre las voces a diferenciar, así como por una actividad de los locutores no uniforme, generalmente alternando segmentos de locutores muy activos aquellos menos relevantes.

Dada la magnitud de los sistemas de Diarización en cuanto a subtareas internas, así como a la gran variedad de soluciones propuestas para cada una, se concentrarán la mayoría de los esfuerzos en aquella tarea considerada como más compleja para este entorno, la tarea de la aglomeración o Clustering, pues es aquella donde las dificultades de este entorno son más críticas.

Finalmente, más allá del estudio de la subtarea de Clustering propiamente dicho, se desarrollará un sistema completo de Diarización, a fin de comparar resultados con aquellos pocos existentes en la bibliografía.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

1.1. Motivación del proyecto

La investigación en el campo de las tecnologías del habla ha sufrido en las últimas décadas un auge enorme, surgiendo gran cantidad de aplicaciones a su alrededor. Sin embargo, muchos de estos sistemas, como requisito de funcionamiento, o como función propia a desarrollar, deben aislar los segmentos sonoros de un único locutor para procesarlos conjuntamente. Realizada esta tarea antaño manualmente, la cantidad de información que se requiere actualmente ha hecho que dicho proceso deba realizarse de una manera automática por un sistema electrónico/informático. Esta labor será llevada a cabo mediante las denominadas técnicas de Diarización, la rama de las tecnologías del habla centrada en la separación de locutores.

Centrándonos en un ambiente concreto, existe un entorno de trabajo sobre el cual se desarrollará todo el proyecto, donde esta clase de sistemas de Diarización son de vital importancia: Los medios de radiodifusión o *Broadcast*. Aparte de la gran cantidad de aplicaciones de reconocimiento de locutor, donde esta técnica es complementaria, la gama de aplicaciones específicas de diarización para Broadcast van desde el subtitulado a la indexación de datos mediante relaciones de locutor activo.

1.2. Marco del proyecto

La línea de trabajo en la que este proyecto queda alojado es acerca de la Segmentación de Audio, es decir, la clasificación de audio en función de su fuente sonora. La tarea de Diarización es una especificación del problema de Segmentación, donde las fuentes a separar son los locutores activos en los diferentes segmentos sonoros.
Este Proyecto Fin de Carrera «Estudio de métodos de Diarización en un entorno de Broadcast» no es el primer trabajo realizado sobre Segmentación de Audio en la Universidad de Zaragoza, sino que sigue la labor impulsada por distintos trabajos como [Castan, 2009] o [Vaquero, 2011]

entre otros, de los cuales hereda y busca complementar.

1.3. Introducción

WHO SPOKE WHEN? Esta pregunta es la empleada en la mayor parte de la bibliografía existente, incluyendo textos como [Vaquero, 2011] o [Anguera et al., 2012], para explicar el término Diarización. Esta rama de las tecnologías del habla pretende la separación de diversas voces, pertenecientes a distintos locutores, a partir de una o varias pistas de audio que contengan a un conjunto de ellas. El sistema busca acumular bajo una etiqueta común todos los segmentos sonoros en los que la voz de un locutor concreto esté presente. El sistema en ningún momento buscará dotar a dicho locutor de una identidad concreta, pues dicha clase de funcionalidades ya son realizadas por otros sistemas, a los que esta tecnología da apoyo.

Si bien es cierto que la Diarización es un concepto relativamente joven dentro de las tecnologías del habla, existen tres grandes líneas de investigación, dependientes de los tres principales entornos en los que dichas técnicas han sido estudiadas.

- El primer entorno de trabajo es el **canal telefónico**. Es la versión más acotada del mismo problema. Se suelen asumir dos locutores activos, así como un medio de variabilidad limitada.
- El segundo entorno es el entorno de **radiodifusión** o *Broadcast*. Este tipo de entorno lo podemos considerar más complejo que el anterior. Se caracteriza por tener un número de locutores desconocido, por superponer música o ruido respecto a la voz así como combinar continuamente segmentos de locutores prioritarios muy presentes en el audio con pequeñas intervenciones de locutores de menor relevancia.
- Por último está el entorno de **reuniones** o *meetings*. Este entorno presenta la versión más general y nada acotada del problema de la diarización. Se caracteriza por tener un número desconocido de locutores, desconociendo a su vez su posición respecto a los micrófonos de grabación, y en un ambiente nada controlado en términos de ruido.

El problema de Diarización se corresponde con un problema de reconocimiento de patrones. Un sistema de reconocimiento de patrones por definición parte de un conjunto de observaciones a su entrada y devolverá como resultado las clases a las que pertenece dichas observaciones. Dicha clasificación puede ser llevada a cabo por distintos métodos y principios.

Se puede observar un diagrama de bloques de este sistema en la figura 1.1. Sin embargo, el problema de Diarización en general no puede resolverse por el sistema genérico expuesto, requiriendo una variante. Esto se debe a que, en la versión más general del problema, no se conocen en principio las clases o locutores ni su número exacto.

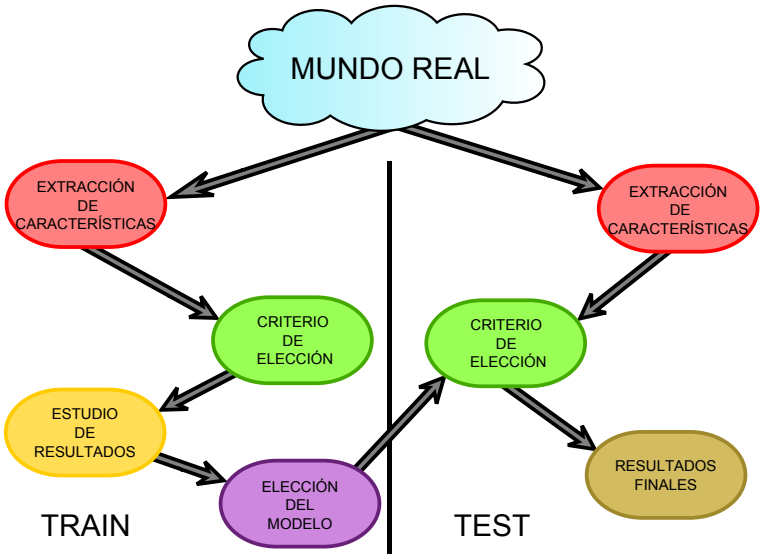


Figura 1.1: Esquema básico de un sistema de Reconocimiento de Patrones

1.4. Objetivos

El objetivo de este proyecto será el estudio de las diversas técnicas existentes actualmente en el ámbito de la Diarización para radiodifusión o *Broadcast*. Sin embargo, la complejidad de los sistemas aconseja dividir el problema, dando lugar a varios subobjetivos:

- Desarrollo de la etapa de Agrupación o Clustering. Dentro de las diferentes funcionalidades de un sistema de Diarización, la etapa de Clustering o Aglomeración es la más crítica por el salto a un entorno de Radiodifusión o Broadcast. Además, esta etapa es muy compleja por sí misma, por lo que se realizará un estudio en profundidad de la misma, centrando la mayoría de esfuerzos en este punto. Por las características del entorno de Broadcast, la etapa de Clustering puede descomponerse en dos funcionalidades diferenciadas, debiendo profundizar en ambas:

  - Estudio de los criterios de fusión. Una etapa de Clustering tendrá como entrada un conjunto de segmentos sonoros que contienen idealmente audio de un único locutor. Un objetivo será establecer los mejores métodos para combinar los segmentos de un mismo locutor, minimizando los errores de combinación.
  - Estudio de la estimación del número de locutores. El entorno de Broadcast implica un número desconocido de locutores, por lo que se deberá estimar dicho valor, a fin de saber en qué momento finalizar la tarea de fusión.

- Integración de las diferentes etapas desarrolladas anteriormente en un sistema real y completo. Ya que la gran mayoría de esfuerzos serán destinados a la tarea de Clustering, al

menos es necesario establecer un sistema completo de Diarización con todos sus elementos, con la finalidad de obtener datos comparables con otros estudios.

1.5. Organización de la memoria

Una vez vista una pequeña introducción al proyecto realizado, así como vistos los objetivos definidos para el mismo, se expondrán las distintas partes en las que el proyecto se descompone:

- **Estudio del estado del arte.** En este capítulo se presentará el Estado del Arte actual dentro de la rama de Diarización. Basándose en una amplia búsqueda bibliográfica se tratará de dar una visión general de esta rama del conocimiento.
- **Diseño de la etapa experimental**. A continuación, el lector encontrará un capítulo dedicado a la explicación de los diferentes experimentos que se han llevado a cabo.
- **Exposición de resultados**. Después se incluirá un capítulo dedicado a los resultados obtenidos de los distintos experimentos.
- **Conclusiones**. Vistos los resultados del capítulo anterior, se formularán unas conclusiones, en los que se tratará de justificar el porqué de los diferentes valores obtenidos.
- **Trabajo futuro** Por último, y vistas ya las conclusiones del trabajo realizado, se propondrán las lineas de trabajo para, partiendo de este trabajo, avanzar en esta rama del conocimiento.

Capítulo 2

Estado del Arte

En este capítulo se va a exponer el actual estado del arte en lo que a técnicas de Diarización se refiere. Pese a ser adelantados parcialmente algunos esbozos en el capítulo precedente, en éste se realizará una presentación formal del mismo. Primero se describirá un sistema completo de Diarización, comentando las tareas a realzar, para después explicarlas detenidamente. Para más información, se puede acudir a algunos documentos mucho más detallados en este sentido tales como [Anguera, 2006], [Tranter and Reynolds, 2006] o [Vaquero, 2011].

Como conclusión de este apartado, se incluye una sección, "Medidas de Prestaciones y Error", referente a las distintas medidas diseñadas para evaluar el funcionamiento de todos estos sistemas.

2.1. Elementos de un sistema de Diarización

La gran mayoría de la bibliografía existente apunta a un esquema general para un sistema de Diarización como el representado en la figura 2.1. Se pueden observar tres principales subtareas a realizar:

- **Extracción de Características**. En general, los datos de entrada, sin tratar, no son útiles para el reconocimiento de patrones. Esto puede deberse a redundancia de información, o por costes computacionales , etc. Por ello deben ser procesados, con la intención de hacerlos lo más aprovechables posibles, extrayendo la información y condensándola de una forma compacta, funcional y sencilla. El resultado de esta etapa será un conjunto de características, una versión de la información de entrada apta para alimentar al sistema de reconocimiento de patrones. La correcta elección de estas características permitirá a los subsistemas posteriores reducir su complejidad en gran medida, por lo que no es desdeñable su mejora.
- **Segmentación**. En las labores de separación de locutor, se requerirá una etapa que sea

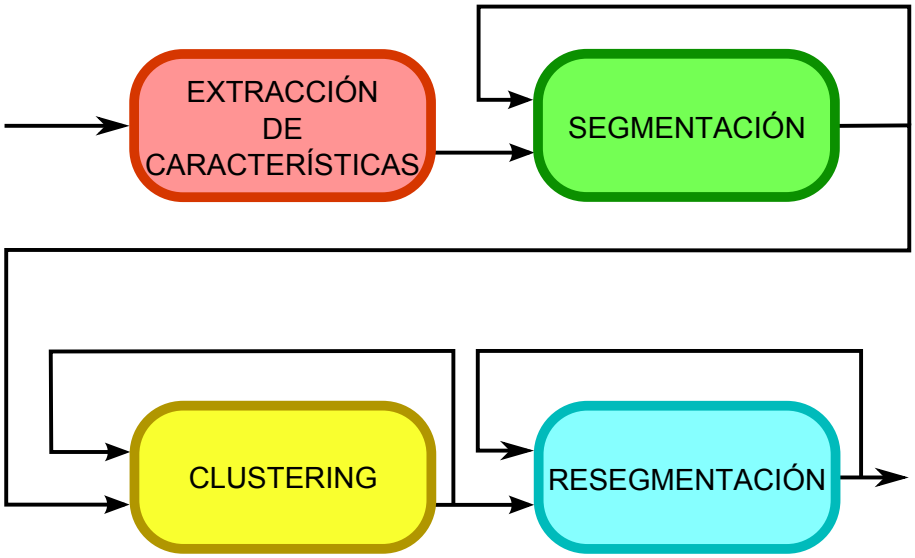


Figura 2.1: Esquema general de un sistema de Diarización

capaz de detectar el instante preciso en el cual se produce un cambio de locutor. Este subsistema es el encargado de dicha función. Puesto que todo locutor puede modelarse, las muestras del mismo locutor tienen que mostrar cierto parecido. Dado un segmento de audio, homogéneo, se irá extendiendo añadiendo datos, hasta que dicha homogeneidad no sea tan elevada. En ese momento el sistema supondrá que ha encontrado una frontera entre dos locutores. Este proceso se repetirá a lo largo del audio de análisis con la intención de detectar todas las fronteras existentes.

- **Agrupación** o **Clustering**. Este último sistema se ocupará de agrupar bajo una misma etiqueta, todos los segmentos producidos por el mismo locutor, de tal manera que cada etiqueta corresponda unívocamente con el audio de un único hablante. El concepto en el que se basa es buscar aquellos segmentos que aunque separados a lo largo del audio, entre ellos podamos asumir la homogeneidad comentada anteriormente, permitiéndonos suponer que proceden del mismo locutor. Si queremos dotar a este esquema un carácter general se deberá aceptar la posibilidad de que el sistema deba afrontar un problema con un número variante y desconocido de locutores, en vista de lo cual este subsistema será el encargado de incluir la funcionalidad necesaria para inferir dicho número.

Una vez explicadas estas etapas fundamentales, se deben mencionar el resto de aspectos y matices no explicados anteriormente, pero que quedan recogidos en la figura 2.1:

- Un punto importante es la realimentación mostrada en el gráfico. A pesar de no ser obligatoria, y en ciertos sistemas no ser aplicada, puede aplicarse con fines de refinado de los resultados mediante pasadas consecutivas.

- Como última tarea, se ha puesto una etapa de Resegmentación. Esta etapa, como su nombre indica, es una segunda segmentación y tiene la simple función de refinar los resultados previos.

Una vez visto el esquema desde un punto de vista general, se procederá a profundizar cada uno de los distintos elementos que lo forman.

2.1.1. Extracción de Características

El proceso de Diarización exige como primera funcionalidad la extracción de diversas características para la correcta separación de los distintos locutores presentes en el audio.

Esta etapa presenta una gran dificultad: **La señal de voz no es estacionaria** estrictamente hablando, e incluso solo podemos asumir un cierto grado de estacionariedad en períodos muy pequeños, del orden de $\mathbf{20 - 50ms}$. Esta naturaleza obliga a desechar las técnicas muy potentes y sencillas, ya que exigen como requisito la estacionariedad de la señal de voz.

Una rama de investigación muy importante es aquella que parte del **modelo de producción** humano. El conocimiento en la forma de modelar los mecanismos de producción del habla ha evolucionado mucho, siendo aplicado a distintas características, entre ellas, MFCC (Mel Frequency Cepstral Coefficients) y PLP (Perceptual Linear Predictive) . Aportan una ventaja añadida al combinar conocimientos acerca de la **producción del habla** con aquellos sobre la **percepción del sonido** por parte de los órganos auditivos, en los que también se inspiran.

Si bien han recibido una amplio apoyo por la comunidad investigadora, amparadas por buenos resultados, su origen, diseñadas para el reconocimiento de discurso (*speech recognition*) independientemente del locutor, ha generado cierto recelo. Por ello se han propuesto otras características, como Perceptual Minimum Variance Distorsionless Response (PMVDR), Smoothing Zero Crossing Rate (SZCR) o Filter-Bank Linear Coefficients (FBLC), todos ellos estudiados en [Huang and Hansen, 2006] además del estudio de prosodías [Friedland et al., 2009], mas estas nuevas características no aportan ninguna mejora sustancial respecto a MFCCs o PLPs, a tenor de los resultados obtenidos.

2.1.2. Segmentación

La segmentación es la etapa cuya finalidad es la búsqueda de las fronteras que delimitan los segmentos procedentes de distintos locutores. Existen en la bibliografía gran cantidad de maneras de clasificar los distintos métodos. Un esquema muy clarificador lo aporta

[Chen and Gopalakrishnan, 1998], ya que clasifica los distintos algoritmos según estén basados en métricas, modelos o silencios:

- **Basados en Métricas**. Se definirá una medida de distancia o parecido entre dos subregiones contiguas de un mismo fragmento de audio.

  La métrica representa para una región de estudio la mejora de modelado por representar como un único locutor (hipótesis H_0) toda la región respecto a emplear dos modelos (hipótesis H_1), conteniendo cada una de las subregiones uno de los locutores .

  Por tanto, para una región de estudio, se escogerán una serie de muestras, candidatas a ser frontera, siéndoles aplicada la métrica definida a las dos subregiones que dichas muestras delimitan. De entre todas las muestras, se escogerá aquella cuyo valor de métrica indique mayor probabilidad de ser frontera. Entonces, se tomará la decisión de si considerarla como tal, una frontera (hipótesis H_1), o si por contra, no indica una transición entre locutores (hipótesis H_0).
  Se trata de la filosofía más robusta y empleada, ya que no asume en ningún momento la existencia de datos a priori. Dentro de este conjunto de medidas destacan medidas como BIC y algunas variantes. Dada su relevancia, se definirá BIC tal que:

  BIC (Bayesian Information Criterion) es una medida del grado de relación entre unos datos $\boldsymbol{\chi}$ y un modelo $\boldsymbol{\Psi}$, candidato a ser generador de dichos datos. BIC se define como

  $$BIC(\Psi) = log(\mathcal{L}(\chi|\Psi)) - \lambda\frac{1}{2}\#(\Psi)log(N) \tag{2.1}$$

  donde el término $\boldsymbol{log(\mathcal{L}(\chi|\Psi))}$ representa la logverosimilitud de los datos $\boldsymbol{\chi}$ respecto al modelo $\boldsymbol{\Psi}$, representando por tanto el grado de relación entre ambos, mientras el término $\boldsymbol{\lambda\frac{1}{2}\#(\Psi)log(N)}$ es un parámetro de penalización dependiente del número de observaciones N, un parámetro de ajuste $\boldsymbol{\lambda}$ y el número de parámetros independientes del modelo $\boldsymbol{\Psi}$. Para segmentación, BIC no es aplicable directamente, ya que solo evalúa la calidad de modelado de unos datos. En este caso, se necesita contraponer la calidad de modelar con un único locutor (H_0) respecto a modelar con dos locutores (H_1). Por lo tanto se emplea una medida desarrollada a partir de BIC, ΔBIC, definida como:

  $$\Delta BIC = BIC(H_1) - BIC(H_0) \tag{2.2}$$

- **Basados en Modelos**. Si se dispone de la suficiente cantidad de datos *a priori*, se pueden generar modelos estadísticos para cada locutor, y se pueden determinar la pertenencia de

los datos a los mismos en función de la probabilidad de éstos respecto a dichos modelos. Aunque conceptualmente válido, en muchos casos no se dispondrá de este conjunto de datos, ya sea por datos insuficientes o por cuestiones de robustez, ya que los modelos dependen de los datos con los que han sido estimados.

- **Basados en Silencio** . Asume la existencia de modelos *a priori* de voz-silencio, de tal manera que se realiza una segmentación voz-silencio, en la que todo silencio se etiqueta como una posible transición. Tiene el inconveniente de que solo asume como posible transición de locutor aquella que sucede a través de un período de silencio, perdiendo por ello toda frontera entre locutores sin silencio de por medio.

Vista la idea global, en el anexo A puede obtenerse más información acerca de cada una de estas opciones de segmentación. Para la definición formal de métricas, se ha elaborado el anexo B, en el cual se describen gran cantidad de métricas diferentes, entre ellas todas aquellas relacionadas con este proyecto.

2.1.3. Agrupación o Clustering

El problema del Clustering consiste en la agrupación de los diferentes segmentos sonoros en un conjunto discreto de clases, *a priori* desconocidas, que deberían representar los diferentes hablantes presentes en un audio. Sistema muy vinculado a la diarización, en este ámbito realiza este proceso a partir del audio procedente de la segmentación, aunque también puede trabajar con audio procedente de diversas grabaciones como en [van Leeuwen, 2010].

Asumiendo que para una segmentación dada, existen un conjunto C de posibles etiquetados c_i, obtenidos mediante diferentes combinaciones de los segmentos obtenidos por la etapa de Segmentación. El reparto correcto, único, será aquel que maximice la verosimilitud de los datos respecto a dicho etiquetado. Por ello, el método óptimo consiste en calcular para cada reparto c_i su verosimilitud, escogiendo aquel que maximice este valor. Sin embargo, esto es inviable computacionalmente, dado que el número de etiquetados posibles crece drásticamente conforme el número de segmentos aumenta. En consecuencia se deberán aplicar métodos aproximados, entre los que destaca la **Agrupación Aglomeraiva Jerárquica o AHC**. Una versión más extensa de esta explicación puede verse en el anexo B.

2.1.3.1. Agrupación Aglomerativa Jerárquica o AHC

La agrupación Jerárquica parte de un etiquetado inicial para un audio dado (el reparto más grueso (un único locutor) o el más fino (cada cluster lo conforma solo un segmento sonoro)), y el sistema iterativamente va fusionando o dividiendo estas agrupaciones hasta llegar al número óptimo de locutores. Las divisiones o fusiones realizadas no serán reevaluadas en una

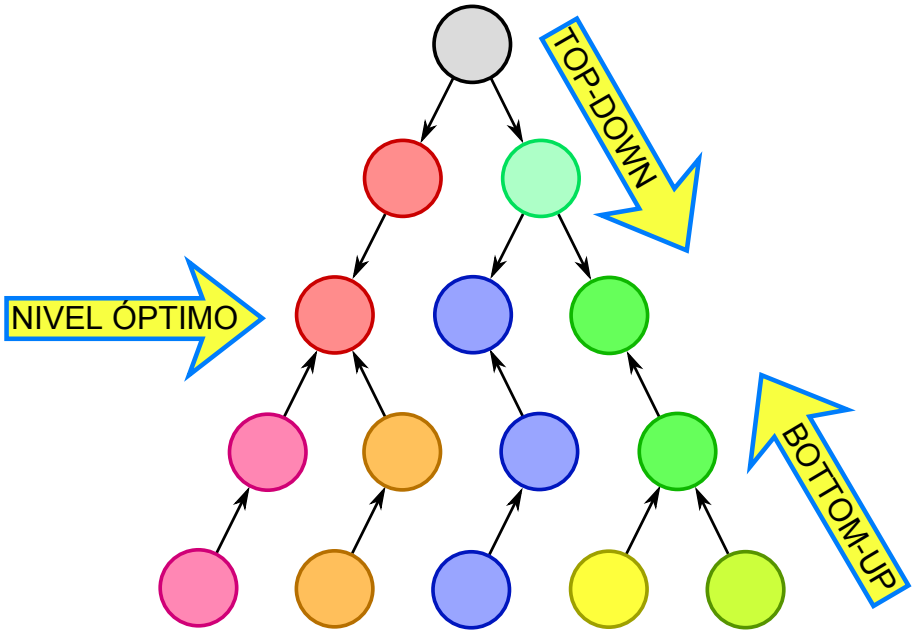


Figura 2.2: Esquema de funcionamiento de estrategias de Clustering Bottom-Up y Top-Down

iteración siguiente, y se arrastrarán los errores cometidos. Este es el precio a pagar por esa reducción del coste computacional. Dentro de la agrupación jerárquica, existen dos filosofías de clusterización

- **Bottom-Up**. En esta estrategia, el sistema partirá de la partición más fina, en la que cada cluster únicamente contendrá a un segmento de audio. A partir de esta segmentación, el susbistema irá fusionando clusters iterativamente. Es la filosofía más empleada, ya que el número de iteraciones está acotado (no puede agruparse sobre el reparto más grueso). Aparte aporta ventajas de cálculo, pues se rige únicamente por una matriz a controlar, y permite estudiar cada fusión por separado. La primera vez que se empleó esta técnica en *speaker Clustering* fue en [Jin et al., 1997].

- **Top-Down**. Esta filosofía parte del otro extremo. Se comenzará con un número limitado de clusters (como caso general una único cluster), siendo divididos iterativamente hasta alcanzar el criterio de parada. A diferencia de la estrategia anterior, presenta una gran desventaja, ya que no existe un nivel límite en la generación de los árboles de decisión, puesto que siempre puede desarrollarse una partición más fina descomponiendo un segmento en dos. Esta filosofía es mucho menos común que la estrategia Bottom-Up. Como ejemplo de utilización se puede citar [Reynolds and Torres-Carrasquillo, 2005].

La figura 2.2 representa las dos formas más comunes de realizar la Agrupación Jerárquica. Mientras el sistema Bottom-Up únicamente comprueba cada par de clusters, solo fusionando el par más cercano por ciclo, la topología Top-Down analiza cada cluster y divide uno solo por

iteración. En ambos casos las decisiones tomadas son locales e irreversibles. Todo sistema de Clustering requiere a su vez de dos elementos diferentes: una medida de similitud y un criterio de parada, que presentamos a continuación.

- **Medida o distancia de similitud**. En el proceso de Clustering se necesita de algún tipo de medida para valorar qué fusión o división de clusters conlleva una mayor ganancia de modelado. Esta ganancia nos la aportan todas aquellas métricas definidas para Segmentación. Se aplicarán las mismas métricas considerando como subregiones el audio etiquetado por cada uno de los clusters, siendo la región de estudio el audio procedente de la unión de ambos.

- **Criterio de parada**. Debido al carácter iterativo de la aglomeración jerárquica, es necesario saber cuando detener dicho proceso al alcanzar el número de clusters igual al número de locutores. En Broadcast, ademas, este valor deberá ser estimado, ya que es a priori desconocido. Para ello, la bibliografía emplea técnicas como:

  - En la bibliografía relacionada se observa una técnica muy común: un **umbral respecto a la métrica de fusión**, empleado en numerosas ocasiones, tales como en [Zhou and Hansen, 2005] o [C.Barras et al., 2004]. Este umbral consistirá en ajustar cuál es la mejora de modelado mínima a obtener por toda fusión correcta, considerando como subregiones los dos clusters a fusionar. En el momento en el que la fusión a realizar alcanza el umbral (toda fusión es incorrecta), el sistema interrumpirá la fusión de clusters.
  - En la bibliografía también aparecen sistemas más complejos, como Mean-Shift ([Fukunaga and Hostetler, 1975]) Mean-Shift es una técnica procedente de Factor Analysis, donde las características producidas por cada locutor son parametrizadas por una gaussiana. Por tanto, contando el número de gaussianas, se estimará el número de locutores. Para encontrar y contabilizar dichas gaussianas se buscarán sus máximos en las características del audio a estudiar, empleándose un método iterativo que, mediante el gradiente, irá aproximándose a dichos picos.

