RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de diarización de locutores sobre señales telefónicas

Gabriel Hernández Sierra

RT_081 **febrero 2016**

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

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de diarización de locutores sobre señales telefónicas

Gabriel Hernández Sierra

RT_081 **febrero 2016**

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

Tabla de contenido

Lista de figuras

Métodos de diarización de locutores sobre señales telefónicas

Gabriel Hernández Sierra

Equipo de Imágenes y Señales, CENATAV - DATYS, La Habana, Cuba

RT_081, Serie Azul, CENATAV - DATYS
Aceptado: 29 de diciembre de 2015

Resumen. La diarización del locutor es la tarea de determinar "¿Quién hablo cuando?", en una señal de voz que contiene una cantidad desconocida de habla y también un número desconocido de locutores. En este reporte técnico ofrecemos una visión general de los enfoques utilizados actualmente en la Diarización del Locutor sobre señales telefónicas, describiendo los principales métodos de referencia en la literatura que enfrentan los retos de: no utilizar información a priori del locutor y procesan la señal de voz en línea, obteniendo respuestas de la diarización a medida que se incrementa el flujo de la señal. El proceso de diarización fue dividido en dos etapas para un mejor entendimiento, la primera se enfoca en segmentar la señal de voz a partir de los puntos de cambios de los locutores y la segunda profundiza en los métodos de agrupamiento de locutores. Finalmente se describe, el protocolo de evaluación que determina el error de los sistemas y los sistemas de referencia en el estado del arte.
Palabras clave: diarización del locutor, segmentación, agrupamiento, compensación de sesión

Abstract. Automatic Speaker Diarization is the task of determining "Who spoke when?" in an speech signal that contains an unknown amount of speech and also an unknown number of speakers. In this technical report we provide an overview of the approaches currently used in speaker diarization on telephone signals. This work describes the main methods in state of the art, with their relative merits and limitations facing the following challenges: the prior information of speaker can not be used and the speech signal is processed online, ie. responses of the diarization are obtained while increases the signal flow. The diarization process was divided into two stages for better understanding, the first focuses on the speech signal segmentation from the change points of the speakers and the second explores the clustering methods of the speakers. Finally, the evaluation protocol that determines the error of systems and reference systems in state of the art are described.
Keywords: speaker diarization, segmentation, clustering, session compensation.

1 Introducción

En la actualidad con los niveles de grabaciones de voz y la necesidad de realizar la tarea de diarización del locutor en el menor tiempo posible, es impracticable utilizar personas para segmentar y agrupar según la identidad de cada locutor en la señal de voz, que además puede contener segmentos musicales, tonos, locutores irrelevantes (operadoras), etc. Este problema fue la raíz que provoco el auge existente en la actualidad por los sistemas de diarización del locutor y aunque mucho se ha desarrollado e investigado en esta área, aún el nivel de conocimiento dista mucho del alcanzado en áreas como el reconocimiento del habla y el reconocimiento del locutor.

En general un documento hablado, es una grabación por un canal de audio que contienen múltiples fuentes de voces, que pueden ser: diferentes locutores, segmentos de música, tipos de ruido y otras características del canal etc. Ejemplos de documentos hablados son: audio de noticieros de radiodifusión,

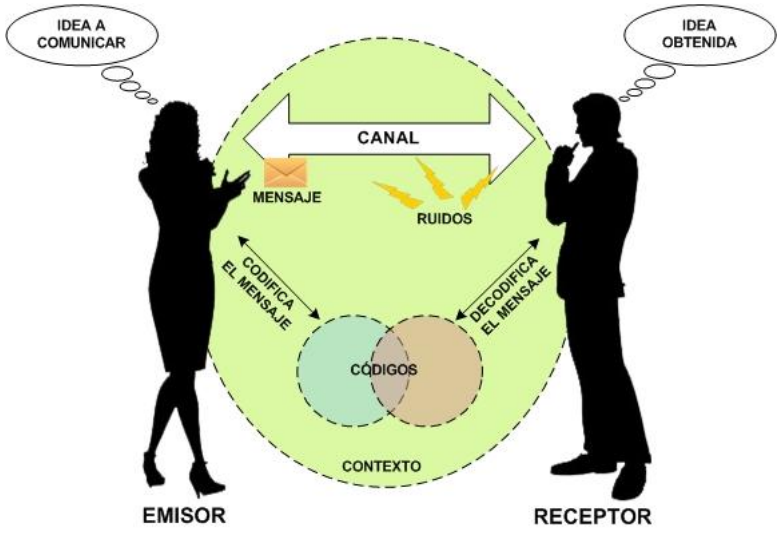

Fig. 1. Modelo de comunicación oral.

conversaciones telefónicas y grabaciones de reuniones (Modelo de comunicación oral, Figura 1). El audio de un noticiero de radiodifusión contiene voces de diferentes locutores, así como segmentos musicales, comerciales y sonidos propios de los reportajes.

La diarización del locutor consiste en dividir un flujo de audio de entrada con múltiples locutores (probablemente solapados en el tiempo) en partes homogéneas y luego agrupar las partes resultantes de acuerdo con la identidad de cada locutor; implicando una combinación en una misma tarea de la segmentación y la agrupación de locutores en documentos hablados. Entonces podemos dividir la diarización del locutor en dos etapas para una mejor comprensión:

1. Segmentación: Su objetivo consiste en buscar los puntos de cambios de los locutores en el flujo de audio (segmentos).
2. Agrupación: Su finalidad es agrupar los segmentos sobre las base de las característica de cada locutor (n-locutores).