2.1.4. Sistemas de Diarización

Una vez vistas las diferentes técnicas, cada cual solucionando y aportando su punto de vista sobre un problema concreto, es necesario subir de nivel estudiando la integración en un único sistema de Diarización de los distintos elementos anteriormente explicados.
Una buena referencia de sistemas se encuentra en [Tranter and Reynolds, 2006]. En la mayoría de sistemas tradicionales se procederá con una segmentación basada en distancias, muchas veces escogiendo BIC como eje de todo el proceso, ya sea segmentación, Clustering y criterio

de parada. En algunos casos las distribuciones escogidas serán gausianas multidimensionales de covarianza completa, como en [Anguera, 2005], o GMM (Gaussian Mixture Models) ([Wooters et al., 2004]). Para terminar, se aplicará una etapa de resegmentación empleando sistemas basados en modelos, construyendo a tal fin un HMM (*Hidden Markov Model* o modelo oculto de Markov). Este paso de resegmentación puede ser aplicado únicamente al final, o intercalarse dentro del algoritmo de Clustering, siendo realizado a cada iteración.
Para terminar, los últimos avances en reconocimiento de locutor han aportado nuevas herramientas. Una muy potente es JFA (**Joint Factor Analysis**), trabajado en [Castaldo et al., 2008], [Kenny et al., 2010] o [Vaquero et al., 2010]. En estos trabajos, se extraen a partir de los MFCCs un conjunto de *speaker factors*, representaciones compactas de los locutores, para convertirse en la entrada a los sistemas ya vistos.

2.2. Medida de prestaciones y error

Como principal medida de prestaciones se ha desarrollado un parámetro, el término DER (Diarization Error Ratio), que computa el porcentaje de audio mal clasificado. Se define como la proporción de audio mal etiquetado respecto al total, o matemáticamente:

$$DER = \frac{Audio_{Mal-Clasificado}}{Audio_{Total}} \tag{2.3}$$

Además, esta medida se descompone en cuatro términos de error, en función de la razón porqué el audio no este bien etiquetado:

- **Missed Speech**. Proporción de audio con voz no considerado con un locutor activo.
- **False Alarm Speech**. Proporción de audio sin voz etiquetado con un locutor activo.
- **Speaker Error**. Proporción de audio atribuido a un locutor incorrecto.
- **Overlapped Speech Error**. Proporción de audio con solapes mal etiquetado, ya sea confundiendo locutores o perdiendo uno de los locutores activos.

pudiendo escribirse:

$$DER = Error_{Miss} + Error_{FalseAlarm} + Error_{Speaker} + Error_{Overlap} \tag{2.4}$$

Tras su implantación, se ha convertido en un estándar de facto, por lo que todo trabajo elaborado lo emplea, además de poder emplear alguna otra medida. Dado que esta medida solo aporta una visión global de error, sin ofrecer más información, también se han desarrollado medidas complementarias mediante las Impurezas de Cluster e Impurezas de Locutor, que estudiarán la tarea de Clustering.

Sea un conjunto de N segmentos $\boldsymbol{\Omega}$, que contiene R locutores distintos, y R<N, se define como frecuencia relativa del locutor $\boldsymbol{r}$ en el segmento $\boldsymbol{n}$ como

$$f_r(n) = \frac{L_r(n)}{L(n)} \tag{2.5}$$

donde $\boldsymbol{L_r(n)}$ es el número de observaciones o tramas en el segmento $\boldsymbol{n}$ que pertenecen al locutor $\boldsymbol{r}$, y $\boldsymbol{L(n)}$ $\sum_{r=1}^{R} \boldsymbol{L_r(n)}$ es el total de observaciones en el segmento $\boldsymbol{n}$. Para un hipotético reparto H que presente S clusters $\boldsymbol{\mathcal{C}_s}$, $\boldsymbol{s = 1, ..., S}$, se puede definir la frecuencia de un locutor $\boldsymbol{r}$ en un cluster $\boldsymbol{\mathcal{C}_s}$ como:

$$f_r(\mathcal{C}_s) = \frac{L_r(\mathcal{C}_s)}{L(\mathcal{C}_s)} = \frac{\sum_{n\in\mathcal{C}_s} f_r(n)L(n)}{\sum_{n\in\mathcal{C}_s} L(n)} \tag{2.6}$$

Donde $\boldsymbol{L_r(\mathcal{C}_s)}$ es el número de tramas de todos los segmentos en el cluster $\boldsymbol{\mathcal{C}_s}$ que pertenecen al locutor $\boldsymbol{r}$, y $\boldsymbol{L(\mathcal{C}_s)}$ es el número total de tramas en el cluster $\boldsymbol{\mathcal{C}_s}$.

A partir de la definición de $\boldsymbol{f_r(\mathcal{C}_s)}$, se puede definir la pureza de cluster para una única agrupación $\boldsymbol{\mathcal{C}_s}$ como la frecuencia del locutor r que obtiene la máxima frecuencia en dicho cluster $\boldsymbol{\mathcal{C}_s}$. Matemáticamente:

$$P_{cluster}(\mathcal{C}_s) = \max_r(f_r(\mathcal{C}_s)) \tag{2.7}$$

y por ende, la pureza del conjunto de todos los segmentos $\boldsymbol{\Omega}$ dada el hipotético reparto H, se define como la media ponderada de las purezas individuales de cada cluster.

$$P_{cluster}(\Omega|H) = \frac{\sum_{s=1}^{S} L(\mathcal{C}_s)P_{cluster}(\mathcal{C}_s)}{L(\Omega)} \tag{2.8}$$

donde $\boldsymbol{L(\Omega)}$ es el total de muestras de voz en todos los segmentos.

Análogamente a la pureza de cluster surge el concepto de pureza de locutor. En primer lugar se calculará la frecuencia de un segmento $\boldsymbol{n}$ en un locutor $\boldsymbol{r}$ como:

$$g_n(r) = \frac{L_r(n)}{L_r(\Omega)} \tag{2.9}$$

donde $\boldsymbol{L_r(\Omega)}$ es el número de tramas que pertenecen al locutor $\boldsymbol{r}$ en la totalidad de segmentos sonoros. A partir de esta definición se puede obtener la frecuenciade un cluster $\boldsymbol{\mathcal{C}_s}$ en un locutor $\boldsymbol{r}$, pudiendo expresarse como:

$$g_{\mathcal{C}_k}(r) = \frac{L_r(\mathcal{C}_s)}{L_r(\Omega)} = \sum_{n\in\mathcal{C}_s} g_r(n) \tag{2.10}$$

Siguiendo con la analogía, la pureza de locutor para un locutor $\boldsymbol{r}$ puede obtenerse como la máxima frecuencia de locutor para cada uno de los clusters existentes. Matemáticamente puede expresarse como:

$$P_{speaker}(r) = \max_{\mathcal{C}_s}(g_{\mathcal{C}_s}(r)) \tag{2.11}$$

y para finalizar, se puede definir la pureza de locutor para la totalidad del segmento $\mathbf{\Omega}$, en función de un reparto hipotético H, mediante la media ponderada de las purezas de locutor para cada hablante:

$$P_{speaker}(\Omega|H) = \frac{\sum_{r=1}^{R} L_r(\Omega) P_{speaker}(r)}{L(\Omega)} \tag{2.12}$$

Al igual que su homóloga para el cluster, esta medida tampoco es capaz de aportar la totalidad de la información por sí misma. Sin embargo, el empleo conjunto de ambas si puede ser considerado una opción muy válida, ya que son dos medidas complementarias.

En la literatura es común emplear estos conceptos, pero mediante su contrapartida, debido en parte a la costumbre de evaluar mediante el error de funcionamiento, las impurezas de cluster y locutor. Éstas se definen así:

$$I_{cluster}(\Omega|H) = 1 - P_{cluster}(\Omega|H) \tag{2.13}$$

$$I_{locutor}(\Omega|H) = 1 - P_{locutor}(\Omega|H) \tag{2.14}$$

Todos estos conceptos se explican con más detalle en el anexo E

Capítulo 3

Diseño de la Etapa Experimental

Vista la parte de documentación, este capítulo expondrá los experimentos a realizar a lo largo de la elaboración del proyecto. Para ello se incidirá en las técnicas empleadas, empezando desde un sistema de referencia hasta las técnicas más novedosas, justificando el porqué de su uso. Se describirán tanto las bases de datos como los evaluadores empleados, aparte de la explicación propiamente dicha de las técnicas a experimentar. Finalmente, se explicará la planificación de la experimentación.

3.1. Bases de datos

Como bases de datos, se procederá al empleo de los datos propuestos para la **Evaluación de Segmentación de Audio de Albayzin en 2010**, procedente de la radiodifusión del canal de televisión 3/24 ([Butko and Nadeu, 2011]).

Esta base de datos se compone de **veinticuatro sesiones** numeradas, formadas por un único canal de audio muestreado a $16KHz$ aunque durante la totalidad del trabajo se optará por emplear un muestreo de $8KHz$, suficiente para almacenar la información más significativa de la voz humana. La base de datos consta de dos particiones:

- **Datos para desarrollo**. Este conjunto de datos se empleará en la fase de desarrollo, estableciéndose con ellos todas las configuraciones de los sistemas. Para este experimento, el conjunto de datos para desarrollo estará conformado por las sesiones numeradas de uno a dieciséis.

- **Datos de Test**. Este conjunto de datos se empleará exclusivamente para la fase de test. Aplicando la configuración de parámetros de nuestro sistema, obtenida en la etapa de desarrollo, se evaluarán los datos de test. Este set de datos lo compondrán las sesiones numeradas de diecisiete a veinticuatro.

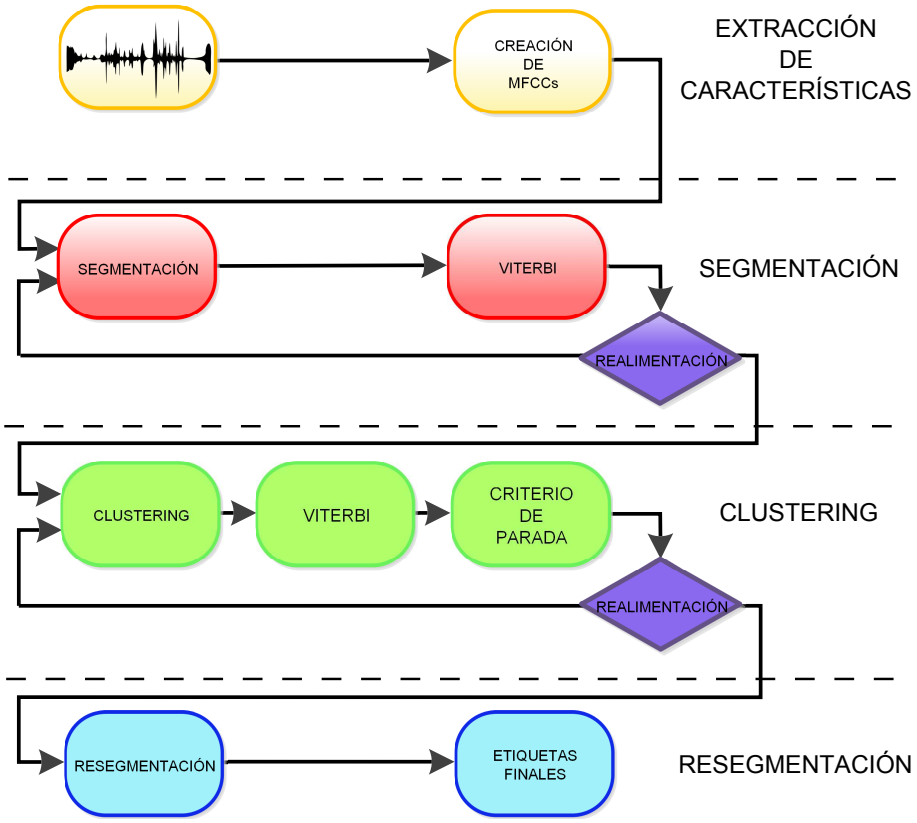


Figura 3.1: Esquema general diseñado para el proyecto

3.2. Evaluación

Como medio de evaluación se ha optado por el empleo de la medida DER, ya que aporta una visión global del error de nuestro sistema. Esta evaluación será llevada a cabo mediante un software evaluador proporcionado por **NIST** (National Institute of Standars and Technology). Este evaluador calcula el término de error DER, exponiendo además dicho valor descompuesto en tres componentes distintas (Miss Error, False Alarm Error y Speaker Error). El error de solape (overlap) queda incluido en los términos citados (en Miss Error si no se detecta un segundo locutor y en Speaker Error si se confunde de cluster). Este trabajo trabaja siempre sobre la hipótesis de detector de actividad vocal o VAD perfecto, por lo que este trabajo se centrará empleará el valor de Speaker Error.

3.3. Estrategias experimentadas

Ya expuestos los datos sobre los que trabajar, en las siguientes secciones se irán mostrando las diferentes soluciones adoptadas a resolver cada una de las etapas del proceso de Diarización. Todas ellas seguirán el modelo de empleo presente en la figura 3.1.

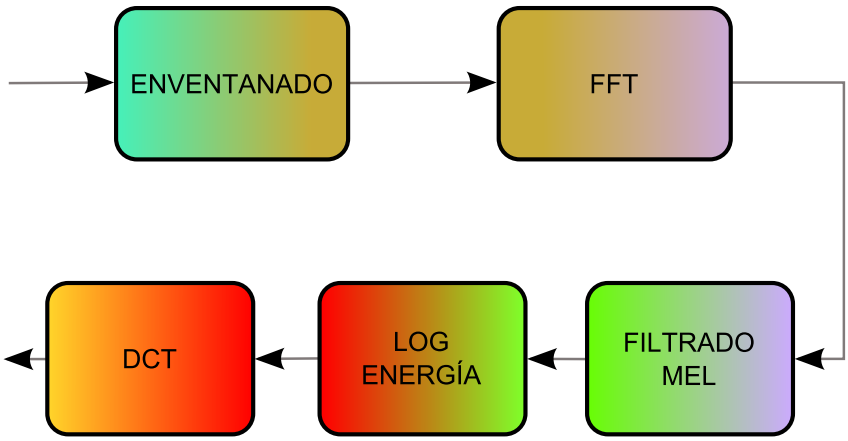


Figura 3.2: Esquema de elaboración de MFCC

Debido a la magnitud y complejidad de los sistemas de Diarización, así como de la cantidad de opciones existentes, se ha optado por focalizar los esfuerzos en la etapa de Clustering, la más crítica en un entorno Broadcast al tener que lidiar con un número de locutores desconocido, por lo que en las otras etapas se han tomado elecciones previas a la experimentación, en vez de estudiar experimentalmente las diferentes variantes existentes.

3.4. Extraccción de características

Pese a ser una elección *a priori* y por tanto, no estudiadas diversas variantes como solución, la elección de las características a emplear puede ser una decisión de vital importancia para los resultados de este trabajo.

Para este problema concreto, se ha optado por la elección de **MFCC** (*Mel Frequency Cepstral Coefficients*). Son una aproximación al dominio cepstral, el cual facilita el estudio del modelo de producción humano, combinado con la aplicación de filtros Mel, que emulan la respuesta del oído, tratando de imitar al ser humano. La formulación matemática puede encontrarse en [Huang et al., 2001], aunque en la figura 3.2 puede verse un esquema de los procesos que aplica. Se trata de una técnica muy contrastada en la bibliografía existente dentro de la rama de tecnologías del habla.

Para este proyecto, se estudiará la influencia de los coeficientes en la separación de locutores. Inicialmente se extraerán para cada ventana de veinticinco milisegundos un total de diecinueve coeficientes, incluido el coeficiente C_0, coeficiente sobre el que existen ciertas discrepancias en la bibliografía , así como el logaritmo de la energía. El desplazamiento de la ventana de estudio será de diez milisegundos. Como primera tarea de experimentación se estudiará cuál es la combinación de coeficientes más oportuna para un medio de Broadcast.

3.5. Segmentación

Esta es otra de las etapas que no van a ser estudiadas en profundidad en este proyecto. No obstante, no puede obviarse. Por lo tanto, en primera aproximación se considerará la etapa de Segmentación como ideal, donde cada segmento solo contiene audio de un locutor, empleando una configuración de **oráculo**, pues se acudirá a la referencia para obtener la segmentación perfecta. La razón principal es independizar la etapa de estudio, Clustering, de cualquier tipo de error procedente de etapas previas (Segmentación).

En la conclusión del trabajo se buscará la elaboración de un **sistema de Dizarización real y completo**. En esa fase se optará por una segmentación basada en distancia, optando por una segmentación basada en $\boldsymbol{\Delta BIC}$, ya que existe una gran bibliografía al respecto, así como la ventaja de ser referencia en otros muchos trabajos. Dado que esta etapa no va a ser estudiada en profundidad, se aplicará una configuración ya empleada en otros trabajos ([Vaquero, 2011]), aunque se dejará la posibilidad a una pequeña etapa de resegmentación mediante ΔBIC, combinando segmentos contiguos dando lugar a otros de mayor longitud.

3.6. Clustering

Una vez finalizada la tarea de segmentación, y siguiendo el esquema 3.1, se debe realizar la labor de Clustering. Dado el carácter general del proyecto, y vistas las características de los diferentes métodos de clustering, se ha decidido focalizar el trabajo en los **basados en métricas**, pues aportan mayor robustez. Dentro de este grupo, se ha optado por la filosofía AHC, en su estrategia Bottom-Up. Aportan grandes ventajas en cuanto a sencillez y modularidad, ambas muy beneficiosas.

3.6.1. Aglomeración Jerárquica (AHC)

Este apartado reflejará todas las técnicas estudiadas en este proyecto en lo referente al proceso de clustering mediante una filosofía AHC, empleando un estilo Bottom-Up, es decir, comenzar con la segmentación más fina, para ir fusionando los distintos segmentos cuando el locutor activo sea común. La figura 3.3 presenta un esquema general de la funcionalidad que un subsistema de este tipo debe realizar para acometer con su deber.

De todos los elementos presentes en nuestro sistema de clustering genérico, se obviará la rama destinada al refinamiento mediante algoritmo de Viterbi, ya que su función es refinar los fallos de la etapa de Segmentación.

Siguiendo el esquema propuesto, se pueden observar dos tareas a realizar:

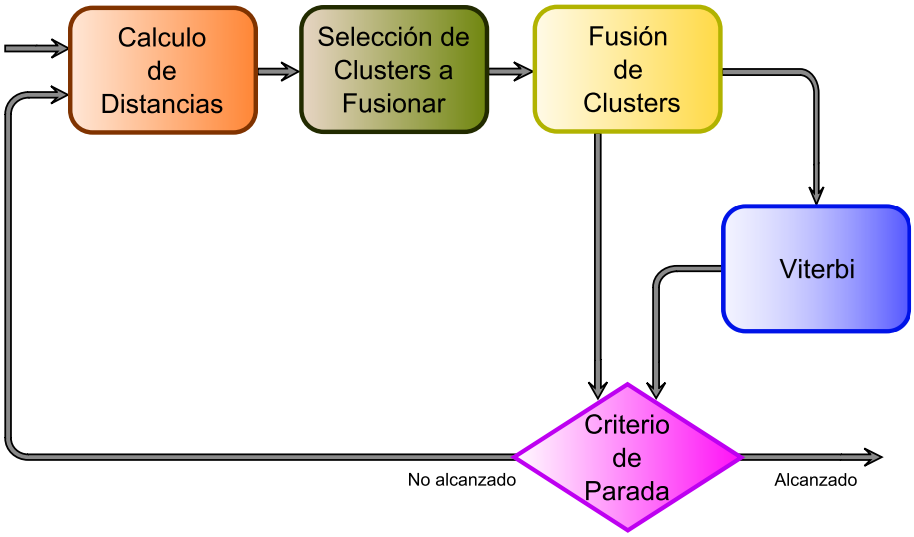


Figura 3.3: Esquema general de un sistema de Clustering mediante AHC

- **Generación de árboles de decisión**. En esta etapa se tratará de perfeccionar las elecciones que el sistema debe tomar acerca de qué clusters fusionar. Una ventaja de la filosofía AHC es la de permitir la elaboración de **árboles de decisión**. Son estructuras de datos que van almacenando toda la información de las distintas elecciones que el sistema va tomando en su proceso iterativo en diferentes niveles, uno por elección. Entonces el criterio de parada solo debe establecer qué decisión es la última que debe ser válida.

- **Elección y comprobación del criterio de parada**. Esta etapa comprobará si el sistema ya ha llegado al número estimado de locutores, en el cual el sistema debe parar de fusionar. En el ámbito de Broadcast, este valor es desconocido, así que el sistema además tendrá que inferirlo. Esta etapa puede diseñarse para actuar iterativamente sobre el sistema, o ser acometida esta labor una única vez, si se aprovechan la estructura de árbol antes comentada.

3.6.2. Generación de árboles de decisión

Como primer paso en el proceso de clustering será establecer un árbol de decisión, que almacenará las diferentes elecciones que paso a paso nuestro sistema tomará.
La experimentación consistirá en la búsqueda de la mejor configuración de coeficientes, empleando una métrica de referencia, y posteriormente, en función de las problemáticas que muestren, se buscarán diferentes estrategias, para corregir los defectos. No obstante, se puede predecir varias fuentes de error, por la naturaleza de los datos:

- **Segmentos cortos**. El empleo de métodos estadísticos trae consigo una problemática que no se puede obviar: Los segmentos más pequeños no conseguirán estimaciones de los

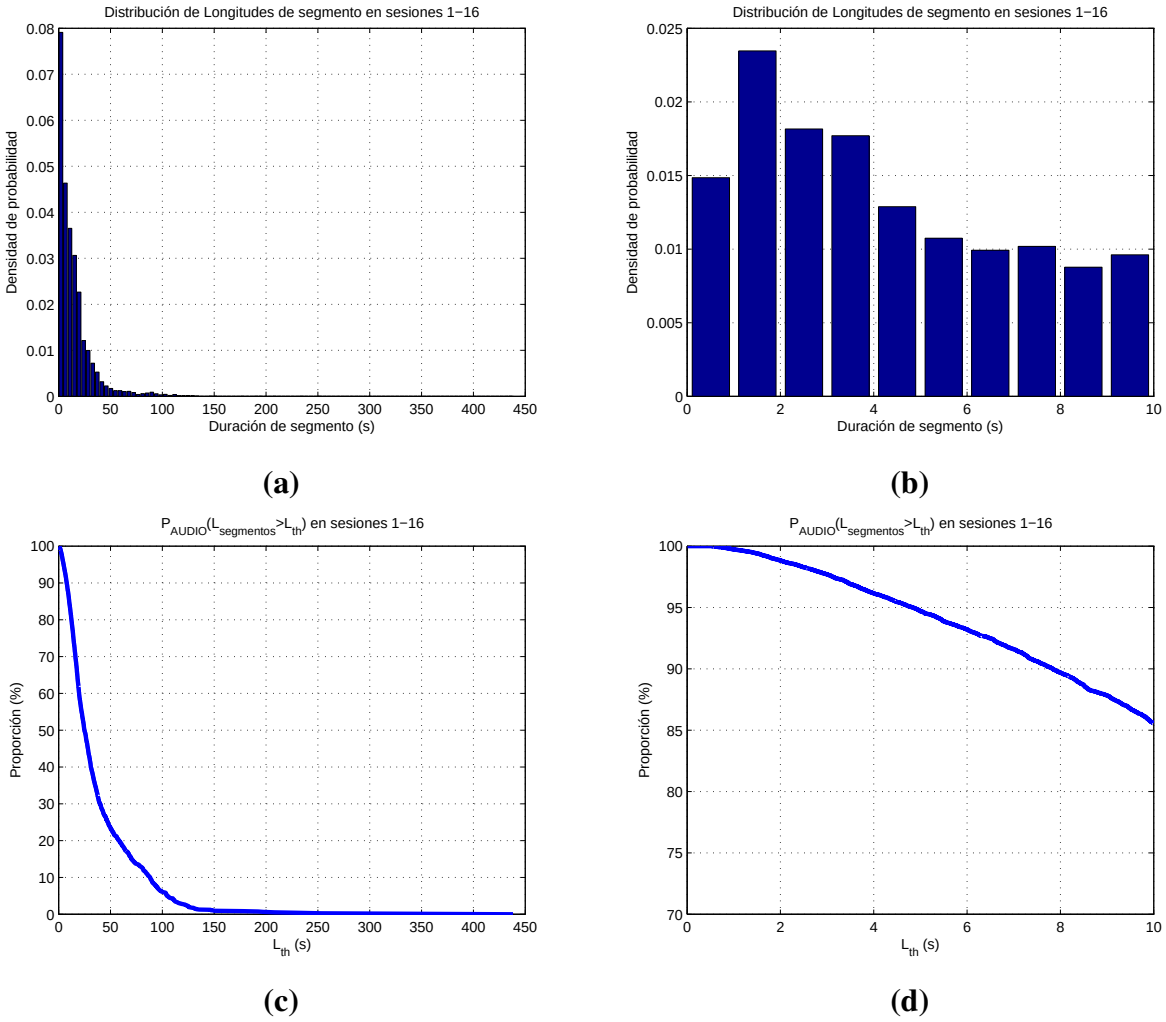


Figura 3.4: Histograma de longitudes de segmento (a), Histograma de longitudes de segmento acotado a 10 s. (b), Proporción de audio contenido en segmentos mayores a L_{th} (c) y Proporción de audio contenido en segmentos mayores a L_{th} acotado a 10 s (d) para sesiones de entrenamiento 1-16, extraidas mediante referencia

parámetros muy robustas para los diferentes modelos, siendo esto más crítico conforme dichos parámetros, muchas veces momentos centrados, requieran un orden mayor. La figura 3.4 muestra que el número de segmentos de este tipo es elevado para un entorno de Broadcast, aunque la cantidad de audio que contienen es baja, por lo que los mayores errores que generen se deberán a la contaminación de clusters con más datos.

- **Segmentos largos**. Si los segmentos cortos no son beneficiosos, los segmentos extremadamente largos tampoco lo son. Las intervenciones excesivamente largas causan variaciones de la voz a lo largo del tiempo, generando modelos más variantes. Estos modelos no tan precisos aumentan la probabilidad de confusión a la hora de fusionar clusters. Además, debido tanto a la cantidad de audio que contienen, sus fallos son más relevantes

que aquellos causados por segmentos cortos.

Todas las métricas mencionadas a continuación y aplicadas durante la experimentación están descritas en el anexo C.

3.6.2.1. Estudio de coeficientes para la referencia

Como primer paso dentro del estudio de la etapa de Clustering, se debe establecer una referencia. Para llevar a cabo esta función, se ha optado por la métrica ΔBIC, basada en BIC. Se trata de una de las técnicas más empleadas en la bibliografía, de escasa complejidad, y con abundante información acerca de ella, con lo que es una referencia idónea.
Con esta técnica se estudiará qué configuración de MFCC es la más apropiada para nuestro estudio de Diarización. Se estudiarán aquellas configuraciones $c_{init} - c_{fin}$, siendo fin un valor entre 11 y 18, mientras init será 0 o 1, estudiando por tanto la idoneidad del coeficiente c_0, sobre cuya idoneidad existen discrepancias en la bibliografía.

3.6.2.2. Tratamiento de segmentos cortos

La predicción de errores por segmentos cortos, motiva a buscar sistemas que minimicen su efecto, ideándose una serie de estrategias para combatirlos:

- **Descarte temporal de segmentos** Esta propuesta constituye la gran aportación de este proyecto. El proceso de Clustering explicado se realizará en dos partes. Primero, y previo descarte de los segmentos más cortos, se realizará una primera aglomeración, generando un árbol de decisión y una inferencia del número de locutores. Después, y tras readmitir los segmentos más cortos, se procederá a una segunda pasada, donde se comprobará como se combinan los segmentos cortos en los diferentes clusters ya establecidos. La figura 3.5 refleja gráficamente el modo de realización del proceso anterior.

  De esta manera se consigue construir etiquetados parciales con segmentos de larga duración, a priori más robustos frente a contaminaciones y errores producidos por segmentaciones no ideales. Posteriormente, con una base sólida, se readmitirán aquellos segmentos más cortos, menos robustos en cuanto a modelar su información, ya que en ese momento no podrán degradar tanto.

- **Empleo de métricas diferentes a ΔBIC** Como se trata de combatir los segmentos cortos, se puede recurrir a técnicas específicas que en teoría mejoran los resultados respecto a las técnicas de referencia. Dentro de este campo están:

  - **Hotelling $\mathbf{T^2}$** . Técnica diseñada para segmentos de corta duración. Su empleo queda a expensas de la validez del descarte temporal de los segmentos más cortos, ya que es poco precisa con segmentos de largos.

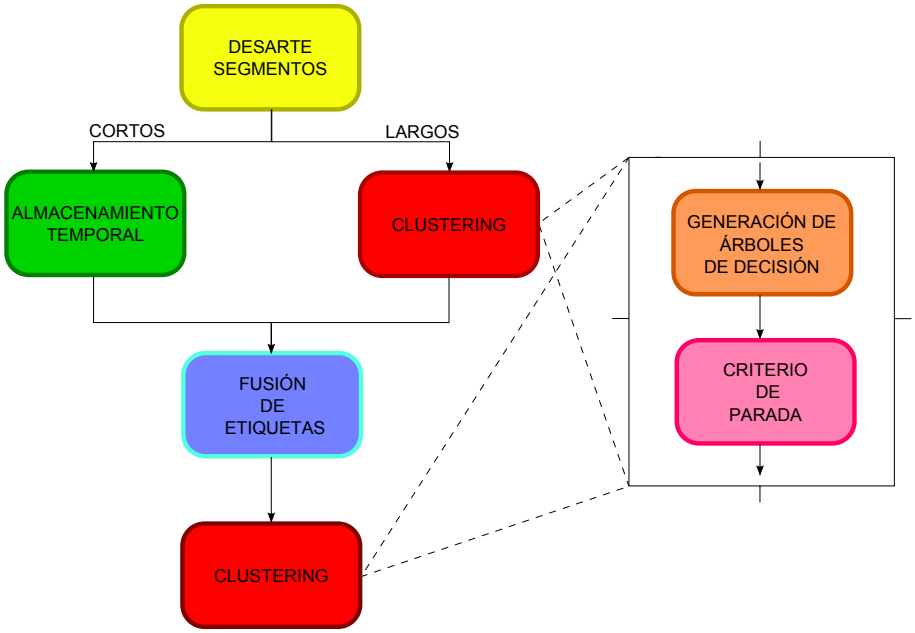


Figura 3.5: Esquema del sistema de Clustering con tratamiento de segmentos cortos

- **Combinación de las técnicas T^2 y ΔBIC**. Citada en [Huang and Hansen, 2006]. Busca aplicar cada técnica , T^2 y ΔBIC, a los segmentos apropiados, cortos y largos respectivamente.

3.6.2.3. Tratamiento de segmentos largos

Para este problema las ideas a aplicar deben ser completamente diferentes. En este caso, los segmentos contienen gran cantidad de información, pero su longitud los hace más sensibles al ruido. Por lo tanto, se deberá tratar de explotar cualquier tipo de relación entre los datos que nos aporte información del locutor, independiente al ruido. Por ello, las diferentes estrategias a estudiar son:

- **Aprovechamiento de la localidad temporal**. El entorno de Broadcast se caracteriza en parte porque en períodos cortos de tiempo unos locutores son mucho más activos que otros. La labor de diarización para dichos períodos exclusivamente se simplifica, ya que disminuye el número de combinaciones posibles a valorar, y la mayoría de las fusiones correctas se deben a unos pocos locutores. Posteriormente, solo deben combinarse los resultados parciales de los diferentes períodos. Como desventaja se limita durante la primera pasada el rango de búsqueda de mejor fusión, obligando además a trabajar sobre etiquetados ruidosos durante la combinación de etiquetados parciales.

- **Empleo de distintas métricas** Tal como se ha aprovechado anteriormente, también se ha optado por probar otras métricas, que siguiendo diferentes filosofías, o variantes que robustecen nuestro sistema, tratarán de obtener mejores resultados. En el caso de que el

sistema de descarte temporal antes comentado sea viable, los sistemas pueden limitarse a ser robustos solo con segmentos largos, en vez de tener un comportamiento global adecuado. Las técnicas a estudiar son:

- **Compensación de la correlación Inter-frame** o **IFCC** siendo el acrónimo de su nombre en inglés (Inte-Frame Correlation Compensation). Es una variante de ΔBIC más robusta mediante la compensación de la correlación entre MFCC calculados a partir de ventanas solapadas.
- **T-test de Tstudent**. Basado en [Nguyen et al., 2008]. Aplicado a dos poblaciones, T-test ha sido escogido para explorar una nueva filosofía. Esta estrategia escoge qué clusters empleando modelos de muy elevada complejidad, empleando modelos globales o UBM (Universal Background Models).

3.6.3. Técnicas para la elaboración de un criterio de parada

Una vez elaborados los árboles de decisión, es necesario establecer en qué nivel del árbol permanecer. Este nivel de permanencia estará ligado al número de locutores presentes en el audio de partida. Estos sistemas se reparten en dos grandes subconjuntos: Aquellos que dadas varias distribuciones (diferentes etiquetados para nuestro caso), escogen aquel que mejor se ajusta a los datos, y aquellos que infieren el número a partir de todas las muestras, sin valoraciones previas. Todos los métodos explicados a continuación pueden hallarse explicados más detenidamente en el anexo D.

3.6.3.1. Criterios de parada a partir de los etiquetados

Esta filosofía trata de generar un criterio de parada a partir de los diferentes etiquetados que contienen los árboles de decisión. El objetivo es estimar, de alguna manera, qué nivel del árbol se ajusta mejor a los datos, pues cada nivel contiene un número de locutores distinto. Dentro de este grupo de técnicas se probarán:

- **Umbral respecto a la distancia entre clusters**. O criterio de información local. Es la primera técnica aplicable en un entorno real. Ha sido muy empleada en la bibliografía por su simplicidad, por lo que será nuestra referencia. Se valdrá de la métrica empleada para fusionar clusters, un criterio de información que para cada fusión realizada indica la mejora de modelar dos clusters conjuntamente respecto a hacerlo por separado. Se ajustará un valor mínimo (**umbral**) de esta mejora que toda fusión a realizar debe cumplir, pues las fusiones correctas implican un beneficio de modelado. Al encontrar una fusión cuya mejora es menor que dicho valor o umbral, se finalizará la tarea de Clustering sin llevar a cabo esta última fusión.

- **Criterios de Información Global**. Empleando criterios de información, independientes a la métrica de fusión, se calculará cuan bien modelados están los datos mediante los diferentes niveles del árbol, escogiendo aquel que mejor modele los datos. A diferencia del umbral, que evalúa exclusivamente los datos de los clusters a fusionar, esta medida evalúa toda la sesión, aparte de recorrer siempre desde la partición más fina a la más gruesa. Dentro del abanico de estadísticos posibles, destaca BIC.

3.6.3.2. Criterios de parada a partir de las características

Las técnicas anteriores presentaban caraterísticas muy interesantes, pero presentaban un gran defecto. Una mala tarea de fusión puede impedir estimar adecuadamente el número de locutores. Por ello, se debe buscar una alternativa: Estimar el número de locutores únicamente a partir de los datos. Dentro de este conjunto de técnicas, se probará: **Mean-Shift**. Su base es estudiar la distribución estadística de las características como si se tratara de una curva de nivel, considerando que todo máximo idealmente corresponde con un locutor en el audio. Se buscarán dichos máximos maximizando el gradiante.

3.7. Resegmentación

Como última tarea, se aplicará una técnica de refinado, denominada Resegmentación. Asumiendo cada locutor modelado por un GMM (Gaussian Mixture Model), la resegmentación consistirá en un HMM (Hidden Markov Model), que emplea como distribuciones los GMM estimados para cada locutor. Además, aplicará el concepto de *tied-states*, que ajustará una mínima permanencia en los locutores para ser mas realista, desarrollado en [Levinson, 1986].

En las primeras fases del proyecto esta etapa se obviará, para no afectar a los resultados de Clustering, ajenos a esta etapa. En las últimas etapas del proyecto, cuando se busque estudiar un sistema completo, sí se incluirá esta etapa.

3.8. Fases de la experimentación

Dada la gran extensión del trabajo a desarrollar, es necesario ordenar toda la carga experimental en diversas fases, con el fin de optimizar el esfuerzo y el trabajo a desarrollar.

Por todo ello, el orden en el cual la experimentación ha sido realizada queda reflejado en la figura 3.6, donde se observan cinco fases diferenciadas, desarrollándose diferentes tareas en cada una de ellas.

Las dos primeras equivaldrán a la creación de una referencia fiable, respecto a la cual comparar. Las dos siguientes consistirán en la profundización de la etapa de Clustering, trabajando en profundidad la técnica de descarte temporal, principal aportación de este proyecto, así como

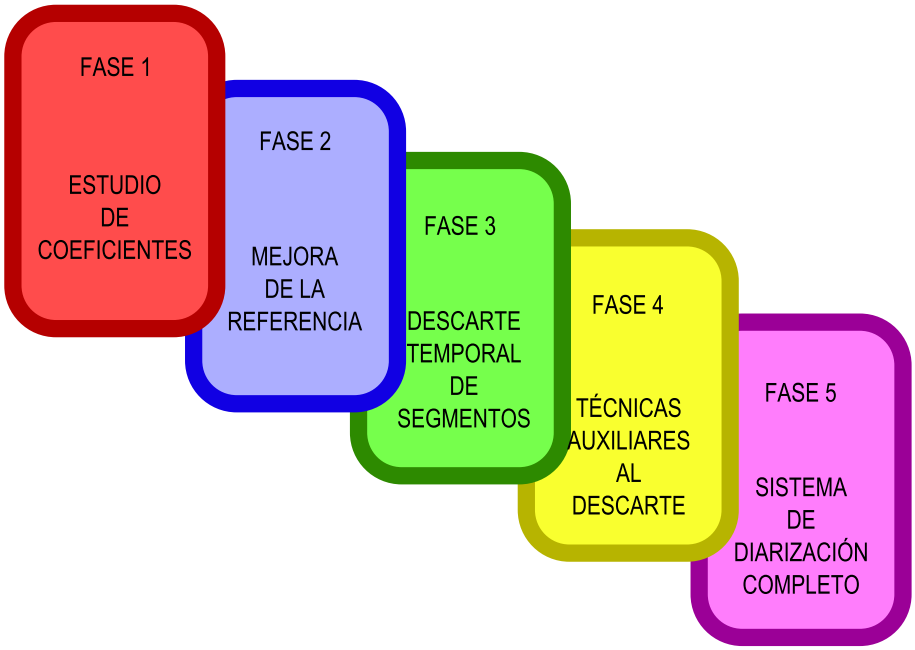


Figura 3.6: Objetivos de cada fase de experimentación

de otras técnicas, y finalizando con una última fase dedicada a la resegmentación, haciendo posible la comparación con otros trabajos.

3.8.1. Estudio de coeficientes

Como primer paso para este trabajo, se deberán estudiar aquellos coeficientes que mejor discriminan los locutores activos. Para este trabajo, se ha restringido un rango de coeficientes inicial bastante amplio, y es preciso hacer un estudio en profundidad. Este proceso se llevará a cabo en esta fase. Además, se comprobarán los efectos de esta configuración de coeficientes para los casos con segmentación ideal y estimación de locutores real, así como para ambas condiciones reales.

3.8.2. Mejora de la referencia

Los resultados obtenidos para referencia muestran una serie de problemáticas que impiden cumplir su papel adecuadamente. Por ello, se buscarán alternativas, que sustituirán a sus predecesoras tomando el papel de referencia en adelante.

3.8.3. Descarte temporal de segmentos

Fijada ya una referencia de calidad, se estudiará la eficacia de la técnica principal de este trabajo: el descarte temporal. Se estudiará su eficacia desde las condiciones más ideales hasta las más reales en lo referente a segmentación y estimación del número de locutores.

3.8.4. Técnicas complementarias al descarte temporal

Tanto si la técnica de descarte temporal funciona como si no, se probarán otras opciones para mejorar las tasas de error de la referencia, ya sea apoyando a un confirmado descarte temporal, o sustituyéndolo en caso de no funcionar.

3.8.5. Resegmentación

Finalmente, como última tarea, se aplicará al sistema resultante una etapa de resegmentación, mediante un algoritmo de Viterbi, refinando los resultados obtenidos, y posibilitando la comparación con el resto de sistemas de la bibliografía.

Capítulo 4

Resultados

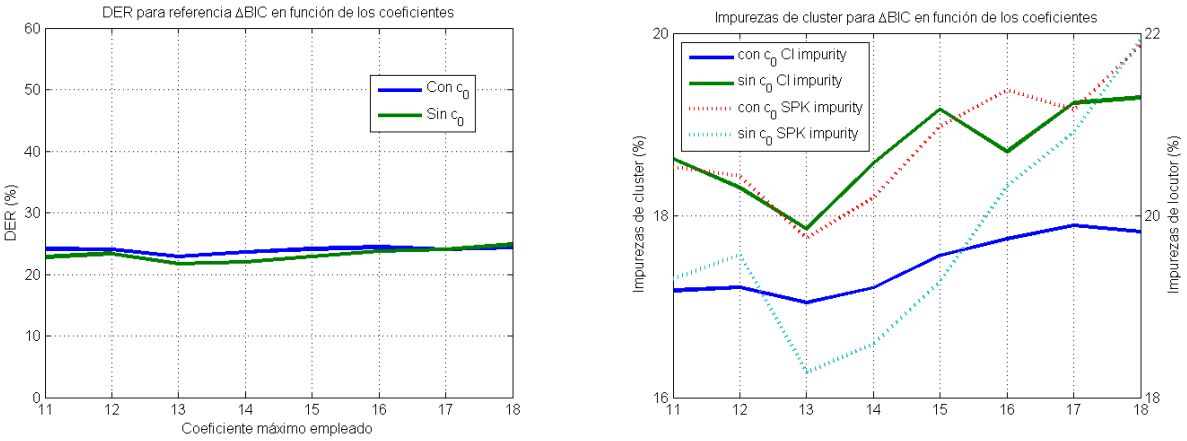
Una vez vistas las diferentes pruebas a realizar, este capítulo será el encargado de albergar los resultados, y donde estos serán analizados.

Los resultados sobre las diferentes etapas de clustering, expuestos según el orden ya descrito en el último punto del capítulo anterior, se estudiarán para tres condiciones de estudio o grados de idealidad, en función de la posibilidad de generar errores de las etapas distintas a la de Clustering:

- Condiciones ideales. Tanto la etapa de segmentación como la de estimación del número de locutores serán llevadas a cabo mediante oráculo, a partir de las referencias disponibles.
- Condiciones intermedias. La etapa de segmentación continúa siendo ideal, mientras el número de locutores deberá ser estimado sin información previa, pudiendo aumentar las tasas de error.
- Condiciones reales. Esta configuración es la más importante, ya que analiza el comportamiento del sistema funcionando en un entorno real. Tanto la etapa de segmentación como de estimación de locutores son susceptibles de cometer errores, aumentando la tasa de error.

4.1. Primera fase: Estudio de coeficientes

Como primera tarea una vez comenzado la carga experimental del proyecto es establecer qué configuración de coeficientes es más productiva para la tarea de Diarización.
Para ello se recurrirá a la métrica de referencia, ΔBIC. De esta forma, se analiza cuan discriminante es cada configuración, obteniendo tanto la configuración más apropiada para radiodifusión como una tasa de error DER a mejorar para el resto del proyecto, referencia para posteriores experimentos.



(a) **(b)**

Figura 4.1: Error e de Diarización o DER (a) e Impurezas (b) para ΔBIC, en función de los coeficientes escogidos

En primer lugar, se procederá a una búsqueda de resultados bajo condiciones ideales para el rango de coeficientes $c_{init} - c_{fin}$, siendo $init = 0, 1$ y $fin = 11, ..., 18$. El coeficiente c_{init} varía para estudiar si el coeficiente c_0 aporta beneficios o no, ya que en la bibliografía existen opiniones encontradas. Los resultados obtenidos son mostrados en la figura 4.1.

Dentro de todas las configuraciones posibles, la subfigura (a) muestra que la mejor configuración es $c_1 - c_{13}$, aunque las diferencias son poco significativas con el resto de configuraciones. Respecto al coeficiente c_{final}, la gráfica refuta la idea general de que aumentar la dimensionalidad del problema tiende a aportar beneficios. Referente al coeficiente c_0, no es preferible su uso, ya que sin él, el sistema no se distrae con fusiones arbitrarias pequeñas, centrándose en aquellas más importantes, aunque las confunda, según los resultados de impurezas.

Aplicando esta configuración de coeficientes para los tres grados de idealidad descritos, los resultados quedan mostrados en la tabla 4.1, convirtiéndose en las referencias con las que comparar los diferentes experimentos, en función del grado de idealidad analizado. Los sistemas aplicados serán:

- ΔBIC como métrica de Clustering.
- Umbral sobre la métrica de fusión como criterio de parada no ideal.
- ΔBIC como métrica de segmentación no ideal.

Los resultados indican que la configuración de coeficientes es relativamente precisa si se permite una estimación de locutores mediante umbral. En cambio, si además se permite una segmentación real, la degradación de resultados es mucho más elevada, por lo que se deberá mejorar este punto.

Configuraciones	Config. Ideal	Config. Intermedia	Config. Real
DER (%)	21.72	24.16	80.48

Tabla 4.1: Resultados de DER para la referencia ΔBIC según el grado de idealidad en las etapas generadoras de error (Umbral como criterio de parada y ΔBIC para segmentación)

4.2. Segunda fase: Mejora de la referencia

Los resultados de la fase anterior muestran que la mayor problemática hace presencia al añadir una segmentación no ideal, haciendo peligrar una referencia mínimamente fiable.

La etapa de segmentación, aplicada a canal telefónico en [Vaquero, 2011], ha sido ajustada para no perder aquellas fronteras que realmente definen fronteras locutor-locutor, pudiendo añadir fronteras falsas. La distribución de estos segmentos, así como la distribución de la cantidad de voz en función de la longitud de los segmentos puede verse en la figura 4.2.

Comparando la distribución obtenida con aquella presentada anteriormente (la figura 3.4), analizada a partir de las referencias, nuestro sistema de segmetación ha tendido a dividir en exceso los segmentos, principalmente aquellos más largos, originando esta degradación. Como única nota positiva, la proporción de fronteras locutor-locutor solo representan un 11 % de las fronteras totales, siendo el resto ajustadas por el VAD, que para todo el trabajo será perfecto, minimizando los daños de nuestra segmentación.

En la situación actual, con la intención de aproximar la distribución ideal, se ha optado por aplicar una etapa básica de resegmentación, basada en ΔBIC, que estudiará generar segmentos más largos a partir de los segmentos contiguos anteriormente obtenidos. Bajo esta estrategia, la distribución de segmentos queda representada en la gráfica 4.3.

Estos últimos resultados indican una gran mejoría como aproximación a la distribución ideal, sobre todo comparando aquella sin resegmentación. La precisión de las fronteras sigue siendo mala, pues depende demasiado del VAD, aunque se ha reducido en gran medida la cantidad de fronteras irreales, dando lugar a segmentos más largos, y más robustos. Con esta configuración, los resultados para los diferentes grados de idealidad quedan reflejados en la tabla 4.2

A pesar de estos cambios, con las mejorías que ocasionan, no son suficientes. Los etiquetados que estas configuraciones aportan indican una mala respuesta del estimador del número de locutores, ya que solo estima un único locutor activo por sesión. Por este motivo se probará otra etapa de criterio de parada, basada en BIC. Con este cambio, se cambia la filosofía

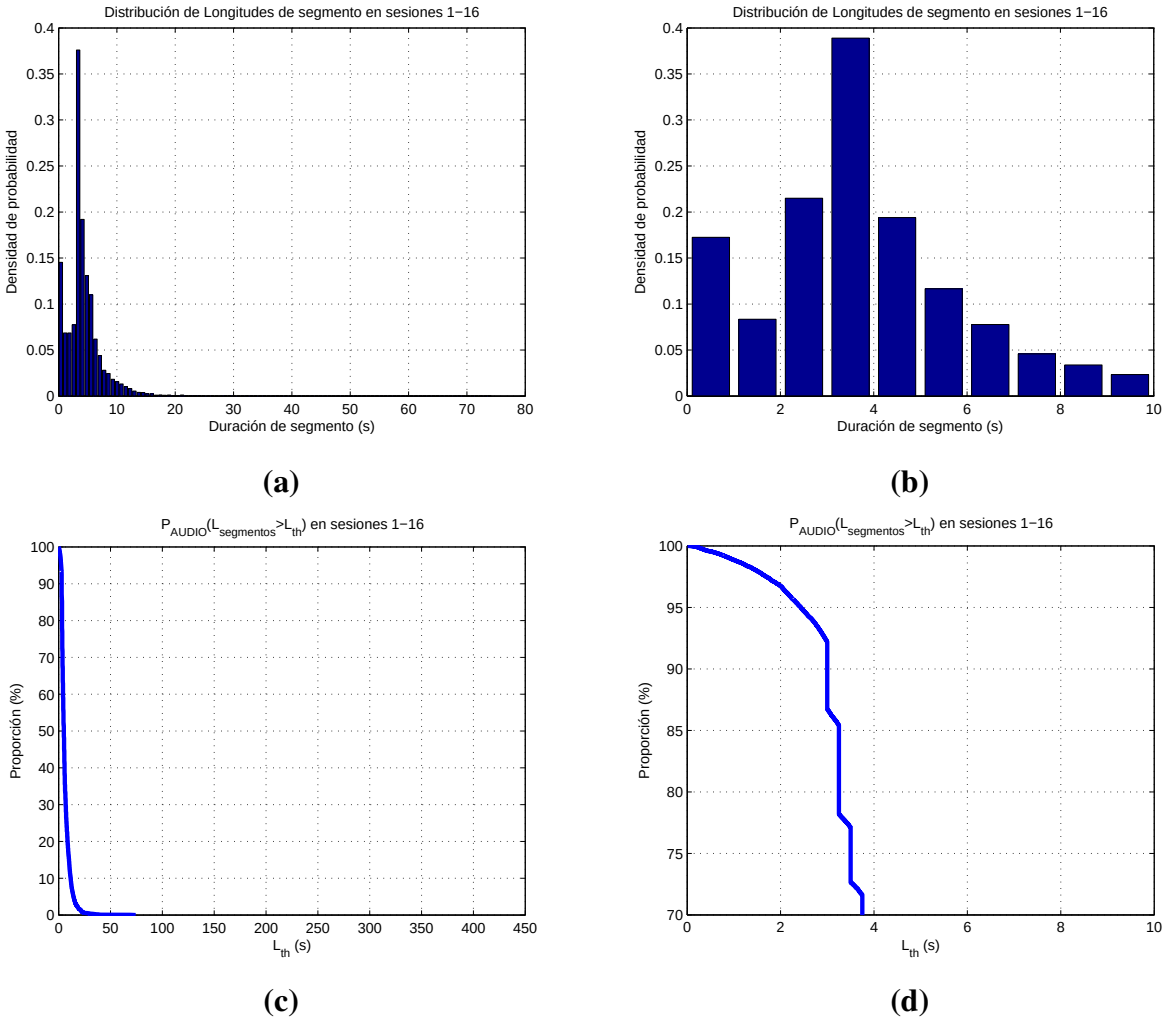


Figura 4.2: Histograma de longitudes de segmento (a), Histograma de longitudes de segmento acotado a 10 s. (b), Proporción de audio contenido en segmentos mayores a L_{th} (c) y Proporción de audio contenido en segmentos mayores a L_{th} acotado a 10 s (d) para sesiones de entrenamiento 1-16, extraidas estrategia de canal telefónico

Configuraciones	Config. Ideal	Config. Intermedia	Config. Real
DER (%)	21.72	24.16	84.88

Tabla 4.2: Resultados de DER para la referencia ΔBIC según el grado de idealidad en las etapas generadoras de error(Umbral como criterio de parada y ΔBIC con resegmentación simple para la labor de segmentación)

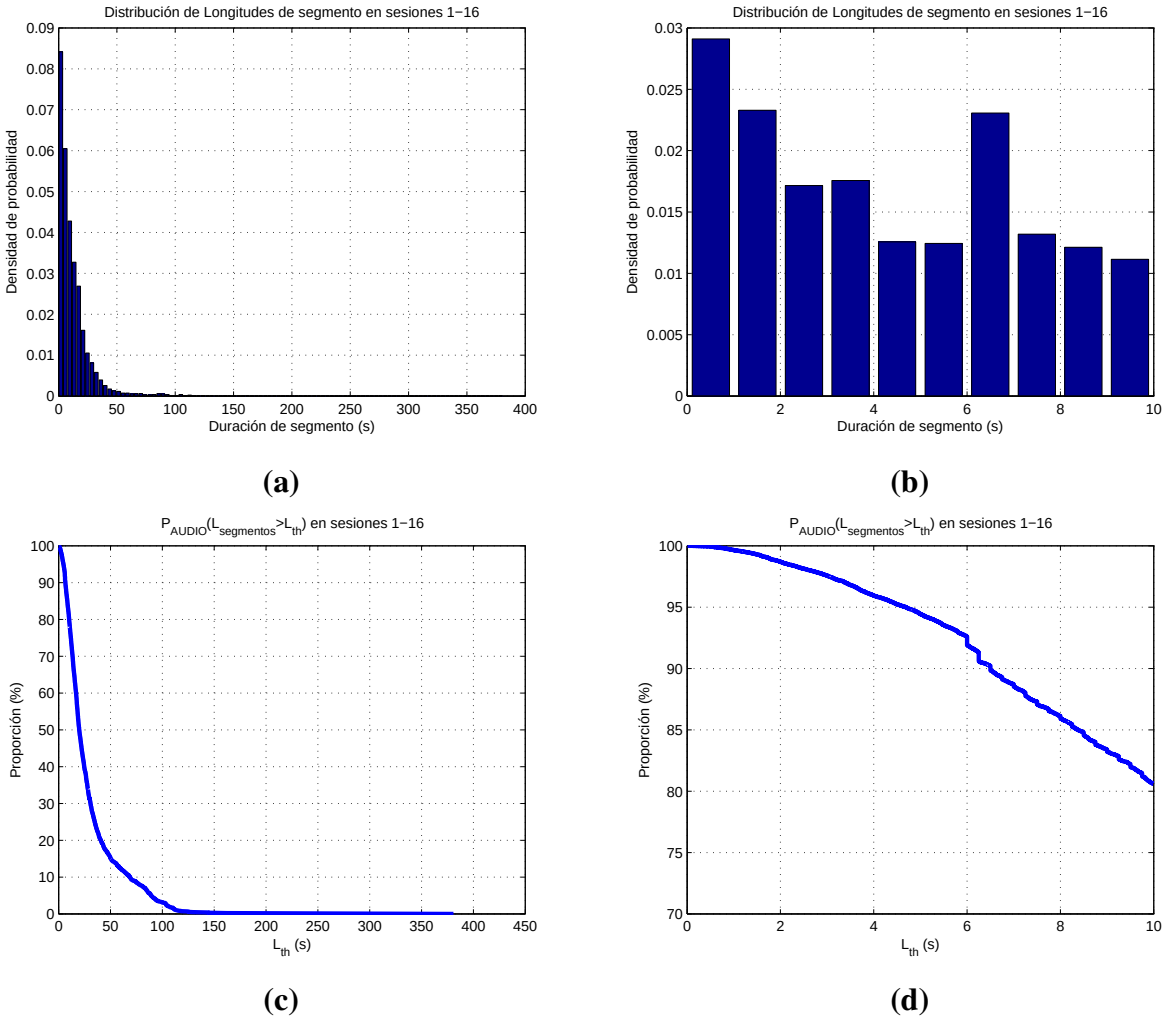


Figura 4.3: Histograma de longitudes de segmento (a), Histograma de longitudes de segmento acotado a 10 s. (b), Proporción de audio contenido en segmentos mayores a L_{th} (c) y Proporción de audio contenido en segmentos mayores a L_{th} acotado a 10 s (d) para sesiones de entrenamiento 1-16, extraidas mediante estrategia de canal telefónico con resegmentación

Configuraciones	Config. Ideal	Config. Intermedia	Config. Real
DER (%)	21.72	24.51	79.80

Tabla 4.3: Resultados de DER para la referencia ΔBIC según el grado de idealidad en las etapas generadoras de error (BIC como criterio de parada y ΔBIC con resegmentación simple para la labor de segmentación)

de estimación de locutores, ya que se pasa de estudiar exclusivamente los clusters fusionados a analizar como repercute la fusión a nivel de sesión de audio. Además, se persigue también mejorar la robustez frente al umbral respecto a la métrica, frágil ante cambios entre datos de entrenamiento y test. Con esta nueva técnica, los resultados obtenidos pueden verse en la tabla 4.3

Este cambio presenta ciertas ventajas. En los experimentos con idealidad intermedia, los resultados son muy parejos a aquellos obtenidos con umbral, y menos dependientes de las diferencias entre datos de entrenamiento y test. Aparte, con condiciones reales, es capaz de discernir al menos la presencia de varios locutores, aparte de tener una distribución de segmentos parecida a la obtenida mediante oráculo.

A consecuencia de los beneficios, así como de las problemáticas que presentaban sus predecesoras, las técnicas que en adelante ejercerán de referencia consistirán en:

- ΔBIC como métrica de Clustering.
- BIC como estimación del número de locutores y criterio de parada no ideal.
- ΔBIC con resegmentación en la etapa de segmentación real.

4.3. Tercera fase: Descarte temporal de segmentos

Establecida una referencia más robusta sobre la que trabajar, se pasará a estudiar la estrategia central de este proyecto, la técnica de descarte temporal, pues se considera de gran importancia la influencia los segmentos de larga duración en la tarea de Diarización. Se buscará aprovechar la mayor cantidad de datos en los segmentos largos, dotando de una mayor robustez a los modelos generados. Con estos segmentos largos se realizará una primera tarea de Clustering, para después añadir a los resultados aquellos más cortos, descartados en un primer momento por ser menores a una longitud mínima, y realizar una segunda pasada.

El rango de estudio D de longitud mínima de segmentos es $D = 0, .., 10$. Se incluye el descarte nulo, para comparar las posibles ventajas de la estrategia. Pruebas preliminares apuntan a que este rango es suficiente.

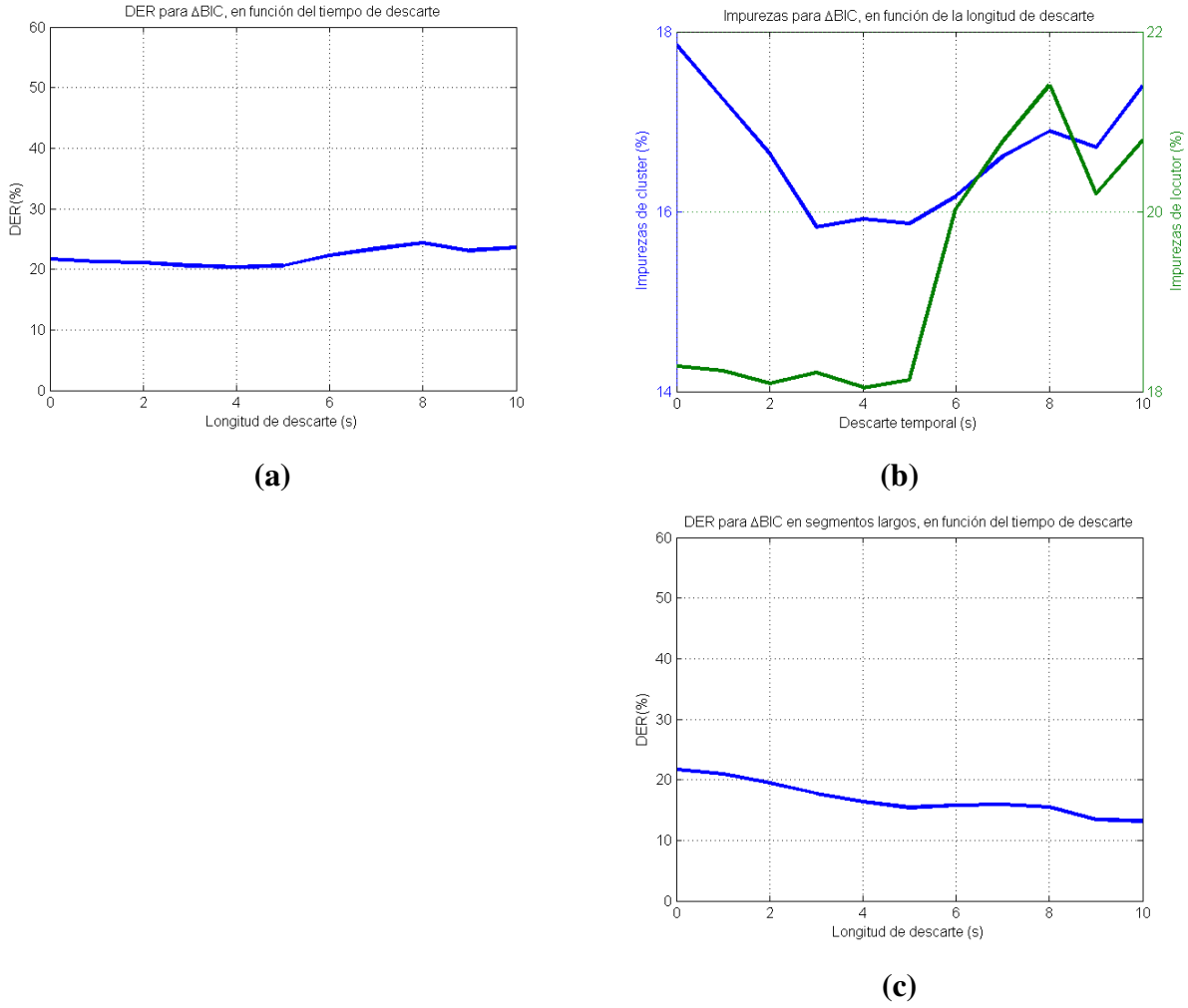
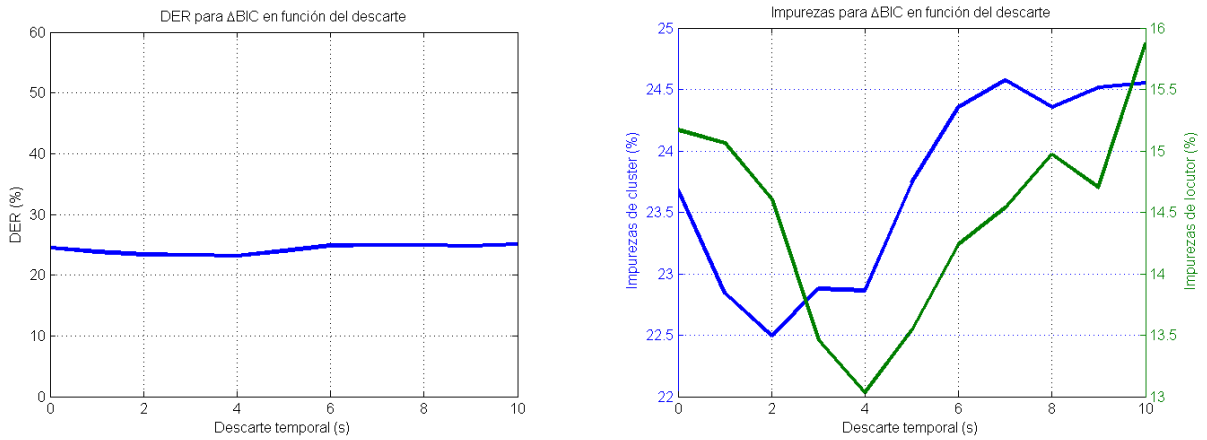


Figura 4.4: DER e Impurezas para ΔBIC a nivel de sesión (a y b respectivamente), y DER para los segmentos largos (c), para segmentación y estimación de locutores ideales, en función del descarte

Empleando esta estrategia, para nuestra configuración de experimentación ideal, los resultados quedan reflejados en la figura 4.4.

Los resultados muestran una pequeña mejoría conforme el descarte aumenta respecto a un descarte nulo, siempre que dicho descarte sea menor a cierto valor. Para descartes mayores, la compleja readmisión de los fragmentos cortos no compensa el beneficio obtenido con aquellos largos, perdiendo prestaciones. Esta mejoría no solo es aplicable a los segmentos más cortos (menores a un segundo), sino que se extiende a segmentos algo más largos. Además, observando, la gráfica 4.4 (c), en conjunción con 3.4, indican que en condiciones ideales la mitad de la tasa de error DER existente se debe a fallos provocados o influenciados por segmentos menores a diez segundos, que en conjunto contienen menos del 20 % del audio total, mostrando que es una fuente importante de error.



(a) **(b)**

Figura 4.5: DER e Impurezas para ΔBIC, en función del descarte a nivel de sesión (a y b respectivamente), para segmentación ideal y estimación de locutores real

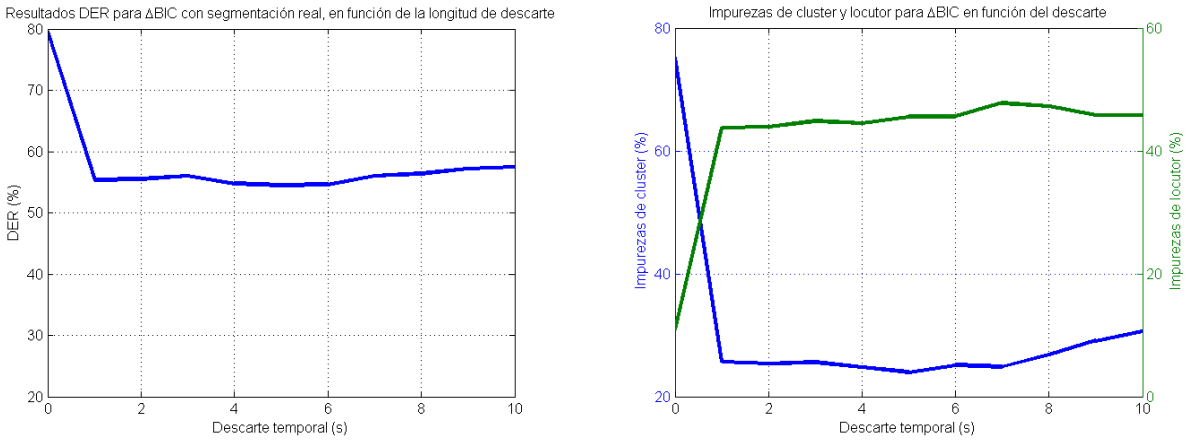
Por lo tanto, se han confirmado varios puntos: Por un lado, se ratifica la mala influencia de los segmentos más cortos, ya sea ensuciando clusters o impidiendo fusiones correctas. Por otro lado, se confirman los beneficios de la estrategia de descarte, tanto obteniendo mejores árboles con los segmentos más largos como dividiendo la dificultad de la tarea de Clustering, aunque cuantitativamente no tienen mucha incidencia.

Realizando este mismo análisis para la configuración de experimentación intermedia , los resultados son mostrados en la figura 4.5.

En este caso, los resultados son muy parecidos a los obtenidos con configuración ideal, salvo por un pequeño incremento, fruto de los errores por estimación de locutores. Este aumento procede de detectar menos locutores que antes, causando los fallos. Esta circunstancia indica que existen fusiones incorrectas a realizar que en teoría aportan mejoras de los etiquetados a nivel de sesión, y por lo tanto el sistema las realiza. Se deduce de ello que las últimas fusiones involucran a clusters muy parecidos, por lo que fácilmente son confundidas.

Finalmente, se ha probado la estrategia de descarte en condiciones reales (segmentación y estimación de locutores reales). Bajo estas condiciones, los resultados quedan reflejados en la figura 4.6

La gran confirmación de esta técnica se produce bajo condiciones reales. A costa del más mínimo descarte, la tasa de error se reduce un 25 %. Esto se debe a que los segmentos más cortos, y más susceptibles a contaminaciones por una segmentación errónea, degeneran en gran medida la tarea de clustering, siendo más oportuno trabajar con aquellos más largos y



(a) **(b)**

Figura 4.6: DER e Impurezas para ΔBIC, en función del descarte a nivel de sesión (a y b respectivamente), para segmentación y estimación de locutores reales

robustos frente a la falta de información y una mala segmentación. Además, esta ventaja cobra más importancia conforme los errores de segmentación son más elevados, ya que estos errores afectan más dañinamente a los segmentos cortos. Como nota final, la gráfica de impurezas indica que el sistema pasa de sobreagrupar (encontrar menos locutores de los debidos) sin descarte a subagrupar (encontrar más locutores de los existentes).

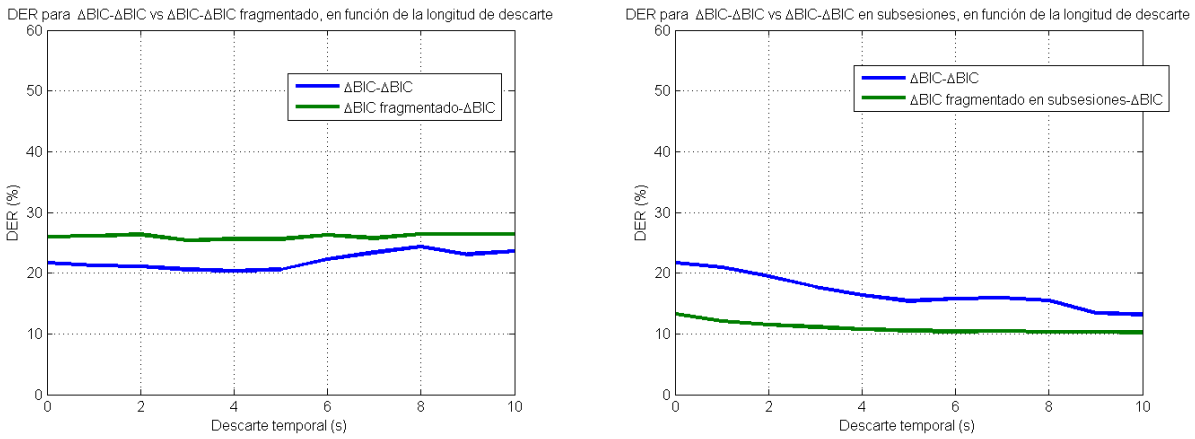
4.4. Cuarta fase: Técnicas auxiliares al descarte

Vistos los beneficios aportados por la técnica de descarte, se estudiarán diferentes técnicas que, sustituyendo o apoyando a ΔBIC, puedan aportar beneficios combinados con la estrategia de descarte. Dado que esta divide la tarea de Clustering en dos subprocesos, uno para segmentos largos y otro para cortos, las diferentes posibilidades también se doblan.

Como nomenclatura a utilizar en adelante se emplearán nombres con una estructura $A - B$ para cada experimentación, tanto en el título como en la leyenda. Esto indicará que el sistema $A - B$ se compone de una etapa basada en A para el trabajo con los segmentos más largos, mientras una etapa basada en B hace lo propio con aquellos más cortos y descartados en un primer momento.

4.4.1. Tratamiento de segmentos largos

La fase anterior ha confirmado que la técnica de descarte implica apoyarse principalmente en la combinación de segmentos largos, más robustos. Sin embargo, el aumento de la tasa de error al pasar a condiciones reales se debe al aumento de fusiones erróneas con este tipo de



(a) **(b)**

Figura 4.7: DER para ΔBIC siguiendo la estrategia de empleo de subsesiones, para las sesiones completas (a), y con las subsesiones (b)

segmentos por lo que se deben estudiar fórmulas que la combatan.

Bajo condiciones ideales de segmentación y estimación del número de locutores, se han pensado diversos métodos para mejorar la fusión de segmentos largos:

- **Emplear la localidad temporal** Se estudiará el audio en segmentos cortos (cinco minutos) solapados. Esto puede permitir mejorar mucho los resultados a nivel de dichos intervalos. Después se combinarán los resultados de los diferentes intervalos. En este caso los resultados están presentes en la figura 4.7.

  Éstos muestran que si bien la idea de aprovechar la predominancia de ciertos locutores es correcta, estudiando las subsesiones por separado (subfigura (b)), al combinar resultados se introduce una degradación, más importante que el beneficio que esta idea genera. Esta degradación puede tener dos procedencias: o la limitación de rango de búsqueda de segmentos, debido al estudio local del audio, o también puede deberse a que el proceso de integración de resultados no parte de unos clusters puros (perfectos).

- **Emplear nuevas medidas más robustas**. ΔBIC es una medida que tiende a comportarse relativamente bien en cualquier situación, aunque no destaque en ninguna. Para trabajar con segmentos de larga duración, la bibliografía presenta:

  - **Compensación de la correlación entre tramas o IFCC (Inter-Frame Correlation Compensation)**. Esta técnica es una evolución de ΔBIC. A costa de añadir una mayor complejidad (un parámetro **r**, multiplicativo a la longitud del segmento)

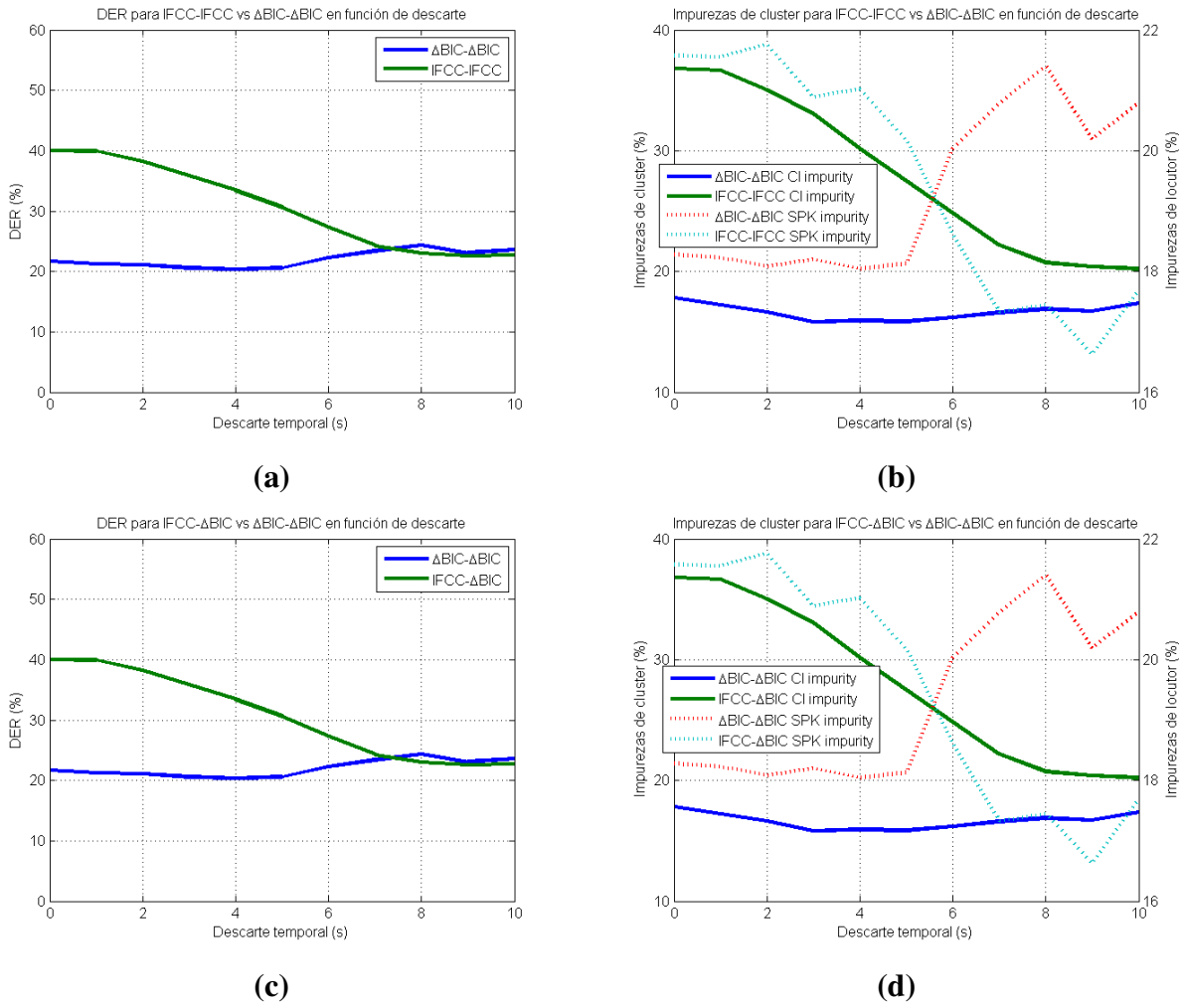


Figura 4.8: DER e impurezas para IFCC-IFCC (a y b), IFCC-ΔBIC (c y d) y IFCC-Combinacion frente a ΔBIC-ΔBIC

buscará buscará compensar la correlación de la extracción de MFCC. Se muestran sus resultados en la figura 4.8.

Los resultados enseñan claramente la gran potencia que presenta esta técnica para fusionar segmentos medianamente largos con respecto a nuestra referencia ΔBIC. Esto se debe a que tiende a priorizar las fusiones con segmentos largos, en vez de estudiar aquellas con segmentos cortos involucrados , llegando a compensar los problemas que éstos últimos generan. Sin embargo, con descartes pequeños el sistema pierde prestaciones. IFCC lo logra disminuyendo tanto las impurezas de cluster como de locutor en gran medida conforme el descarte aumenta. Estudiando su comportamiento con los segmentos no descartados, los resultados pueden verse en la gráfica 4.9

Esta figura nos indica que, conforme aumenta la longitud de descarte, la mejora

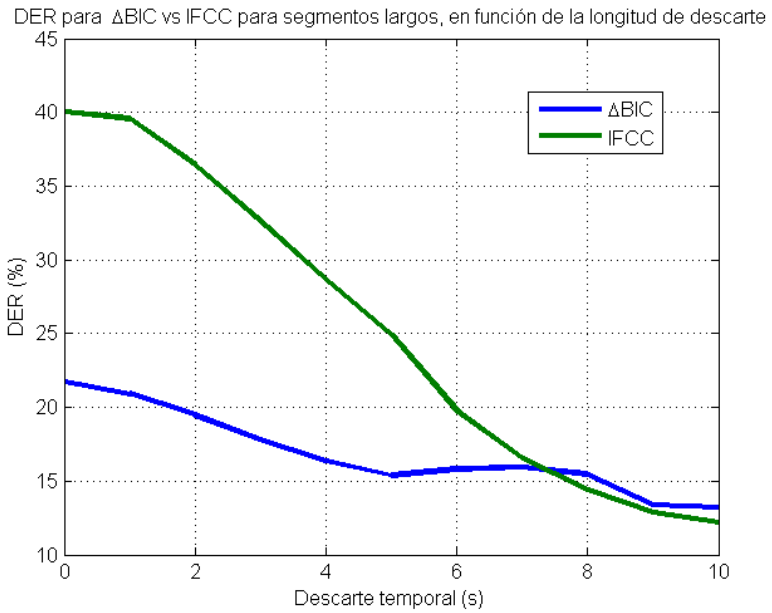


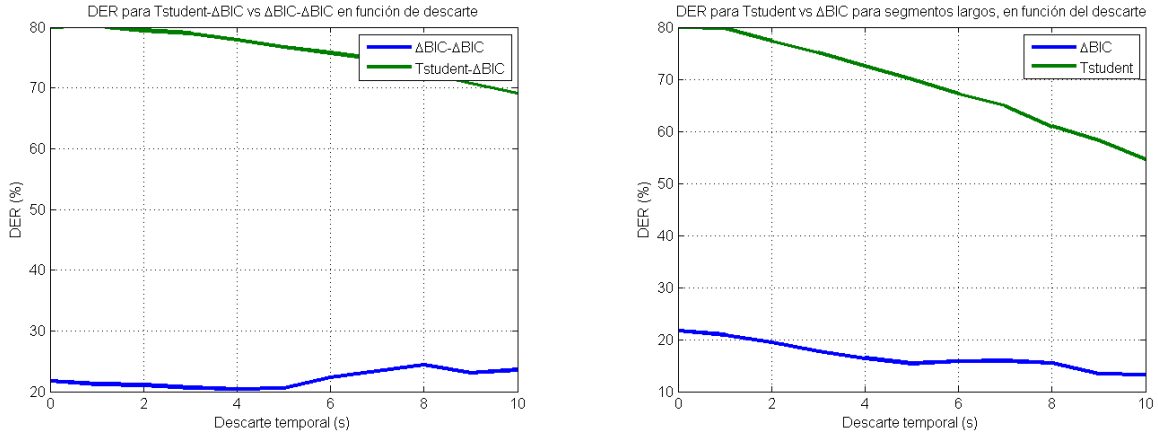
Figura 4.9: DER para ΔBIC y Compensación de la Correlación en clustering de segmentos largos, en función del descarte

es notable, llegando a mejorar los resultados de ΔBIC. La complicación viene por tanto de la tarea de readmisión, muy problemática con descartes elevados.

- **T-test de Tstudent** Otra filosofía de trabajo a probar, en combinación con ΔBIC, que realizará las tareas de readmisión. Se busca emplear modelos más complejos de la verosimilitud de las características. Los resultados con esta técnica pueden observarse en la figura 4.10.

  A la vista de los resultados, pocas conclusiones pueden sacarse, ya que los resultados son muy negativos. Esta técnica tal como se ha empleado no tiene utilidad, al menos con la configuración actual: Emplear una gaussiana multidimensional como PDF a elegir, para combinarla con un UBM (Universal Background Model, en este caso un GMM de 1024 gaussianas). En defensa de esta técnica cabe destacar que el autor del artículo establecía como PDF a elegir un GMM de al menos treinta y dos componentes para su funcionamiento óptimo. Se esperaba degradación por el empleo de un modelo tan sencillo, pero los resultados han sido completamente inesperados. Estos resultados hacen descartar el empleo de T-student como criterio de parada en la siguiente fase, pues requiere de esta medida tan poco precisa.

Los diferentes resultados solo dejan una opción viable para dar el salto a condiciones menos ideales, IFCC. Aporta beneficios con los segmentos largos, aunque presenta un peor comportamiento con los segmentos más cortos.



(a) **(b)**

Figura 4.10: DER para la técnica T-student en comparación a la referencia ΔBIC para la sesion completa (a) y para los segmentos largos (b), en función de la longitud de descarte

Aplicando esta técnica para una configuración de idealidad intermedia, los resultados que se obtienen son los presentes en la figura 4.11

Vuelven a verse los mismos comportamientos que en el caso de la referencia ΔBIC, apreciando una pequeña pérdida de prestaciones respecto al caso de configuración ideal. El sistema tiende a sobreagrupar, detectando menos locutores de los necesarios, y tratando de compensar fusiones erróneas.

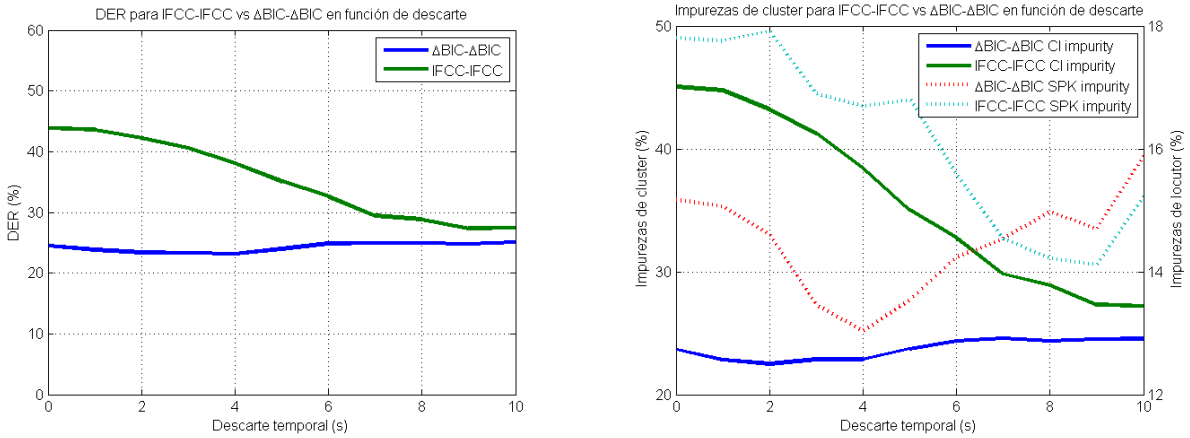
Finalmente, el salto a las condiciones reales, con segmentación y estimación de locutores reales aporta unos resultados mostrados en la figura 4.12

Aqui los resultados reafirman la principal utilidad de IFCC: Trabaja adecuadamente con segmentos largos. En conjunción con la técnica de descarte, que evita la mala influencia de los segmentos cortos, IFCC escoge adecuadamente los segmentos a fusionar, pues se apoya principalmente en los clusters con más información y más robustos. Además, una mejora en la pureza del proceso de fusión simplifica la tarea de estimación del número de locutores, según la gráfica de impurezas.

4.4.2. Tratamiento de segmentos cortos

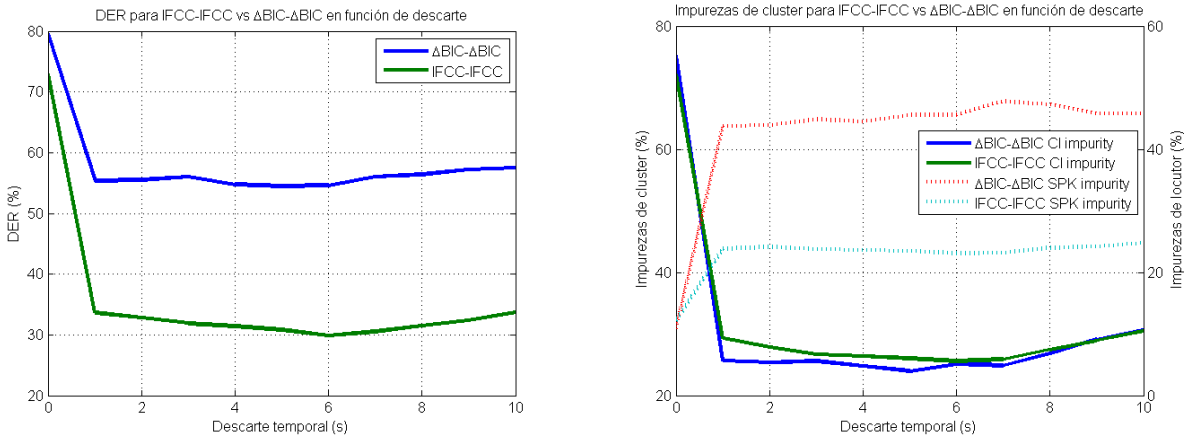
Causantes principales de las altas degradaciones en condiciones ideales, las pruebas en condiciones reales han disminuido su relevancia. No obstante, la técnica de descarte permite trabajar directamente sobre estos segmentos, pudiendo mejorar su influencia en las tasas de error.

Dado que presentan problemas a consecuencia de su pequeña cantidad de información, se



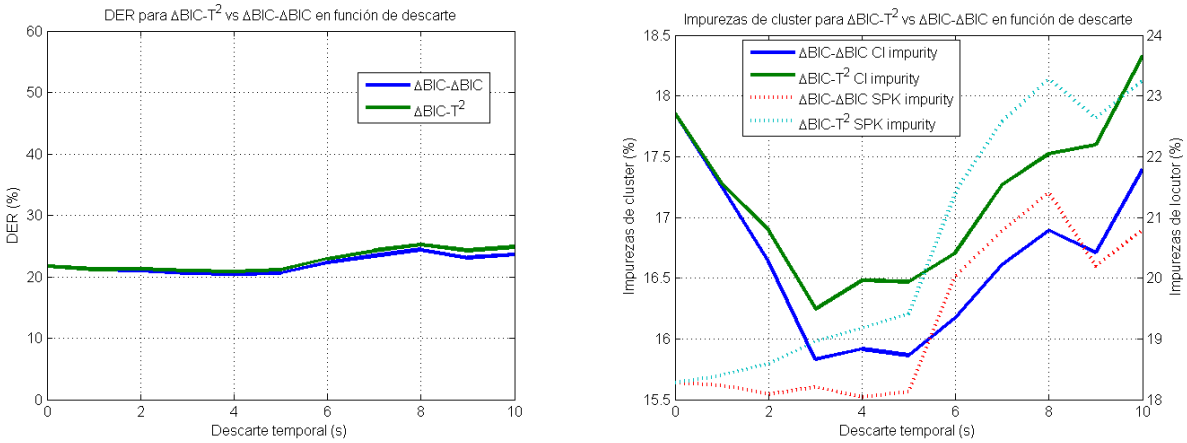
(a) **(b)**

Figura 4.11: DER e Impurezas para IFCC a nivel de sesión (a y b respectivamente), y DER para los segmentos largos (c), para segmentación ideal y estimación de locutores real, en función del descarte



(a) **(b)**

Figura 4.12: DER e Impurezas para IFCC a nivel de sesión (a y b respectivamente), y DER para los segmentos largos (c), para segmentación y estimación de locutores reales, en función del descarte



(a) **(b)**

Figura 4.13: DER para sistema ΔBIC-T2, en función de longitud de descarte

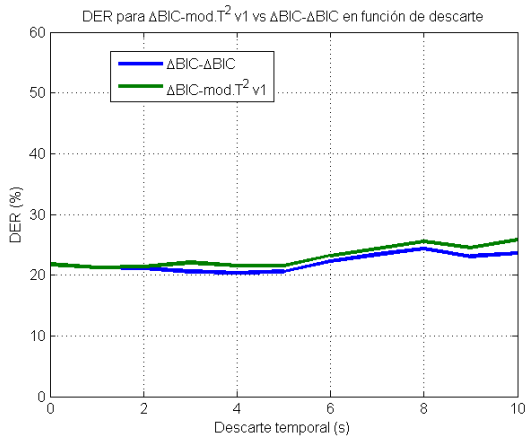
han propuesto diferentes estadísticos menos complejos, más aptos para segmentos con pocos datos:

- **Hotelling T^2**, en adelante T^2. Propuesta en [Zhou and Hansen, 2005]. Aunque específica para segmentos muy cortos (1-2 segundos), se probará también con descartes mayores, debido a los beneficios del descarte. Bajo estas condiciones, los resultados obtenidos pueden verse en la gráfica 4.13

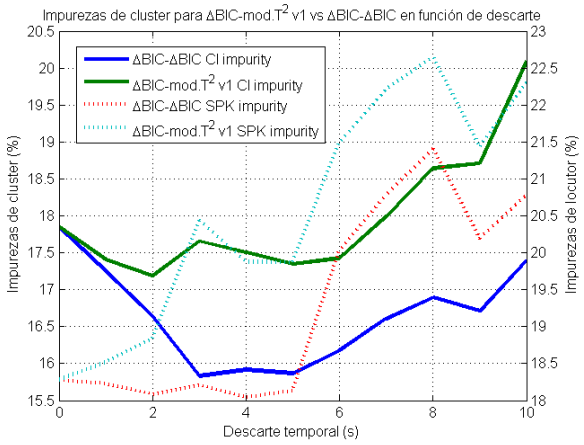
  Observando los resultados se observa que la distancia T^2 no funciona en ningún momento mejor que Δ BIC, presentando resultados muy parejos hasta una longitud de descarte de un segundo, y a partir de esa longitud, se genera una diferencia aproximada de 0.5 %, que en ningún momento se reduce. Los problemas pueden deberse al alto nivel de error para descartes pequeños (vease 4.4 (c)) donde reside el potencial de T^2.

- **Combinación de ΔBIC con T^2**. Si T^2 es la medida más apropiada para segmentos muy cortos (1-2 segundos), no lo es tanto con otros más largos, a los que deberá enfrentarse para descartes mayores. Por esto se emulará a [Huang and Hansen, 2006], donde combina en un mismo proceso T^2 y ΔBIC. No obstante, debido a que el artículo referido es sobre segmentación, combinando dos métricas distintas, queda a nuestra elección la forma de adaptar dicha idea, ya que las métricas no son comparables entre ellas. Se estudiarán tres modificaciones sobre esta idea, realizadas sucesivamente y tratando de resolver los problemas generados por la versión anterior. Los resultados de todas las versiones podrán verse en la figura 4.14.

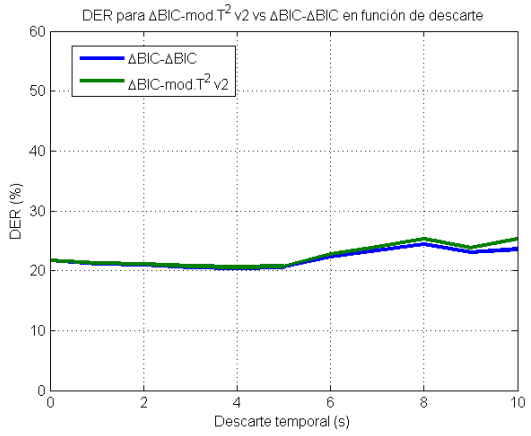
  - La primera modificación o versión consiste en el establecimiento de dos longitudes de segmento, A y B. Todo segmento con longitud inferior a A será aglomerado me-



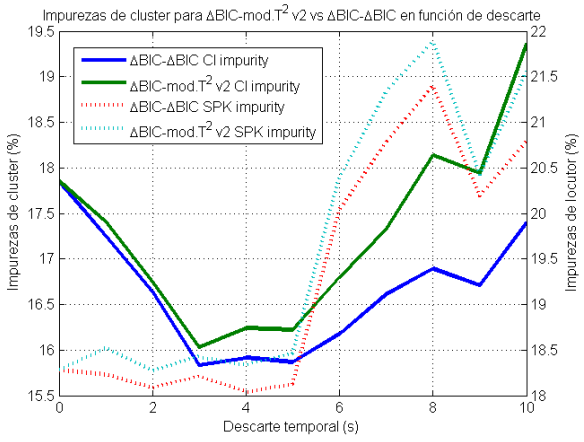
(a)



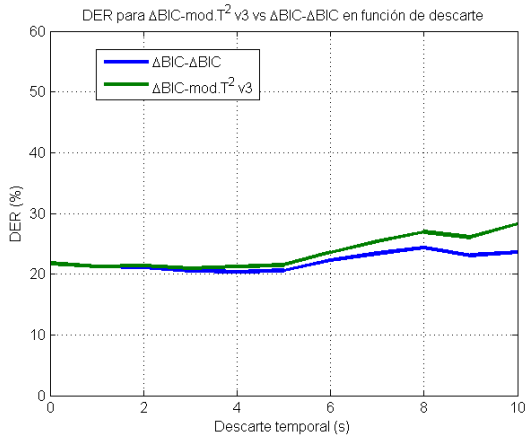
(b)



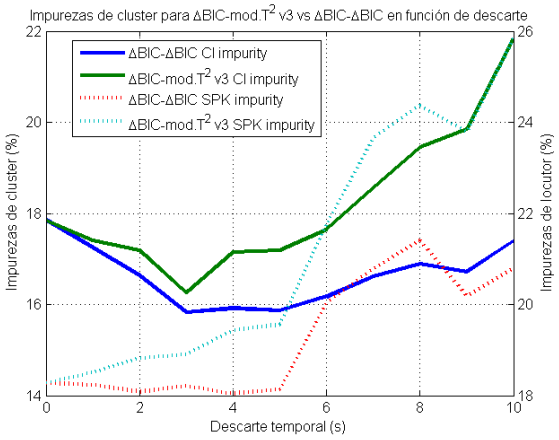
(c)



(d)



(e)



(f)

Figura 4.14: DER e Impurezas para las modificaciones version1 (a y b), version2 (c y d) y version3 (e y f) respecto a ΔBIC

diante T2, mientras todo segmento mayor a B será aglomerado mediante ΔBIC con matriz de covarianza completa. Para aquellos segmentos de longitud comprendida entre A y B serán aglomerados mediante ΔBIC, aunque su matriz de covarianza será diagonal. En todos los casos, tanto la longitud A como la longitud B serán inferiores o iguales a la longitud de descarte de datos. A la hora de jerarquizar fusiones, se ha escogido fusionar primero T2 a los clusters más grandes (procedentes de la etapa de fusión con segmentos largos, o readmitidos largos), imposibilitando cualquier fusión entre estos segmentos tan cortos. Esta posibilidad sí se dotará a todos los segmentos fusionados mediante ΔBIC, en sus dos variantes. Por último, pese a que toda fusión requiere una actualización de los diferentes parámetros de las distancias, se ha optado por obviar esta fase ya que se asume que los segmentos son lo suficientemente cortos como para influir muy poco en ambos valores, y pueden degradar la pureza conseguida en el árbol básico.

- La segunda modificación o versión es una variación del experimento anterior. Se ha observado que el sistema funciona relativamente bien hasta aplicar ΔBIC con covarianza diagonal. Por ello se replicará el experimento anterior, en las mismas condiciones, salvo por el hecho de que solo existirá una distancia, C, eliminando el rango de longitudes donde esta opción se aplicaba. Por ello se diferenciará entre T2 y ΔBIC con covarianza completa.
- La última modificación o versión es una evolución de la anterior. Los resultados indican que para los descartes de segmentos más largos se empeora la respuesta respecto a descartes menores. Se considera que puede deberse al hecho de no actualizar ni distancias ni parámetros de los diferentes clusters. Por este motivo se realizará la segunda modificación, aunque durante la aglomeración mediante ΔBIC sí se realizará la actualización anteriormente comentada.

Viendo las tres evoluciones en común, se puede ver que los resultados, incluso para la mejor evolución (la versión 2) en ningún caso consiguen mejorar a la referencia. Esto puede deberse tanto a los problemas de T^2 respecto a los segmentos cortos, como los problemas de flexibilidad que implica no tener libertad completa de fusión, ya que ambas métricas son incomparables.

Los resultados para tratamiento de segmentos cortos no aportan ningún candidato capaz de mejorar los resultados obtenidos con la referencia ΔBIC, incluso en condiciones ideales donde la segmentación es óptima. Dado que en muchos casos la principal problemática es el nivel de error procedente de la combinación de los segmentos no descartados, el uso de estas técnicas en condiciones no ideales no es viable, ya que en ningún caso se ha logrado reducir sino incrementar dichos valores.

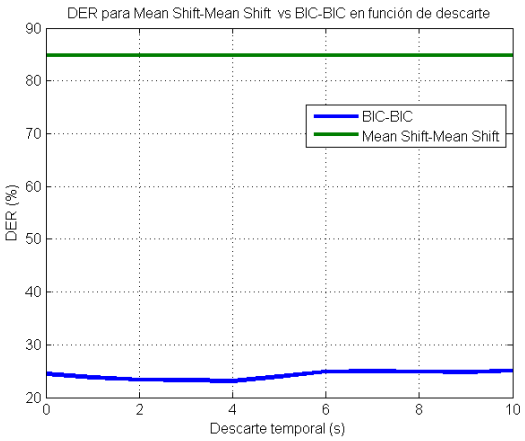


Figura 4.15: Resultados de DER para técnica Mean Shift-Mean Shift respecto a BIC, en función de la longitud de descarte

4.4.3. Criterios de Parada

Hasta el momento, los mejores resultados, incluso en las mejores condiciones es mayor a un 20 %, siendo una tasa de error relativamente alta. Con estas tasas de error, estimar el número de locutores en función de los etiquetados existentes puede no ser la mejor elección, ya que los etiquetados no son excesivamente fiables. Otra opción es inferir dicho valor a partir de las características exclusivamente. La medida a aplicar bajo esta premisa es **Mean Shift**. Los resultados para esta técnica, bajo una segmentación ideal, se muestran en la figura 4.15.

Según los resultados esta técnica no es eficiente, ya que el sistema tiende a detectar un único locutor. Esto se debe a que aproximando cada locutor por una gaussiana , estas gaussianas están muy solapadas, haciendo al sistema incapaz de diferenciarlas por métodos distancias euclídeas. Cabe mencionar que esta técnica fue desarrollada para otras características, los $i-vectors$, desarrollados a partir de MFCC, donde estas gaussianas sí pueden diferenciarse. Por ello, se seguirá optando por un sistema con criterio de parada basado en BIC.

4.5. Quinta fase: Sistema de Diarización completo

A lo largo de los experimentos previos y pensando en sistemas aplicables a la realidad, se ha visto que los mejores resultados son obtenidos con una configuración como la siguiente:

- Segmentación mediante ΔBIC con resegmentación sencilla.
- Clustering con IFCC como métrica de fusión.
- Clustering con BIC como criterio de parada, estimando el número de locutores.

Sin embargo, los resultados mencionados han sido obtenidos sin una etapa de Resegmentación propiamente dicha, para refinar resultados. Como la gran mayoría de sistemas de Dia-

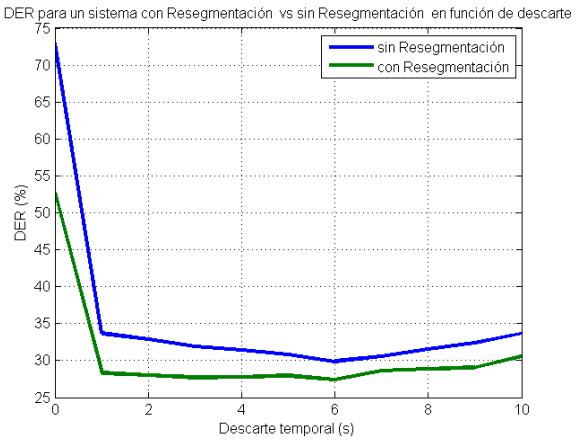


Figura 4.16: Resultados de DER para empleo de Resegmentación respecto a resultados previos, en función de la longitud de descarte

rización publicados lo contienen, sobre la configuración mencionada se aplicará dicha etapa, empleando algoritmos de Viterbi para modelar un HMM (Hidden Markov Model). Los resultados para esta experimentación se muestran en la figura 4.16.

El uso de esta etapa de resegmentación proporciona una mejoría de cierta significancia, siendo más marcada esta mejoría para una situación de descarte nulo, donde el sistema está más contaminado. Esto reafirma la potencia de esta etapa, ya que se trabaja sobre modelos de locutor procedentes de los etiquetados realizados por la etapa de Clustering, están muy contaminados para este trabajo. Como única contrapartida se da el tiempo de procesamiento que estas técnicas exigen, ya que computacionalmente son muy pesadas.

Capítulo 5

Conclusiones y Lineas futuras

Este capítulo será el lugar donde quedarán plasmadas las diferentes conclusiones a extraer de los experimentos realizados. Dado que la labor experimental presenta cinco fases muy marcadas, con objetivos diferentes, es muy útil seguir una estructura común a la ya empleada.

5.1. Primera Fase - Estudio de Coeficientes

El estudio de coeficientes indica que, si bien todos los resultados son muy parejos y con diferencias poco significativas, para estos datos, solamente con los coeficientes menores al c_{13} se obtienen los mejores resultados, perdiendo prestaciones en caso de aumentar la dimensionalidad. En lo referente al coeficiente c_0, el estudio ha comprobado que la suma de log-energías (sentido físico que aporta c_0) no es útil para discriminar locutores, sino que además tiende a confundir al sistema. Esta diferencia de prestaciones no es significativa, con lo que tampoco puede confirmarse o denegarse su utilidad categóricamente. Además, aun bajo segmentación y estimación de locutores ideales, los resultados no son aplicables a un sistema comercial. Conforme se tiende a un sistema real, las degradaciones se hacen presentes, siendo aquella debida a la segmentación la más crítica, y de gran magnitud.

5.2. Segunda Fase - Mejora de la referencia

La segmentación propuesta en [Vaquero, 2011] no es nada precisa para Broadcast. El sistema tiende a segmentar en exceso y perder fronteras. Solo la acción del VAD perfecto alivia los resultados, aunque dista mucho de la ideal. Una técnica de resegmentación es una buena estrategia para aproximar la distribución de segmentos ideal.

Además, en condiciones de segmentación real un criterio de parada como umbral es completamente ineficaz, ya que los errores de segmentación, imprevisibles, pueden modificar las condiciones de test respecto a las de entrenamiento. Métricas que analicen las sesiones al completo y comprueben todas las combinaciones posibles de locutores, tal como se ha probado

con BIC, dan mayor robustez que el umbral, por lo que son más oportunas en condiciones tan desfavorables.

5.3. Tercera Fase - Descarte temporal de segmentos

La estrategia de descarte es acertada, siendo más beneficiosa conforme más reales son las condiciones que el sistema deba afrontar. Conforme los segmentos están más contaminados por una segmentación incorrecta, más viable es fiarse exclusivamente de los segmentos más largos, pues son más robustos frente a estos errores. Aunque en todas las condiciones de experimentación ha aportado mejoras en los resultados, los mayores beneficios han sido obtenidos en condiciones reales, donde nuestra etapa de segmentación es muy poco precisa,convirtiéndose esta estrategia en una técnica para minimizar la pérdida de prestaciones que estos errores generan.

5.4. Cuarta Fase - Técnicas auxiliares al descarte

Dentro de las técnicas para segmentos largos, destaca la modificación de BIC con compensación de correlación (IFCC). Presenta un gran comportamiento con los segmentos más largos, ya que el sistema tiende a priorizar sus fusiones respecto a ΔBIC, por lo que en las condiciones más reales, con segmentación y estimación de locutores real, el sistema se apoya en estos segmentos, más robustos frente a los errores de frontera presentes. El resto de técnicas, tal como han sido probadas, no presentan ninguna mejora, sino que empeoran los resultados.

Para los segmentos cortos, las nuevas métricas tampoco aportan ningún beneficio. Esto se debe a que en condiciones ideales y en su región más fiable, con descartes bajos, los etiquetados realizados con segmentos no descartados son poco fiables, haciendo trabajar sobre un nivel de error muy alto a estas técnicas. Fuera de ese rango, ya no son tan fiables, pese a la reducción de ese nivel de error. Dado que la causa de su no validez es la cantidad de errores por combinar los segmentos no descartados en condiciones ideales, en condiciones más hostiles como las reales estos errores aumentarán en gran medida, haciendo estas filosofías todavía más inviables.

5.5. Quinta Fase - Sistema de diarización completo

A costa de costes computacionales y temporales elevados, la etapa de resegmentación es capaz de reducir la tasa de error DER significativamente. Además, se trata de una técnica bastante robusta, ya que es capaz de aportar mejoras significativas incluso en las peores condiciones de contaminación de modelos (en este trabajo, las mejoras más relevantes han sido con los descartes más bajos, los resultados más contaminados).

5.6. Balance final

Como conclusión final, si bien los resultados no son aplicables actualmente a ninguna aplicación comercializable debido a su baja precisión, los resultados obtenidos están en el orden de los aportados por otros trabajos del estado del arte ([Gupta et al., 2008], [Zelenák et al., 2012] o [Charlet et al., 2013]) en la misma situación, incluso pudiendo admitir aquellos obtenidos en la última fase del proyecto, donde la degradación ha sido muy elevada.

Además, dichos resultados han sido obtenidos por métodos mucho menos costosos a nivel computacional, salvo por la etapa de resegmentación, común a todos ellos, ya que se han empleado tanto técnicas como estadísticas más sencillas que aquellas utilizadas en la bibliografía. Por tanto, se abre la puerta a nuevos sistemas, donde combinando las técnicas de otros trabajos, más complejas, unidas a las ideas aportadas por este proyecto, se puede reducir el término de error a unos valores de funcionamiento válidos, posibilitando por ello los productos comerciales basados en esta tecnología.

5.7. Lineas futuras

Como último punto de este trabajo se marcarán las posibles líneas de trabajo que este proyecto marca para investigaciones futuras.

- Durante todo el trabajo se ha empleado como función de densidad de probabilidad un modelo de gaussiana multidimensional, ya que para otros entornos era suficiente. Para el entorno de Broadcast, no lo es. Vistas las oportunidades que aporta cada técnica, puede comprobarse los beneficios de un aumento de la complejidad mediante el uso de GMM de 16 a 32 gaussianas.

- Este trabajo se ha centrado en el estudio de la etapa de Clustering. A pesar de ser la más compleja a priori, así como de obtener resultados del orden de aquellos expuestos en el estado del arte, el salto al sistema completo provoca una degradación extrema que se deberán combatir. Dado que para este salto se ha tomado una configuración propia de canal telefónico, se deberán probar diferentes técnicas y configuraciones de las mismas con la intención de averiguar cual de ellas es más propicia para Broadcast.

- Un gran hueco dejado por este trabajo es el estudio de JFA. Es una de las técnicas más precisas actualmente, por lo que seria necesaria su experimentación. Los resultados obtenidos para canal telefónico indican su potencia a la hora de etiquetar, por lo que parece razonable probar su eficacia.

- Puede ser interesante desarrollar estimaciones locales de fiabilidad de la tarea de Diarización, a ser posible que no dependan de una referencia. De esta manera, en fase de ejecución normal un sistema podría valorar automáticamente la calidad de sus resultados. Esto permite flexibilizar los sistemas ya que podrían adaptarse a los resultados obtenidos, aplicando sistemas más complejos si fuese necesario, o incluso, refinar mediante una tarea manual.

Apéndice A

Segmentación de audio

En la memoria principal se han descrito en términos generales las diferentes tendencias en lo que respecta a la tarea de segmentación necesaria en el proceso de Diarización. Esta tarea tenía la función de encontrar las fronteras o transiciones entre los diferentes locutores, aislando segmentos donde solo un locutor estuviese presente. Existen tres estrategias diferenciadas:

- **Basados en Métricas**. Se define una medida de distancia de parecido entre dos subregiones contiguas de un mismo fragmento de audio, y se computa para cada muestra de dicha región esta distancia. En la muestra de esa región que indique un máximo de la distancia, el punto más factible de ser frontera, se tomará la decisión de si considerarla como tal, una frontera, o si por contra, a pesar de ser el punto más probable a contener una transición, todo el audio pertenecerá al mismo locutor. Se trata de la filosofía más robusta y empleada, ya que no asume en ningún momento la existencia de datos a priori.

- **Basados en Modelos**. Si se dispone de la suficiente cantidad de datos *a priori*, se pueden generar modelos estadísticos para cada locutor, y se pueden determinar la pertenencia de los datos a los mismos en función de la verosimilitud de éstos respecto a dichos modelos. Aunque conceptualmente válido, en muchos casos no se dispondrá de este conjunto de datos, ya sea por datos insuficientes o por cuestiones de robustez, ya que los modelos dependen de los datos con los que han sido estimados.

- **Basados en Silencio** . Asume la existencia de modelos *a priori* de voz-silencio, de tal manera que se realiza una segmentación voz-silencio, en la que todo silencio se etiqueta como una posible transición.

A.1. Sistemas basados en métricas

Esta clase de sistemas son los más empleados en la gran mayoría de los trabajos actualmente realizados. En gran medida se debe a que estos sistemas no necesitan ningún tipo de

información a priori para realizar su labor, con lo que se conseguirán sistemas más robustos y susceptibles de ser aplicados a un entorno más general, aun a costa de perder eficiencia respecto a los métodos que si requieren este tipo de información. Por lo tanto deben ser incluidos dentro de los sistemas no supervisados.

En estos sistemas se sigue el siguiente principio. Se calculará una distancia entre dos segmentos de audio contiguos con la finalidad de determinar si son lo suficientemente homogéneos como para pertenecer al mismo locutor. Sean dos segmentos de audio contiguos $\mathbf{i}$ y $\mathbf{j}$, cuyas secuencias de características sean los conjuntos $\mathbf{X_i}$ y $\mathbf{X_j}$, de longitudes $\mathbf{N_i}$ y $\mathbf{N_j}$ respectivamente. Así mismo, sea el conjunto $\boldsymbol{X_{ij} = X_i \cup X_j}$ la secuencia de características resultante del cálculo de las mismas en caso de darse la fusión de los segmentos $\mathbf{i}$ y $\mathbf{j}$ anteriormente expuestos. Se formularán dos hipótesis a comparar: $\mathbf{H_0}$ o hipótesis nula, en la cual ambos segmentos pertenecen al mismo locutor, y la hipótesis $\mathbf{H_1}$, según la cual existe una frontera de locutor en el límite entre los segmentos $i - esimo$ y $j - esimo$ respectivamente. Esta filosofía buscará aquella distancia que mejore las prestaciones de este test de hipótesis. Dicha métrica buscará determinar el grado de diferencias entre los sets de características $\mathbf{X_i}$ y $\mathbf{X_j}$ así como también establecer el grado de parecido presente entre ambos, comparándolo con el segmento unión $\mathbf{X_{ij}}$. Estos sistemas comparan esta distancia con un umbral $\boldsymbol{\epsilon}$, para establecer el límite de cada hipótesis tal que

$$D_{ij} \underset{H_0}{\overset{H_1}{\gtrless}} \epsilon \tag{A.1}$$

donde $\mathbf{D_{ij}}$ será la distancia entre los segmentos $\mathbf{i}$ y $\mathbf{j}$, teniendo también en cuenta el segmento unión de ambos.

Vista la definición formal del método, se procederá a una aclaración a nivel funcional. Siguiendo la tendencia de la mayoría de sistemas, los cuales presentan ventanas de tamaño variable, se procederá tal como se muestra en la figura A.1.

El sistema realizará un estudio localizado en una ventana de exploración. Este estudio parte de la premisa de que dicha ventana contiene únicamente una transición o frontera entre fuentes sonoras. Bajo esta premisa se calcularán una serie de distancias en función de la existencia de una hipotética frontera existente en dicha ventana. Sea una ventana $\mathbf{w}$, en la cual para un conjunto de muestras $\mathbf{b}$, hipotéticas fronteras, se han calculado la distancias $\boldsymbol{D(w_i(b), w_j(b)) = D_w(b)}$, siendo $\boldsymbol{D(w_i(b), w_j(b))}$ la distancia existente entre los subsegmentos $\boldsymbol{w_i(b)}$ y $\boldsymbol{w_j(b)}$, obtenidos como los segmentos acústicos existentes en la ventana de análisis a ambos lados de la frontera hipotética $\mathbf{b}$. Basándonos en la suposición de frontera única, solo podrá existir una única frontera $\mathbf{b}$ en la ventana $\mathbf{w}$, por lo que primero se deberá

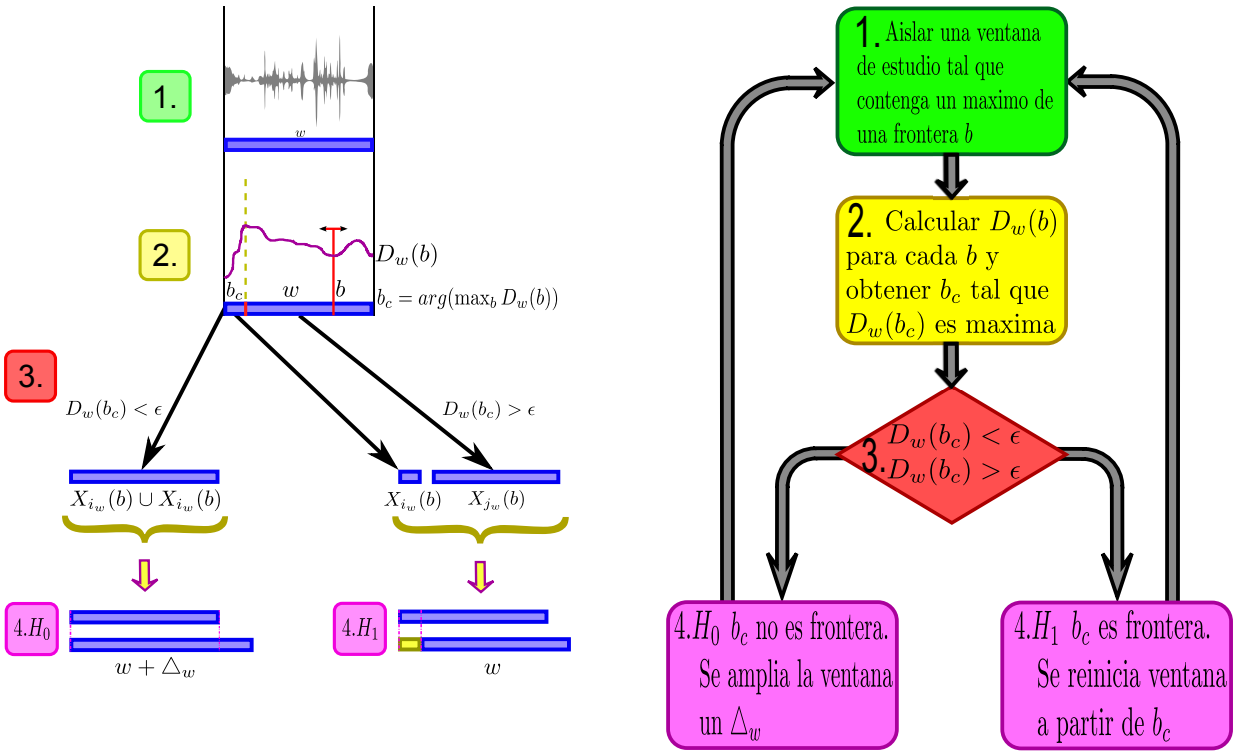


Figura A.1: Esquema general de un sistema de Segmentación mediante distancia

determinar cual es dicha hipotética frontera que tiene más opciones.

$$b_{hipotesis} = arg(\max_b D_w(b)) \tag{A.2}$$

Una vez hallada dicha frontera hipotética, se someterá al clasificador descrito anteriormente en la ecuación A.1, a partir de la cual se establecerá si ese punto, es frontera o no. En caso negativo, es decir, el clasificador opta por la hipótesis nula $\mathbf{H_0}$ y por tanto la ventana no contiene ninguna frontera , se procederá a aumentar dicha ventana un tamaño fijo determinado por $\triangle_{\boldsymbol{w}}$ para repetir el proceso. En caso de que el test hubiese dado positivo, es decir, el clasificador se hubiese decantado por la hipótesis $\mathbf{H_1}$ y en consecuencia existiese una frontera en dicho punto $\mathbf{b_{hipotesis}}$, el sistema habría reiniciado el proceso, inicializando la ventana de análisis a su tamaño inicial y desplazándola hasta hacer coincidir el inicio de la misma con la frontera recientemente encontrada. Este proceso, originalmente fue descrito por [Chen and Gopalakrishnan, 1998], también puede verse por la literatura con ciertas variaciones. Dentro de las técnicas más empleadas para estas tareas de segmentacion se encuentran **BIC** o mejor dicho, su versión diferencia (**ΔBIC**), así como la **divergencia KL2** o **GLR**. Todas estas técnicas son muy comunes a lo largo e la bibliografía, siendo empleadas en cantidad de trabajos. Otras distancias estudiadas pueden ser la distancia de **Mahalanobis** o **Bhattacharyya** en [Hung et al., 2000], donde se compara con la divergencia KL2, así como [Huang and Hansen, 2006], en la que estudia la distancia $\boldsymbol{Hotelling\ T^2}$ combinándola con

BIC, principalmente para segmentación de audio en ventanas pequeñas. No se deben olvidar tampoco algunas modificaciones de las medidas anteriormente expuestas. Aquí podemos citar al BIC cruzado (**Cross-BIC**) en [Anguera, 2005], o una modificación de **BIC con compensación de la correlación entre tramas**, expuesto en [Stafylakis et al., 2013].

La definición formal de las técnicas será llevada a cabo en el anexo C

A.2. Sistemas basados en modelos

Frente a los sistemas basados en métricas, existe otra posibilidad: Los sistemas basados en modelos. Esta clase de sistemas requieren de ciertos datos *a priori*, con los que construir modelos estadísticos, así como establecer umbrales. Siempre que dispongamos de dichos datos, el proceso de segmentación pasa a ser un sencillo problema de decodificación.

En general, estos sistemas generan distintos modelos para clasificar el audio en diferentes clases. Dichos modelos suelen construirse con **GMMs** (Gaussian Mixture Models o modelos de mezcla de gaussianas), y el proceso de clasificación se suele realizar mediante un criterio de máxima verosimilitud (ML o *Maximum Likehood*), como por ejemplo, una **decodificación Viterbi**.

Como caso especial a la segmentación mediante modelos está la Resegmentación. Se trata de un método de refinamiento de resultados, mediante modelos, de los datos obtenidos por una segmentación previa, muchas veces llevada a cabo mediante métricas. Incluso puede realizarse generando los modelos a partir de los datos segmentados en vez de emplear datos *a priori*.

Dentro de los sistemas basados en modelos también se pueden incluir los **modelos basados en silencio**. Esta clase de sistemas se caracteriza por requerir una segmentación voz/no-voz robusta. Sin embargo, esta clase de sistemas no son populares, ya que presentan grandes inconvenientes. Por un lado, solo pueden determinar la existencia de una transición de locutor si existe un paso por un segmento de silencio, una condición nada razonable, y por el otro lado, un segmento de silencio no siempre representa una transición entre locutores.

Apéndice B

Métodos de Clustering

Pese a haber dado en la memoria principal las nociones necesarias acerca de Clustering para comprender el trabajo posterior, existen aclaraciones que complementan lo expuesto anteriormente, y permiten una visión más concreta del trabajo realizado.

El problema del Clustering consiste en la agrupación de los diferentes segmentos sonoros en un conjunto discreto de clases, *a priori* desconocidas, que deberían representar los diferentes hablantes presentes en un audio. Sistema muy vinculado a la diarización, en este ámbito realiza este proceso a partir del audio procedente de la segmentación. Sin embargo, en un contexto más general puede incluso llegar a trabajar con audio procedente de diversas grabaciones, situación típica en ciertos estudios. Esta circunstancia se puede observar en [van Leeuwen, 2010].

En este capítulo se presentará el problema de Clustering, así como algunas de las aproximaciones más populares y empleadas en la literatura relacionada.

El problema más sencillo de agrupamiento queda definido así: Sean N=2 segmentos de voz χ_N (χ_1 y χ_2), y hay K=2 posibles hipótesis de reparto H_K (H_0 y H_1). H_0 es la hipótesis nula, la cual dice que ambos segmentos pertenecen al mismo locutor, mientras la hipótesis H_1 defiende que los segmentos pertenecen a personas distintas.

El problema real de Clustering simplemente es una generalización del problema anterior, donde el conjunto de segmentos Ω, de tamaño **N>2** $\{\chi_1, \chi_2 ... \chi_N\} \in \Omega$ se tiene como entrada, y se quiere agrupar dichos segmentos según su locutor. Además se sabe que el reparto solución correcta es único, y estará incluido en un conjunto M de posibles adjudicaciones hipotéticas $\{H_1, H_2, ...H_M\}$. Este conjunto reflejará la totalidad de combinaciones de Clustering posibles, pasando por la solución H_1, la solución más gruesa según la cual todos los segmentos pertenecen al mismo locutor, hasta la solución H_M, la más fina, según la cual cada segmento sonoro contiene a un locutor, el cual no está presente en ningún otro segmento. Para resolver este problema, se asume la existencia de un modelo generativo Ψ, permitiéndonos obtener un resultado o verosimilitud, para cada hipotético reparto. Cada sección de audio $Audio_k$ se com-

pone de S agrupaciones o clusters no solapadas $\{\boldsymbol{\mathcal{C}_1, \mathcal{C}_2, ...\mathcal{C}_S}\}$, que en conjunto contienen la totalidad de segmentos $\boldsymbol{\Omega}$. Por lo tanto, la verosimilitud del segmento $\boldsymbol{Audio_k}$ será

$$\mathcal{L}(Audio_k) = \prod_{j=1}^{S} \mathcal{L}(\mathcal{C}_j(k)) \propto \prod_{j=1}^{S} P(\mathcal{C}_j(k)|\Omega, \Psi)$$

donde se han asumido tanto el conocimiento de la naturaleza del modelo $\boldsymbol{\Psi}$, sus parámetros, así como que los clusters son independientes y están idénticamente distribuidos, por lo que podemos expresar la verosimilitud de un hipotético segmento $\boldsymbol{Audio_k}$ como el producto de verosimilitudes de los conjuntos no solapados del mismo.

Asumiendo que la verosimilitud obtenida por este método nos permite comparar hipotéticos repartos, el problema se resolvería evaluando las **M hipotéticas soluciones** posibles, escogiendo como correcta aquella con mayor verosimilitud.

Esta solución, **óptima** desde el punto de vista de comprobar todas las opciones, **no es realizable** en la mayoría de los casos, ya que este número de hipótesis crece rápidamente si el número de segmentos a agrupar aumenta.

Con vistas a evitar todo lo posible esta excesiva carga computacional, se han desarrollado diferentes métodos, que si bien no son la solución óptima, si podemos considerarlos **subóptimos** y suficientemente robustos para nuestros intereses. La técnica más popular es la Agrupación Aglomerativa Jerárquica (*Aglomerative Hierarchical Clustering o AHC*). Este sistema reduce la carga computacional realizando elecciones a nivel local.

B.1. Agrupación Aglomerativa Jerárquica o AHC

La agrupación Jerárquica parte de un conjunto de particiones de un audio dado (el reparto más grueso o el más fino), y el sistema iterativamente va fusionando o dividiendo estas agrupaciones hasta llegar al número óptimo de locutores. las divisiones o fusiones realizadas no serán reevaluadas en una iteración siguiente, y se arrastrarán los errores cometidos. Este es el precio a pagar por esa reducción del coste computacional.

Dentro de la agrupación jerárquica, existen dos filosofías de clusterización

- **Bottom-Up**. En esta estrategia, el sistema partirá de la repartición más fina, en la que cada partición únicamente contendrá a un segmento de audio, acotado en sus extremos por dos transiciones de locutor, y por ende, procedente de un único locutor. A partir de esta segmentación, el susbistema irá fusionando particiones iterativamente hasta obtener tantos clusters como locutores estimados. Es la filosofía más empleada, dada la alta sinergia que tiene con el proceso de segmentación anteriormente empleado. Además computacionalmente no tiene tampoco mucha complicación. Solo se requiere una matriz

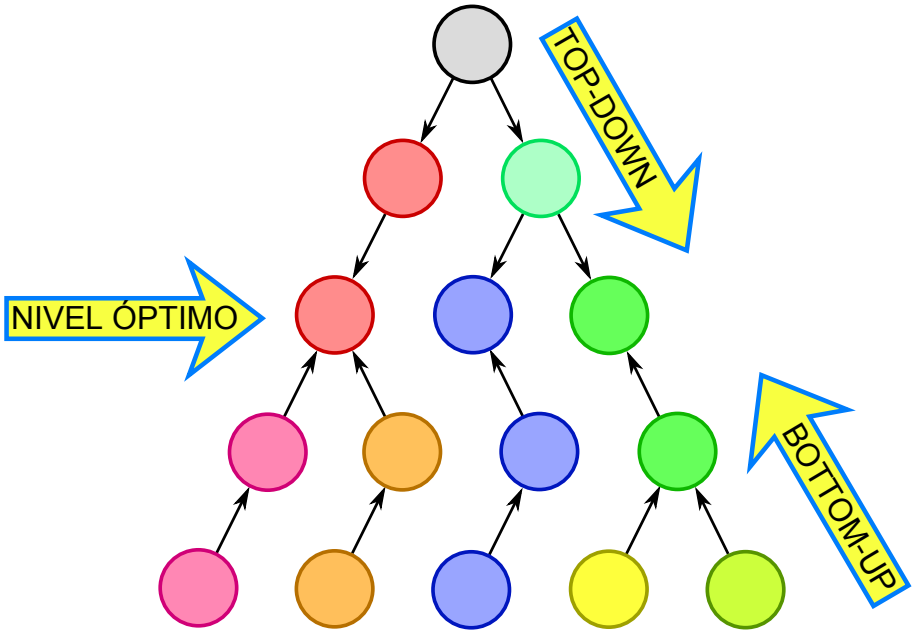


Figura B.1: Esquema de funcionamiento de estrategias de Clustering Bottom-Up y Top-Down

que contenga las distancias entre los distintos clusters. Se escogerán aquellos dos cuya distancia relativa sea menor, para después actualizar la matriz de distancias con el nuevo escenario.

Esta estrategia proporciona ciertas ventajas. Permite ser realizada de una manera iterativa, hasta alcanzar el número de locutores a encontrar. Así mismo, el proceso se puede alargar hasta llegar a la partición más gruesa (solo existe un único locutor), de tal manera que se pueden generar árboles binarios que representen el proceso seguido. En dicho árbol el criterio de parada solo tendría que determinar el nivel en el cual concluir el proceso. La primera vez que se empleó esta técnica en *speaker Clustering* fue en [Jin et al., 1997].

- **Top-Down**. Esta filosofía parte del otro extremo. Se comenzará con un número limitado de particiones (como caso general una única partición), y las particiones existentes serán divididas iterativamente hasta alcanzar el criterio de parada. A diferencia de la estrategia anterior, presenta una gran desventaja, ya que no existe un nivel límite en la generación de los árboles de decisión, puesto que siempre puede desarrollarse una partición más fina descomponiendo un segmento en dos. Esta filosofía es mucho menos común que la estrategia Bottom-Up. Como ejemplo de utilización se puede citar [Reynolds and Torres-Carrasquillo, 2005].

La figura B.1 representa las dos formas más comunes de realizar la Agrupación Jerárquica. Mientras el sistema Bottom-Up solo comprueba cada par de clusters, solo fusionando el par más cercano por ciclo, la topología Top-Down analiza cada cluster y divide uno solo por

iteración. En ambos casos las decisiones tomadas son locales e irreversibles, pues se espera alcanzar un óptimo global, aunque no puede asegurarse este hecho. Todo sistema de Clustering requiere a su vez de dos elementos diferentes: una medida de similitud y un criterio de parada. A continuación se presentarán y describirán las métricas y sistemas de parada existentes en la bibliografía.

- **Medida o distancia de similitud**. En el proceso de Clustering se necesita de algún tipo de medida para decidir qué clusters fusionar (topología Bottom-Up) o dividir (Top-Down). Evaluando este requerimiento, se llega a la conclusión de que todas las medidas anteriormente expuestas para la segmentación, valoradas de la forma adecuada, son válidad para esta tarea. Siguiendo este fundamento, la bibliografía está plagada de casos en los cuales las distintas técnicas anteriormente vistas son empleadas para Clustering: Desde $\triangle BIC$ en [Chen and Gopalakrishnan, 1998], pasando por [Siegler et al., 1997] empleando la **divergencia KL2**, y terminando por medidas no tan comunes como **CLR** (*Cross Likehood Ratio*) en [Reynolds et al., 1998].

- **Criterio de parada**. Debido al carácter iterativo de la aglomeración jerárquica, se necesita un sistema que determine cuándo se ha llegado al número de locutores presentes en el audio. Se trata de uno de los aspectos más críticos de todo el proceso de Clustering, ya que una mala decisión en este aspecto puede degradar en gran medida el resultado, aun a pesar de realizar las uniones correctas. Si bien en algunos sistemas no es tan necesario, debido a que se conoce *a priori* el número de locutores (véase canal telefónico con dos locutores), otros deben contenerlo imperativamente. Buceando por la bibliografía relacionada, se observa una técnica muy común: un **umbral** respecto a la métrica empleada. Esta técnica ha sido empleada en numerosas ocasiones, tales como en [Zhou and Hansen, 2005] o [C.Barras et al., 2004]. En este sentido, y debido al amplio uso de $\triangle BIC$ o **incremento de BIC** (evolución de BIC, explicada junto a BIC en el anexo A), se ha trabajado mucho respecto al tema de los umbrales. En muchos trabajos se ha optado por un umbral tal que $\triangle BIC < 0$ para cada par de clusters. Este umbral simplemente reflejaría que cualquier unión no obtendría ninguna mejora en caso de realizarse. Este sistema se emplea por ejemplo en [Chen and Gopalakrishnan, 1998] Si ya se quieren sistemas más complejos, se puede recurrir a [Nguyen et al., 2008], donde se opta por un criterio basado en $T - test\ de\ student$, o sistemas basados en gradientes como en [Senoussaoui et al., 2013].

B.2. Otras formas de Clustering

Si bien es cierto que la aglomeración jerárquica abunda en la bibliografía, también es verdad que no es la única solución del problema de Clustering. En los últimos tiempos ha surgido

un cierto interés en una técnica denominada **Variational Bayes**. Se trata de una técnica que permite aprender los parámetros y complejidad del modelo empleado en función de los datos de entrenamiento. Algunos ejemplos de esta técnica son [Valente and Wellekens, 2004] y [Kenny et al., 2010].

En caso de conocimiento del número de locutores también se pueden emplear tanto la Cuantificación Vectorial o algoritmos de **K-Means**. Otro frente que se ha abierto recientemente es el de **Joint Factor Analysis** ([Castaldo et al., 2008] o [Vaquero et al., 2010] entre otros), un método de modelado mediante el cual se extraen características compactas de los locutores presentes en pequeños segmentos.

Apéndice C

Métricas de parecido

Este anexo ha sido diseñado para albergar la totalidad de la información acerca de las diferentes técnicas para estudiar la semejanza de segmentos. todas estas técnicas realizarán su labor mediante la formulación de una métrica, la cual deberá depender de dicho parecido. Todas estas técnicas pueden ser empleadas tanto en las tareas de Segmentación como en la tarea de Clustering, ya sea decidiendo qué fusiones de clusters realizar, o incluso aportando para decidir el número de locutores.

- **Bayesian Information Criterion o BIC**. Medida propuesta por [Schwarz, 1978], surgió como medio para estudiar el modelo que mejor representaba unos datos. Dado un conjunto de N muestras independientes $\boldsymbol{\chi}$ procedentes todas de un mismo proceso aleatorio, así como un modelo estadístico $\boldsymbol{\Psi}$, candidato a ser el proceso generador de dichos datos, BIC es una medida que mide el grado de relación entre los datos $\boldsymbol{\chi}$ y el modelo $\boldsymbol{\Psi}$, teniendo el cuenta el grado de adaptación que dicho modelo puede aportar. BIC se define como

$$BIC(\Psi) = log(\mathcal{L}(\chi|\Psi)) - \lambda\frac{1}{2}\#(\Psi)log(N) \tag{C.1}$$

  donde el término $\boldsymbol{log(\mathcal{L}(\chi|\Psi))}$ representa la logverosimilitud de los datos $\boldsymbol{\chi}$ respecto al modelo $\boldsymbol{\Psi}$, representando por tanto el grado de relación entre ambos, mientras el término $\boldsymbol{\lambda\frac{1}{2}\#(\Psi)log(N)}$ es un parámetro de penalización dependiente de la longitud N, un parámetro de ajuste $\boldsymbol{\lambda}$ y el número de parámetros independientes del modelo $\boldsymbol{\Psi}$. Este último término penalizador busca el descarte de aquellos modelos que, a costa de una gran complejidad, pueden dar verosimilitudes o su versión logarítmica muy elevadas, aun no siendo la mejor representación de los datos. Por ello nos permite comparar la fidelidad a los datos de modelos con complejidades (grados de libertad) distintas. El valor del parámetro $\boldsymbol{\lambda}$ está ligado a $\boldsymbol{\epsilon}$, como se comentará posteriormente.

Por lo tanto, BIC es un método muy empleado en los procesos de elección del modelo, ya que nos permite escoger uno lo más potente posible, aunque a la vez en su versión más simple. Simplemente, aquel modelo cuyo BIC sea el mayor, será el modelo más propicio para representar unos datos.

Pese a su gran utilidad, BIC actualmente no satisface nuestros requerimientos, pues no debemos obviar el hecho de que nuestro propósito no es averiguar cual es el modelo que mejor se ajusta a los datos, sino averiguar si es mejor modelar dos segmentos de audio $\mathbf{i}$ y $\mathbf{j}$ por separado o conjuntamente. Para lograr este objetivo tendremos que calcular los valores de dos BIC. Un BIC representará la hipótesis nula o $\mathbf{H_0}$, bajo la cual la secuencia de características $\boldsymbol{\chi_{total}} = \boldsymbol{\chi_i} \cup \boldsymbol{\chi_j}$ puede ser modelada mediante un único modelo $\boldsymbol{\Psi_{total}}$, mientras otro BIC modelará la hipótesis $\mathbf{H_1}$, según la cual existe una frontera $\mathbf{b}$ entre segmentos procedentes de distintos locutores, y por lo tanto, existiendo dos modelos, $\boldsymbol{\Psi_i}$ y $\boldsymbol{\Psi_j}$, uno para cada hablante, contenido supuestamente en el segmento de audio $\mathbf{i}$ y $\mathbf{j}$ respectivamente. En la comparación de ambos BIC tendremos por un lado $\boldsymbol{BIC(H_0)}$, que presentará un modelo más sencillo, dado que busca modelar todos los datos a partir de un único modelo, mientras el $\boldsymbol{BIC(H_1)}$ tratará de obtener una clara mejora en el término de verosimilitud para justificar el incremento en el número de parámetros de los modelos, ya que se habrán empleado dos modelos distintos. La comprobación final se realizará mediante la variación de BIC o $\boldsymbol{\triangle BIC}$:

$$\triangle BIC = BIC(H_0) - BIC(H_1) = R(i,j) - \lambda P \tag{C.2}$$

donde $\boldsymbol{R(i,j)}$ representa la diferencia de log-verosimilitud entre el modelo global y los modelos locales $\boldsymbol{i}$ y $\boldsymbol{j}$, mientras el término $\boldsymbol{P}$ representa la penalización por la mayor complejidad de la hipótesis $\boldsymbol{H_1}$ respecto a $\boldsymbol{H_0}$.

Si bien es cierto que estas definiciones son de carácter general, también es cierto que es muy común emplear como modelos distribuciones gaussianas multidimensionales con covarianza completa, asumiendo por tanto que $\boldsymbol{\chi_k}$ sigue una distribución normal multidimensional de media $\boldsymbol{\mu_k}$ y matriz de covarianza $\boldsymbol{\Sigma_k}$, o expresándolo matemáticamente, $\boldsymbol{\chi_k \sim \mathcal{N}(\mu_k, \Sigma_k)}$. En dicha situación, puede comprobarse que el término $\boldsymbol{R(i,j)}$ puede expresarse tal que:

$$R(i,j) = \frac{N}{2} log(|\Sigma_{i\cup j}|) - \frac{N_i}{2} log(|\Sigma_i|) - \frac{N_j}{2} log(|\Sigma_j|) \tag{C.3}$$

mientras el término de penalización P, presente en la ecuación C.2 y resultante a partir de la diferencia de los términos de penalización individuales de cada BIC, descritos en la

ecuación C.1, quedará bajo dichas condiciones como:

$$P = \frac{1}{2}(p + \frac{1}{2}p(p+1))log(N) \tag{C.4}$$

donde el valor p indica la dimensión del espacio de características.

Además de su empleo con distribuciones gaussianas multidimensionales, también es bastante común en la bibliografía encontrar sistemas basados en BIC empleando como modelos GMM (*Gaussian Mixture Models* o Modelos de Mezclas de Gaussianas). Para estos casos, no existe simplificación de ningún tipo, debiendo ser calculados ambos valores de BIC al uso normal. La función de densidad de probabilidad de un GMM se describe como:

$$f(x) = \sum_{k=1}^{N} w_k \mathcal{N}(\mu_k, \Sigma_k) \tag{C.5}$$

donde w_k es el peso que tendrá cada gaussiana dentro del modelo, siendo la suma de los distintos pesos igual a la unidad ($\sum_{k=1}^{N} w_k = 1$).

Uno de los problemas que presenta esta técnica proviene del parámetro de ajuste $\boldsymbol{\lambda}$, ligado al umbral $\boldsymbol{\epsilon}$ de nuestro clasificador, siendo éste otro parámetro a fijar. En teoría, el ajuste del parámetro $\boldsymbol{\lambda}$ serviría para hacer coincidir el valor del umbral $\boldsymbol{\epsilon}$ con cero, punto de trabajo donde tanto la hipótesis H_0 como la hipótesis $\boldsymbol{H_1}$ representan con la misma fidelidad los datos. Ya existen propuestas para eliminar dicho problema, como en [Ajmera and Wooters, 2003]. En dicho trabajo se ha optado por escoger un modelo $\boldsymbol{\Psi}$ para la hipótesis $\boldsymbol{H_0}$ con el doble de parámetros que los modelos sencillos $\boldsymbol{\Psi_i}$ $\boldsymbol{\Psi_j}$ pertenecientes a la hipótesis $\boldsymbol{H_1}$, de tal manera que se cancele el término de penalización, y por tanto, la dependencia del valor $\boldsymbol{\lambda}$. Esta idea es muy factible de ser implementada en GMMs, de tal manera que el modelo de la hipótesis $\boldsymbol{H_0}$ esté formado por el doble de gaussianas que los modelos de la hipótesis $\boldsymbol{H_1}$

- **Generalized Likehood Ratio o GLR**. Esta medida fue propuesta inicialmente por [Willsky and Jones, 1976], aunque no se usó para estos fines hasta mucho más tarde. Dadas dos secuencias de características $\boldsymbol{\chi_i}$ y $\boldsymbol{\chi_j}$ procedentes de dos segmentos $\boldsymbol{i}$ y $\boldsymbol{j}$, se calcula GLR como un ratio de verosimilitudes , aquella proveniente de las características dadas, siendo calculada bajo la suposición de que ambos segmentos pertenecen al mismo locutor ($\boldsymbol{H_0}$), y la verosimilitud obtenida bajo la suposición de pertenencia a hablantes diferentes ($\boldsymbol{H_1}$). Se trata de una variante de *Likehood Ratio* o LR, ya que mientras en GLR las verosimilitudes se computan a partir de los datos disponibles, en LR se requiere de datos *a priori* para calcular los modelos.

Asumiendo que cada locutor *n-esimo* puede ser modelado mediante una PDF (*Probability Density Function* o Función de Densidad de Probabilidad FDP) cuyos parámetros están englobados en $\boldsymbol{\Psi_k}$, GLR se calcula como sigue:

$$GLR\left(\frac{H_0}{H_1}\right) = \frac{\mathcal{L}(\chi_{i,j}|\Psi_{i,j})}{\mathcal{L}(\chi_i|\Psi_i)\mathcal{L}(\chi_j|\Psi_j)} \tag{C.6}$$

donde $\boldsymbol{\mathcal{L}}$ representa verosimilitud. Como PDFs más empleadas con esta distancia son la distribución gaussiana multidimensional, o una GMM. La distancia también puede verse en su versión logarítmica ($\boldsymbol{D(i,j) = log(GLR)}$)

- **Kullback Leibler Divergence o KL2**. Partiendo de dos distribuciones P y Q, la divergencia KL se define como

$$KL(P,Q) = E_x\left[log\frac{p(x)}{q(x)}\right] \tag{C.7}$$

donde p y q representan las densidades de P y Q respectivamente, mientras el operador $\boldsymbol{E_x}$ denota la esperanza matemática. La divergencia KL mide el extra de bits que requeriría codificar muestras de P empleando un código basado en la distribución Q. Sin embargo, la divergencia KL no es simétrica, por lo que no podemos considerarla estrictamente una norma. Como solución a este efecto se emplea la versión simétrica, denominada KL2, definida como:

$$KL2(P,Q) = KL(P,Q) + KL(Q,P) \tag{C.8}$$

Esta medida fue empleada por primera vez en [Siegler et al., 1997], donde la distancia KL2 fue usada para segmentación acústica, en un entorno de radiodifusión de noticias. Su facilidad de implementación, pudiendo ser adaptada sencillamente a un esquema como el reflejado en la figura A.1, así como su rápida computación la han convertido en una medida muy popular. Generalmente se emplean gaussianas como PDFs, debido a la no existencia de una fórmula cerrada de GMMs, aunque la distancia KL2 puede ser aproximada por una cota superior de cálculo sencillo, la cual sustituirá a la medida real, si las GMM provienen de una adaptación a partir de un UBM (*Universal Background Model*), un modelo GMM de carácter universal a gran escala, tal como se describe en [Do, 2003].

- **Hotelling $\mathbf{T^2}$**. Esta distancia, fue propuesta en [Zhou and Hansen, 2005]. En dicho documento se expone la distancia $\boldsymbol{T^2}$ como solución a la segmentación de segmentos cortos. Dicha distancia se basa en el hecho de que las estadísticas serán más fiables cuanto más datos tengan. Por lo tanto, en segmentos muy cortos, para el empleo de otras distancias

(véase por ejemplo $\triangle \boldsymbol{BIC}$), se deberán calcular los parámetros de los modelos pertinentes para cada segmento de audio, recurriendo por ello en muchos casos a estadísticas de segundo orden. En el caso de segmentos muy pequeños, esta estimación de la estadística de segundo orden es muy poco precisa. Como solución se propuso la distancia $\boldsymbol{T^2}$, que presentaba la siguiente definición:

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)(\sum)^{-1}(\mu_1 - \mu_2) \tag{C.9}$$

Donde $\boldsymbol{\mu_1}$ y $\boldsymbol{\mu_2}$ son las medias de los dos segmentos localizados a ambos lados de la frontera hipotética $\boldsymbol{b}$, N es el total de muestras en el la ventana de búsqueda y $\sum^{\boldsymbol{-1}}$ es la inversa de la matriz de covarianzas de la ventana completa. Este matiz acerca de la covarianza es el que hace a esta distancia una medida interesante para segmentos pequeños, pues no mide la covarianza a cada lado de la hipotética frontera, sino la covarianza total, donde se cuenta con más muestras para ser más robusta. Esta distancia fue propuesta en su artículo original para segmentación, pero tal como se ha comentado anteriormente, muchas de las medidas desarrolladas para segmentación son susceptibles de emplearse en clustering.

- **Combinación de BIC con T2** Esta distancia se expone en [Huang and Hansen, 2006], donde se combina el uso de la distancia Hotelling T^2 con BIC, también con la finalidad de segmentación. En principio puede ser más útil que la distancia T^2 únicamente, ya que esta fue diseñada exclusivamente para segmentos cortos, mientras la combinación propuesta *a priori* puede trabajar tanto con segmentos cortos (T^2) como con los largos (BIC). Sin embargo, la ventaja presentada para segmentación no es aplicable a clustering, ya que requiere comparar las distancias de BIC y $\boldsymbol{T^2}$, con la finalidad de decidir qué clusters fusionar. Esta comparación *a priori* no tiene sentido, por lo que para aplicar este segundo artículo, se debería proceder a la elaboración de algún método para comparar dichas distancias.

- **Compensación de la correlación Inter-frame**. Esta distancia resulta de una modificación de BIC, como se apunta en [Senoussaoui et al., 2013]. BIC en su definición exige que las muestras para calcular las verosimilitudes sean independientes. Esto en realidad no es así, ya que en la generación de las características, se trabaja con ventanas de veinticinco milisegundos , existiendo un desplazamiento de diez milisegundos entre ellas, y por tanto, un solape de quince milisegundos entre ventanas contiguas. Este solape genera una correlación entre los datos y por tanto, no existe la independencia que BIC exige. Este error ha sido cometido continuamente en la bibliografía, y aun a pesar de existir, los distintos resultados positivos obtenidos lo han hecho parecer un error de poca importancia. Sin embargo, dicho error existe y se mantiene ahí. Esta estrategia está centrada en

eliminar o al menos minimizar dicha fuente de error. Para ello se postula una variación de BIC en caso de distribuciones gaussianas multidimensionales como sigue:

$$\begin{aligned} \triangle BIC &= \tfrac{rN}{2} log(|\Sigma_{i,j}|) - \tfrac{rN_i}{2} log(|\Sigma_i|) - \\ &- \tfrac{rN_j}{2} log(|\Sigma_j|) - \lambda \tfrac{1}{2}(p + \tfrac{1}{2}p(p+1)) log(rN) \end{aligned} \quad . \tag{C.10}$$

Se puede observar que la fórmula es casi igual a la descrita para BIC en caso de distribuciones gaussianas. La única diferencia es un parámetro $\boldsymbol{r}$, situado junto a los diferentes valores de longitudes de segmento. Este valor trata de reflejar una pérdida de cantidad de información independiente, debido al solape anteriormente expuesto, por lo que su valor será siempre inferior a la unidad. En todo caso, sus creadores han aproximado su valor a $\mathbf{0,3}$ en [Senoussaoui et al., 2013]. En este trabajo, en el empleo de esta técnica se propondrá ajustar los valores tanto de $\boldsymbol{\lambda}$ como de $\boldsymbol{r}$ que optimicen la respuesta de nuestro sistema.

- **Distancia T-Student**. Esta distancia, propuesta por [Nguyen et al., 2008], trata de realizar la tarea de clustering de una manera diferente a las técnicas convencionales. Estas técnicas, BIC, GLR, CLR, NCLR, etc. , trabajan bajo la asunción de que si $\boldsymbol{\lambda_1}$ y $\boldsymbol{\lambda_2}$ son modelos del mismo locutor, el valor de la verosimilitud $\boldsymbol{\mathcal{L}(X|\lambda_1)}$ y $\boldsymbol{\mathcal{L}(X|\lambda_2)}$ serán próximos, donde $\boldsymbol{X}$ refleja la totalidad de N observaciones $\boldsymbol{X = x_1, x_2, ..., x_N}$. A diferencia de lo anterior, esta métrica propone una modificación: Si $\boldsymbol{\lambda_1}$ y $\boldsymbol{\lambda_2}$ son modelos de un mismo locutor, la población de valores de verosimilitud para los datos respecto al modelo $\boldsymbol{\lambda_1}$ ($\boldsymbol{\mathcal{L}(x_i|\lambda_1), \forall x_i \in X}$) estará próxima a los valores para el modelo $\boldsymbol{\lambda_2}$ ($\boldsymbol{\mathcal{L}(x_i|\lambda_2), \forall x_i \in X}$). Para ello se recurre al T-test de Student, reflejado en la ecuación

$$T_d = d(S_f(X), S_g(X)) = \frac{|m_1 - m_2|}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}} \tag{C.11}$$

donde $\boldsymbol{m_1}$, $\boldsymbol{\sigma_1}$, $\boldsymbol{n_1}$, $\boldsymbol{m_2}$, $\boldsymbol{\sigma_2}$, $\boldsymbol{n_2}$ son respectivamente las medias, desviaciones estándar y tamaño de las distribuciones $\boldsymbol{S_f(X)}$ y $\boldsymbol{S_g(X)}$. Dichas distribuciones así mismo se definen como:

$$S_f(X) = \{f(x_i)|x_i \in X\} \tag{C.12}$$

$$S_g(X) = \{g(x_i)|x_i \in X\} \tag{C.13}$$

pudiendose definir las distribuciones $\boldsymbol{f(x)}$ y $\boldsymbol{g(x)}$ como:

$$f(x) = log\mathcal{L}(x|\lambda_{C1}) - log\mathcal{L}(x|\lambda_{UBM}) \quad \text{(C.14)}$$

$$g(x) = log\mathcal{L}(x|\lambda_{C2}) - log\mathcal{L}(x|\lambda_{UBM}) \quad \text{(C.15)}$$

siendo $\boldsymbol{X = \{x_1, x_2, ..., x_N, y1, y2, ..., y_M\}}$, con $\boldsymbol{x_i}$ e $\boldsymbol{y_i}$ las observaciones dependientes de los locutores $\boldsymbol{C1}$ y $\boldsymbol{C2}$ respectivamente, $\boldsymbol{\lambda_{Ci}}$ es el modelo estimado $\boldsymbol{i}$ a partir de los datos del locutor $\boldsymbol{C_i}$, y $\boldsymbol{\lambda_{UBM}}$ es un modelo universal.

Por tanto, como única decisión a tomar será la distribución del modelo $\boldsymbol{\lambda_{Ci}}$. En lo referente a este proyecto, se ha optado por una distribución gaussiana multidimensional, ahorrando en el aspecto computacional, puesto que esta medida presenta coste de calculo elevado.

Apéndice D

Métodos empleados como criterio de parada

Durante el trabajo se emplearán diversos criterios de parada, los cuales deberán establecer el número de locutores activos en el audio de estudio. A continuación se prensentan aquellos métodos que han sido aplicados durante el desarrollo del proyecto:

- **Oráculo**. Es el primer criterio de parada a emplear. Se basa en el conocimiento exacto y *a priori* del número de locutores. No tiene un sentido en un ámbito real, pero será ámpliamente empleado en ciertas fases del proyecto, por lo que es necesario citarlo. Mediante este método se conseguirá establecer exactamente el número de locutores presentes en las sesiones, no introduciendo por ello un error extra.

- **Umbral sobre las metricas empleadas para fusión**. Es la primera técnica aplicable en un entorno real. Ha sido muy empleada en la bibliografía. La idea principal sobre la que este sistema se sustenta es: Las métricas empleadas para fusión muestran la diferencia de parecido entre la hipótesis de ser dos locutores (H_1) frente a ser un único locutor (H_0). El sistema irá fusionando cada vez los clusters más parecidos de entre los posibles, aunque cada vez dicho parecido será menor. Llegará un punto en el cual esta diferencia será grande ya que se habrá alcanzado el criterio de parada y obligatoriamente se están fusionando locutores diferentes. En teoría, este valor tendría que ajustarse a cero, reflejando que no hay ninguna mejora, modificando para ello los diferentes parámetros de ajuste de las métricas. También puede darse el caso de que los valores de ajuste estén adaptados a otros valores, por razones no relevantes al caso. En ese caso, en vez de ajustar los parámetros para anular el umbral, se puede ajustar el umbral para alcanzar los valores de los parámetros de ajuste. Este ajuste se logrará por tanto por prueba y error. Se trata de una medida muy dependiente de los datos de entrenamiento, y a veces dependiente de los parámetros de los modelos de fusión empleados. Esto la hace poco robusta. A su favor tiene ser una medida muy sencilla, por lo que ha sido ámpliamente utilizada en el caso

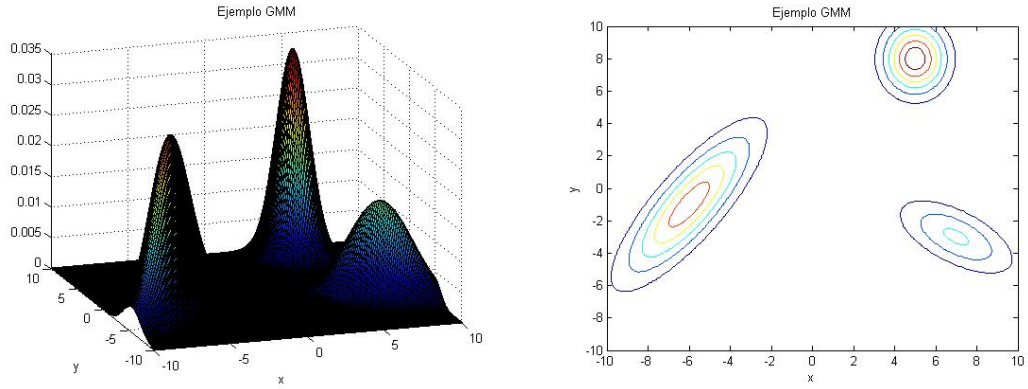


Figura D.1: Histograma (a) y Mapa de curvas de nivel (b) para un GMM en un espacio vectorial de dimensión dos

de no querer profundizar más.

- **Mean-Shift**. Esta técnica, propuesta en [Fukunaga and Hostetler, 1975], y evolucionada en [Senoussaoui et al., 2013], es una de las técnicas más evolucionadas que se van a probar dentro del campo de criterios de parada, buscando inferir el número de locutores. Considerando todas las características obtenidas del audio original, se puede considerar el audio como un proceso estocástico regulado por una variable aleatoria cuya función de densidad de probabilidad (FDP) puede modelarse como un modelo de mezcla de gaussianas (GMM). Dado que el audio en principio contiene varios locutores, cada locutor está modelado por su propia función de densidad de probabilidad , pudiendo ser esta modelada por un GMM también. Un ejemplo de esta FDP queda reflejado en la figura D.1, donde se muestra un GMM perfecto, siendo la dimensión de las características dos. Suponiendo que cada locutor fuese modelado únicamente por una única gaussiana, en el ejemplo anterior habría únicamente tres locutores, llegando a esta conclusión ya fuese por métodos visuales (una persona por métodos manuales) o contando los máximos (un ordenador automáticamente). A estos máximos el ordenador llegaría partiendo de semillas aleatorias, las cuales se aproximan a estos máximos siguiendo las direcciones de mayor inclinación mediante un proceso iterativo. Esta inclinación será calculada mediante el algoritmo de Mean-Shift. Huelga decir que varias semillas pueden converger en un único máximo, por lo que solo deberemos contabilizar una.

  Método nativo y desarrollado principalmente para Factor Analysis, se puede asumir que las observaciones (*speaker factors*) para un locutor tienen una distribución estadística de tipo normal. También se probará sobre MFCCs, asumiendo en la misma premisa. Esta aproximación es más gruesa que la anterior, y por lo tanto, repercutirá en los resultados,

aunque esta técnica presenta como requisito que los datos sean modelados mediante una FDP con un único máximo local para cada locutor.

No obstante, el problema anteriormente presentado era demasiado sencillo, aunque ejemplificaba la idea a desarrollar. En un caso real, la dimensión de la FDP será mayor que dos (entre once y diecinueve, según los MFCC que empleemos). Además, el ejemplo presentado es un ejemplo de FDP perfecto, partiendo de las fórmulas. En un entorno real, existirán máximos próximos entre sí, en función de como queden situadas las características en el espacio. De todos esos máximos, algunos pueden deberse a dos locutores muy parecidos, mientras otros se deben simplemente al efecto del ruido, o que los puntos no están uniformemente distribuidos, aumentando el número de locutores respecto al valor correcto. El sistema deberá ser configurado para que sea capaz de discernir entre ambos casos lo más fiablemente posible

Esta distancia presenta dos parámetros de ajuste a estudiar: El cálculo del gradiente se realiza en un entorno del punto, por lo que debe acotarse la ventana de estudio, de amplitud $\boldsymbol{w}$. Además, debe existir una ventana de fusión. Esta ventana se crea con la intención de minimizar el problema de los máximos ruidosos. Todos los máximos que queden incluidos en dicha ventana de fusión serán contabilizados únicamente como un único máximo, y por ende, representarán a un único locutor. La distancia $\boldsymbol{w_p}$ representa la amplitud de dicha ventana. Se probarán también dos tipos de distancias, tanto para el cálculo de las ventanas de análisis como de fusión: La distancia euclídea [Fukunaga and Hostetler, 1975] como referencia, y la distancia coseno, propuesta en [Senoussaoui et al., 2013] como evolución.

- **T-student**. También propuesta en [Nguyen et al., 2008], combina un sistema de parada basado en la distancia *T-student* con una distancia *T-test*. Este sistema genera a partir de cada cluster una serie de subclusters, calcula la distancia *T-test* entre ellos y en función de dichas distancias, calcula las distancias *inter-cluster* e *intra-cluster*, definidas a continuación. Sea $\boldsymbol{C_i}$ una manera de clusterizar los datos X en $\boldsymbol{K_i}$ clusters ($\boldsymbol{C_i = \{C_1^{(i)}, C_2^{(i)}, ..., C_{K_i}^{(i)}\}}$). Denominando $\boldsymbol{d(x_m, x_n)}$ la distancia anteriormente descrita en la sección *T-test* entre los subsegmentos $\boldsymbol{x_m}$ y $\boldsymbol{x_n}$, se define la distancia de cluster $\boldsymbol{D(C_i, C_j)}$ como:

$$D(C_i, C_j) = \{d(x_m, x_n) | x_m \in C_i, x_n \in C_j \forall m, n\} \tag{D.1}$$

$$D_{intra} = \bigcup_{i=1}^{K} D(C_i, C_i) \tag{D.2}$$

$$D_{inter} = \bigcup_{1 \leq i < j \leq K} D(C_i, C_j) \tag{D.3}$$

siendo $\boldsymbol{D_{intra}}$ la población de distancias intra-cluster y $\boldsymbol{D_{inter}}$ la población de distancias inter-cluster.

En función de estas distancias intra e inter cluster, se calculará una distancia, la cual será la que regule si se ha alcanzado el punto óptimo o no.

Este sistema, ya que está basado en la distancia *T-test de Student*, presenta también la misma dependencia de la distribución de los datos como en el caso de métrica. Además, se deberá establecer la longitud de los diferentes subsegmentos. A diferencia de otros métodos, esta técnica no requiere el ajuste de un umbral.

- **BIC** Tal como indica su nombre, esta técnica está englobada dentro de los denominados criterios de información, una forma de valorar la validez de un modelo estadístico con respecto a unos datos. Para la funcionalidad de estimación de locutores también es válido, ya que , teniendo distintos etiquetados con diferente número de locutores, nos indica cuál es el que mejor modela los datos. La definición formal aplicada, obviando la definición propia de BIC, expuesta en el anexo C es: Sea $X = x_1..X_N$ un set de datos, y existan C_j posibles aglomeraciones o etiquetados, conteniendo un número de locutores diferente. con $j = 1..K$, pudiendo modelarse cada aglomeración mediante un modelo Ψ_j. Además, se aceptará por simplicidad que la aglomeración C_j con modelo Ψ_j refleja un valor de j locutores activos. Bajo esta premisa, el número de locutores N_{loc} presente en el conjunto de datos X será aquel que para dicho set maximice:

$$N_{loc} = arg\max_{j} BIC(\Psi_j) \tag{D.4}$$

Apéndice E

Métodos de evaluación

En este anexo se van a estudiar las distintas métricas desarrolladas a lo largo del tiempo con el fin de **evaluar** la calidad de estos sistemas.
En los inicios proliferaron algunas métricas, tales como los ratios de pérdida y falsa alarma respecto al número de cambios de locutor. Estas medidas estudiaban la proporción de transiciones de locutor que el sistema no detectaba (*miss*) o generaba artificialmente (*false alarm*). Sin embargo, estas medidas, no mostraban la precisión de dichas fronteras. Es decir, no solo importa localizar la existencia de una frontera, sino también saber localizarla en el tiempo lo más precisamente posible. Este fallo ha sido arrastrado durante cierto tiempo, estando presente en algunos de los artículos más representativos de la Diarización, como pueda ser [Chen and Gopalakrishnan, 1998]. Otro problema procede del hecho de que todas las fronteras no son igual de importantes. Es mucho peor no encontrar una frontera entre dos segmentos de audio grandes pertenecientes a dos locutores, a olvidar la frontera que separa segmentos pequeños, en cuyo caso no será tan crítico. Teniendo en cuenta ambas problemáticas, surgió una solución, propuesta en [J. L. Gauvain et al., 1999]: Medir el porcentaje del tiempo incorrectamente clasificado.
Como evolución de la idea de Gauvain, surgió una medida, la más utilizada en la actualidad. Se trata del DER (**Diarization Error Rate**). Dada una hipótesis de etiquetado procedente de un sistema de diarización, DER se define como el tiempo total incorrectamente asignado, dividido entre el tiempo total a estudiar en el audio de entrada. Si bien el término DER es una medida global de error, debido a su definición puede descomponerse en cuatro términos, cada una referente a una fuente de error diferente, de tal manera que el $\boldsymbol{DER_{TOTAL}}$ será la suma de los diferentes términos de error.

- **Voz no localizada** (*Missed Speech*). Se define como voz no localizada a todos aquellos segmentos sonoros, que conteniendo voz, no han sido etiquetadas como tal. Este error procede de los sistemas voz/no-voz dependientes de la segmentación de audio. Formal-

mente puede definirse como:

$$E_M = \frac{T_{voz}(no - detectada)}{T_{voz}} \tag{E.1}$$

Donde $\boldsymbol{T_{voz}(no - detectada)}$ es el tiempo de audio incorrectamente etiquetado como sin voz.

- **Falsa alarma de voz** (*False Alarm Speech*). Se define como falsa alarma de voz todos aquellos segmentos/muestras de audio, etiquetados como voz, en realidad no la contienen. Como su contrapunto Voz no localizada, este error también procede de las técnicas de segmentación de audio. Se define como:

$$E_{FA} = \frac{T_{no-voz}(voz)}{T_{voz}} \tag{E.2}$$

  Si definimos al tiempo de no-voz etiquetado como voz bajo la variable $\boldsymbol{T_{no-voz}(voz)}$

- **Error de locutor** (*Speaker Error*). Este error engloba la fracción total de tiempo que contiene voz y ha sido etiquetada a un locutor incorrecto. No tiene en cuenta errores de solapes. Se define como:

$$E_{SPK} = \frac{T_{voz}(loc_i \neq loc_j)}{T_{voz}} \tag{E.3}$$

  Donde $\boldsymbol{T_{voz}(loc_i \neq loc_j)}$ representa el tiempo de audio erróneamente atribuido al locutor $\boldsymbol{i}$, cuando debería haber sido asignado al locutor $\boldsymbol{j}$.

- **Error de solape de discurso** (*Overlapped Speech Error*). Este error engloba a todos aquellos segmentos sonoros, en los que simultáneamente diferentes locutores hablan, superponiéndose sus voces, siendo indetectados por el sistema, no asignando dichos segmentos a todos los locutores activos. Aunque en teoría este término exista, en algunas aplicaciones no se computa, dado que suele incluirse su valor en alguno de los términos de error anteriormente descritos. Aun así puede expresarse así:

$$E_{OV} = \frac{T_{audio-superpuesto}}{T_{voz}} \tag{E.4}$$

Una vez vistos todos los términos, el error DER se expresa así:

$$DER = \frac{T_{mal-etiquetado}}{T_{voz}} = E_M + E_{FA} + E_{SPK} + E_{OV} \tag{E.5}$$

Dada la definición formal del DER, deben tenerse en cuenta ciertos aspectos importantes. El problema del solape es un problema real aunque muy lejano a ser resuelto, por lo que a la hora de analizar la precisión de estos sistemas, el término de solapamiento tiende a obviarse. En esta situación, la medida DER puede separarse en dos términos distintos. Por un lado tenemos

$\boldsymbol{E_M + E_{FA}}$, una medida de la calidad del proceso de Segmentación de Audio, realizado para el detector de actividad vocal o VAD. En el otro extremo tenemos el término $\boldsymbol{E_{SPK}}$ que refleja el error de diarización. Para nuestro propósito nos centraremos principalmente en este último término.

Pese a la utilidad manifiesta de medida de error que nos aporta el DER, solo nos presenta una medida de audio mal etiquetado, sin aportar nada acerca de su **distribución**. Como consecuencia, surgieron distintas medidas de pureza, generalmente aplicadas sobre locutores y clusters. También pueden verse estos conceptos a partir de sus contrapuestos (impurezas) en [van Leeuwen, 2010]

Sea un conjunto de N segmentos Ω, que contiene R locutores distintos, y R<N, se define como frecuencia relativa del locutor $\boldsymbol{r}$ en el segmento $\boldsymbol{n}$ como

$$f_r(n) = \frac{L_r(n)}{L(n)} \tag{E.6}$$

donde $\boldsymbol{L_r(n)}$ es el número de observaciones o tramas en el segmento $\boldsymbol{n}$ que pertenecen al locutor $\boldsymbol{r}$, y $\boldsymbol{L(n)}$ $\sum_{\boldsymbol{r=1}}^{\boldsymbol{R}} \boldsymbol{L_r(n)}$ es el total de observaciones en el segmento $\boldsymbol{n}$. En el caso de que un segmento contenga únicamente a un único locutor $\boldsymbol{i}$, $\boldsymbol{f_r(n) = 1}$ si $\boldsymbol{r = i}$ o $\boldsymbol{f_r(n) = 0}$ en cualquier otro caso. Para un hipotético reparto H que presente S clusters $\boldsymbol{\mathcal{C}_s}$, $\boldsymbol{s = 1, ..., S}$, se puede definir la frecuencia de un locutor $\boldsymbol{r}$ en un cluster $\boldsymbol{\mathcal{C}_s}$ como:

$$f_r(\mathcal{C}_s) = \frac{L_r(\mathcal{C}_s)}{L(\mathcal{C}_s)} = \frac{\sum_{n\in\mathcal{C}_s} f_r(n)L(n)}{\sum_{n\in\mathcal{C}_s} L(n)} \tag{E.7}$$

Donde $\boldsymbol{L_r(\mathcal{C}_s)}$ es el número de tramas de todos los segmentos en el cluster $\boldsymbol{\mathcal{C}_s}$ que pertenecen al locutor $\boldsymbol{r}$, y $\boldsymbol{L(\mathcal{C}_s)}$ es el número total de tramas en el cluster $\boldsymbol{\mathcal{C}_s}$. Como nota adicional la frecuencia del locutor $\boldsymbol{r}$ en el cluster $\boldsymbol{\mathcal{C}_s}$ puede obtenerse como la media ponderada de la frecuencia relativa del locutor $\boldsymbol{r}$ para todos los segmentos $\boldsymbol{n}$ en $\boldsymbol{\mathcal{C}_s}$, donde los pesos vienen dados por el total de tramas. En el caso de que los segmentos sean pesados por igual, $\boldsymbol{f_r(\mathcal{C}_s)}$ se reduce a la media de $\boldsymbol{f_r(n)}$ para cada $\boldsymbol{n \in \mathcal{C}_s}$.

A partir de la definición de $\boldsymbol{f_r(\mathcal{C}_s)}$, sean cuales sean los pesos escogidos, se puede definir la pureza de cluster para una única agrupación $\boldsymbol{\mathcal{C}_s}$ como la frecuencia del locutor r que obtiene la máxima frecuencia en dicho cluster $\boldsymbol{\mathcal{C}_s}$. Matemáticamente:

$$P_{cluster}(\mathcal{C}_s) = \max_r(f_r(\mathcal{C}_s)) \tag{E.8}$$

y por ende, la pureza del conjunto de todos los segmentos $\boldsymbol{\Omega}$ dada el hipotético reparto H, se define como la media ponderada de las purezas individuales de cada cluster.

$$P_{cluster}(\Omega|H) = \frac{\sum_{s=1}^{S} L(\mathcal{C}_s) P_{cluster}(\mathcal{C}_s)}{L(\Omega)} \tag{E.9}$$

donde $\boldsymbol{L(\Omega)}$ es el total de muestras de voz en todos los segmentos.

Se desprende por tanto que la pureza de cluster solo refleja para cada cluster que un locutor domina dicha agrupación, mas no refleja la posibilidad de que un locutor pueda estar repartido entre muchos clusters. Por ello surgió la pureza de locutor, la cual puede definirse de un modo análogo a la pureza de cluster. En primer lugar se calculará la frecuencia de un segmento $\boldsymbol{n}$ en un locutor $\boldsymbol{r}$ como:

$$g_n(r) = \frac{L_r(n)}{L_r(\Omega)} \tag{E.10}$$

donde $\boldsymbol{L_r(\Omega)}$ es el número de tramas que pertenecen al locutor $\boldsymbol{r}$ en la totalidad de segmentos sonoros. A partir de esta definición se puede obtener la frecuenciade un cluster $\boldsymbol{\mathcal{C}_s}$ en un locutor $\boldsymbol{r}$, pudiendo expresarse como:

$$g_{\mathcal{C}_k}(r) = \frac{L_r(\mathcal{C}_s)}{L_r(\Omega)} = \sum_{n \in \mathcal{C}_s} g_r(n) \tag{E.11}$$

Esta frecuencia aumentará si los segmentos que contienen un locutor $\boldsymbol{r}$ son fusionados en un único cluster, hasta el punto donde $\boldsymbol{c_k(r) = 1}$, si todos los segmento de dicho locutor pertenecen a ese cluster. Siguiendo con la analogía, la pureza de locutor para un locutor $\boldsymbol{r}$ puede obtenerse como la máxima frecuencia de locutor para cada uno de los clusters existentes. Matemáticamente puede expresarse como:

$$P_{speaker}(r) = \max_{\mathcal{C}_s}(g_{\mathcal{C}_s}(r)) \tag{E.12}$$

y para finalizar, se puede definir la pureza de locutor para la totalidad del segmento $\boldsymbol{\Omega}$, en función de un reparto hipotético H, mediante la media ponderada de las purezas de locutor para cada hablante:

$$P_{speaker}(\Omega|H) = \frac{\sum_{r=1}^{R} L_r(\Omega) P_{speaker}(r)}{L(\Omega)} \tag{E.13}$$

Al igual que su homóloga para el cluster, esta medida tampoco es capaz de aportar la totalidad de la información por sí misma. Sin embargo, el empleo conjunto de ambas si puede ser considerado una opción muy válida, ya que son dos medidas complementarias.

En la literatura es común emplear estos conceptos, pero mediante su contrapartida, debido en parte a la costumbre de evaluar mediante el error de funcionamiento, las impurezas de cluster y locutor. Éstas se definen así:

$$I_{cluster}(\Omega|H) = 1 - P_{cluster}(\Omega|H) \tag{E.14}$$

$$I_{locutor}(\Omega|H) = 1 - P_{locutor}(\Omega|H) \tag{E.15}$$

En nuestro proceso de diseño y optimización se buscará reducir en la manera de lo posible ambos valores, tanto $\boldsymbol{I_{cluster}}$ como $\boldsymbol{I_{locutor}}$. El problema radica en que, si uno baja, el otro término tiende a subir. Sin embargo, en función de la aplicación final, puede no ser tan importante reducir uno de los términos, con lo que puede tratarse de optimizar el otro. Si se diese el caso de considerar ambos términos con igual importancia, se define un concepto análogo al EER (*Equal Error Rate*) de tareas de detección. Se denomina $\boldsymbol{EI(\Omega)}$ **(Equal Impurity)**, y se define como el punto de operación en el cual un reparto $\boldsymbol{H_{EI}}$ genera el mismo valor de impurezas tanto para cluster como para locutor. Matemáticamente se puede expresar como:

$$EI(\Omega) = I_{cluster}(\Omega|H_{EI}) = I_{locutor}(\Omega|H_{EI}) \tag{E.16}$$

Bibliografía

[Ajmera et al., 2002] Ajmera, J., Bourlard, H., and Lapidot, I. (2002). Improved unknown-multiple speaker clustering using hmm. In *Technical Report IDIAP*.

[Ajmera and Wooters, 2003] Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proc. of the IEEE workshop on Automatic Speech Recognition and Understanding*, pages 411–416.

[Anguera, 2005] Anguera, X. (2005). Xbic: Real-time cross probabilities measure for speaker segmentation. Technical report, ICSI.

[Anguera, 2006] Anguera, X. (2006). *Robust speaker diarization for meetings*. PhD thesis, Universitat Politecnica de Catalunya.

[Anguera et al., 2012] Anguera, X., Bozonnet, S., and Evans, N. (2012). Speaker diarization: A review of recent research. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 20, pages 356–370.

[Boakye et al., 2008] Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G. (2008). Overlapped speech detection for improved speaker diariztion in multiparty meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4353–4356, Las Vegas, NV.

[Butko and Nadeu, 2011] Butko, T. and Nadeu, C. (2011). Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results and discussion. *EURASIP Journal on Audio, Speech and Music Processing 2011*.

[Castaldo et al., 2008] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. (2008). Stream-based speaker segmentation using speaker factors and eigenvoices. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4133–4136, Las Vegas, NV.

[Castan, 2009] Castan, D. (2009). Estudio de métodos para la clasificación de música y voz. Master's thesis, Universidad de Zaragoza.

[C.Barras et al., 2004] C.Barras, Gauvain, J. L., Meignier, S., and Zhu, X. (2004). Improving speaker diarization. In *RT 04 Fall Workshop*.

[Charlet et al., 2013] Charlet, D., Barras, C., and Liénard, J.-S. (2013). Impact of overlapping speech detection on speaker diarization for broadcast news and debates. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7707–7711, Vancouver, BC.

[Chen and Gopalakrishnan, 1998] Chen, S. S. and Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 127–132.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 28, pages 357–366.

[Do, 2003] Do, N. N. (2003). Fast approximation of kulback-leibler distance for dependence trees and hidden markov models. In IEEE, editor, *Signal Processing Letters*, pages 115–118.

[Friedland et al., 2009] Friedland, G., Vinyals, O., Huang, Y., and Müller, C. (2009). Prosodic and other long-term features for speaker diarization. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 17, pages 985–993.

[Fukunaga and Hostetler, 1975] Fukunaga, K. and Hostetler, L. D. (1975). The estimation of the gradient of a density function,with application in pattern recognition. In *IEEE Transactions on Information Theory*, volume IT-21, pages 32–40.

[Gupta et al., 2008] Gupta, V., Boulianne, G., Kenny, P., Ouellet, P., and Dumouchel, P. (2008). Speaker diarization of french broadcast news. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4365–4368, Las Vegas, NV.

[Huang and Hansen, 2006] Huang, R. and Hansen, J. H. L. (2006). Advances in unsupervised audio clasification and segmentation for the broadcast news and ngsw corpora. In *IEEE Transactions on Speech and Audio Processing 2006,*, volume 14, pages 907–919.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing. A guide o theory, algorithm and system development*. Prentice hall PTR.

[Hung et al., 2000] Hung, J.-W., Wang, H.-M., and Lee, L.-S. (2000). Automatic metric-based speech segmentation for broadcast news via principal component analysis. *Interspeech*, pages 121–124.

[Imseng and Friedland, 2010] Imseng, D. and Friedland, G. (2010). Tuning-robust initialization methors for speaker diarization. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*.

[J. L. Gauvain et al., 1999] J. L. Gauvain, L. L., Adda, G., and Jardino, M. (1999). The limsi 1998 hub-4e transcription system. In *Proc. of the DARPA Broadcast News Workshop*, pages 99,104.

[Jin et al., 1997] Jin, H., Kubala, F., and Schwartz, R. (1997). Automatic speaker clustering. In *DARPA Speech Recognition Workshop*, pages 108–111.

[Kenny et al., 2007] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Speaker and session variability in gmm-based speaker verification. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 15, pages 1448–1460.

[Kenny et al., 2010] Kenny, P., Reynolds, D., and Castaldo, F. (2010). Diarization of telephone conversations using facor analysis. *IEEE Journal on Selected Topics in Signal Processing*, 4:1059–1070.

[Levinson, 1986] Levinson, S. E. (1986). Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, pages 29–45.

[Nguyen et al., 2008] Nguyen, T. H., Chng, E., and Li, H. (2008). T-test distance and clustering criterion for speaker diarization. Technical report, Nanyang Technological University.

[Reynolds and Torres-Carrasquillo, 2005] Reynolds, D. and Torres-Carrasquillo, P. (2005). Approaches and applications of audio diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 953–956, Philadelphia, PA.

[Reynolds et al., 1998] Reynolds, D. A., Singer, E., Carlson, B. A., O'Leary, G. C., McLaughlin, J., and Zissman, M. A. (1998). Blind clustering of speech utterances based on speaker and language characteristics. In *ICSLP*.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. In *The Annals of Stadistics 1978*, volume 6, pages 461–464.

[Senoussaoui et al., 2013] Senoussaoui, M., Kenny, P., Dumouchel, P., and Stafylakis, T. (2013). Efficient iterative mean shift based cosine dissimilarity for multi-recording speaker clustering. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7712–7715, Vancouver, BC.

[Siegler et al., 1997] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. (1997). Automatic segmentation,classification and clustering of broadcast news audio. In *Proc. DARPA Speech Recognition Workshop*, pages 97–99.

[Sinha et al., 2005] Sinha, R., Tranter, S. E., Gales, M. J. F., and Woodland, P. C. (2005). The cambridge university march 2005 speaker diarisation system. *Interspeech*.

[Stafylakis et al., 2013] Stafylakis, T., Kenny, P., Gupta, V., and Dumouchel, P. (2013). Compensation for inter-frame correlations in speaker diarization and recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7731–7735, Vancouver, BC.

[Tranter and Reynolds, 2006] Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. In *IEEE Transactions on Audio, Speech and Audio Processing*, volume 14, pages 1557–1565.

[Valente and Wellekens, 2004] Valente, F. and Wellekens, C. J. (2004). Variational bayesian speaker clustering. In *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*.

[van Leeuwen, 2010] van Leeuwen, D. (2010). Speaker linking in large data sets. *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*.

[Vaquero, 2011] Vaquero, C. (2011). *Robust Diarization for Speaker Characterization*. PhD thesis, Universidad de Zaragoza.

[Vaquero et al., 2010] Vaquero, C., Ortega, A., Villalba, J., Miguel, A., and Lleida, E. (2010). Confidence measures for speaker segmentation and their relation to speaker verification. *Interspeech*, 2010:2310–2313.

[Willsky and Jones, 1976] Willsky, A. and Jones, H. (1976). A generalized likehood ratio approach to the detection and estimation of jumps in linear systems. In *IEEE Transactions on Automatic Control*, pages 108–112.

[Wooters et al., 2004] Wooters, C., Fung, J., Peskin, B., and Anguera, X. (2004). Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In *In RT04F Workshop*.

[Zelenák et al., 2012] Zelenák, M., Schulz, H., and Hernando, J. (2012). Speaker diarization of broadcast news in albayzin 2010 evaluation campaign. *EURASIP Journal on Audio, Speech and Music Processing 2012*.

[Zhou and Hansen, 2005] Zhou, B. and Hansen, J. H. L. (2005). Efficient audio stream segmentation via the combined t2 stadistic and bayesian information criterion. In *IEEE Transactions on Speech and Audio Processing*, volume 13, pages 467–474.