La diarización del locutor más sencilla es la detección del habla no-habla, donde la clase “no habla” es en general una clase que consiste en música, silencios y ruidos, no requiere en general otra división posterior por tipos. Otras aplicaciones pueden necesitar tener más detalles de las clases, como: localizar la música, detectar habla de banda estrecha, etiquetar el habla solo por el género del locutor, etc. La diarización habla no-habla es muy útil como etapa de pre-procesamiento para muchas tecnologías del habla como el reconocimiento automático del habla, de locutores y del idioma.

Una diarización del locutor más compleja consiste en marcar donde ocurren los cambios de locutores en el habla detectada y la asociación posterior de los segmentos de habla procedentes del mismo locutor, respondiendo a “¿quién hablo cuando?”. Posibles utilidades de la diarización son:

- Sistema de indexación de locutores, permite a los usuarios acceder directamente a los segmentos relevantes y de interés dentro de un audio dado, facilitando a otros procesos posteriores como el resumen y el análisis.
- Su combinación con el reconocimiento automático del habla, los meta-datos extraídos con la diarización del locutor proporcionan información complementaria para las transcripciones, incluyendo el

locutor de turno. Esta información tributa en una mayor legibilidad en las transcripciones automáticas del habla, al segmentar el habla (y el texto transcrito) por locutores, respondiendo a la pregunta ¿Quién y cuándo dijo que?.
- El volumen de habla agrupada por locutor, brinda información que puede ser usada en la adaptación no supervisada del locutor en sistemas de reconocimiento de habla (sistemas de transcripción habla-texto).
- Cuando se utiliza junto con los sistemas de reconocimiento de locutor, permitiendo agrupar los segmentos de voz de cada una de las fuentes para proporcionar la verdadera identidad de cada hablante en el flujo de audio.

Algunos ejemplos de referencia de la utilidad de las técnicas de diarización del locutor se muestran a continuación. Gish en 1991 ([1]) llevó a cabo la segmentación y la agrupación de locutores en grabaciones de radio entre controladores y pilotos de tránsito de aeropuertos. En [2,3] se enfocaron en realizar la diarización de conversaciones telefónicas mixtas, como un primer paso para la verificación del locutor en las evaluaciones NIST [4].

Hay tres dominios principales que se han utilizado para la investigación en la diarización del locutor y su desarrollo a lo largo de los años. Estos son: audio de noticieros de radiodifusión [5,6], conversaciones telefónicas [7] y reuniones grabadas [8]. Con el creciente flujo de información recogido cada año, la diarización del locutor ha recibido mucha atención por parte de la comunidad científica, como se manifiesta en las evaluaciones dedicadas a ella bajo el auspicio del Instituto Nacional de Estándares y Tecnología (NIST-SRE[1] [9]). Esto ha conllevado al desarrollo de nuevos algoritmos de segmentación y agrupación de locutores, junto con los desafíos únicos para la diarización que presentan cada uno de los dominios de aplicación, tales como: *diferente número de locutores presentes en el audio*, *el tiempo de respuesta*, la calidad de las grabaciones, la cantidad y los tipos de ruido de fondo, la reverberación, la duración y la secuencia de segmentos de los locutores, la cantidad de habla que es probable que se superponga, y si el habla es con guión o espontánea; constituyendo un campo abierto a la investigación científica. El trabajo presentado en esta investigación se centrará en el dominio de las *conversaciones telefónicas*.

Dadas las aplicaciones prácticas de los sistemas diarización del locutor y los problemas existentes, el **objetivo principal** de este trabajo consiste en realizar un estudio sobre las diferentes técnicas utilizadas tanto en la etapa de segmentación como en el agrupamiento en la diarización del locutor, principalmente en aquellas que permitan obtener una *respuesta mientras se produce el flujo de voz.*

Resultados esperados:

Cumpliéndose con el objetivo de este trabajo y junto con los resultados de la tesis “Métodos de representación y verificación del locutor con independencia del texto” desarrollada en nuestro centro, arribaríamos a herramientas competentes para de forma **dinámica**:

- segmentar señales con voces de varios locutores
- agrupar los segmentos por su identidad
- reconocer la identidad de cada locutor

y enfrentar problemas reales como la variabilidad del canal.

El proceso básico de diarización del locutor comprende generalmente tres etapas principales, la actividad de detección de voz, la segmentación del locutor y la agrupación de los locutores, como se muestra en la Figura 2.

[1] NIST-SRE: evaluaciones de reconocimiento de locutores llevadas a cabo por el instituto Nacional de Estandarización de los EEUU.

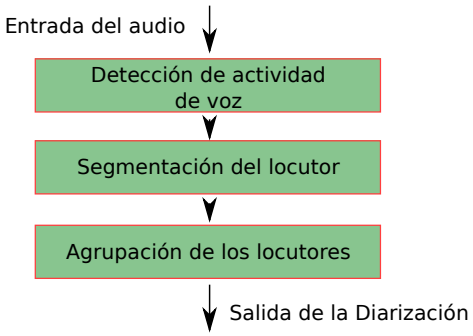

Fig. 2. Proceso básico de diarización de locutores.

Actividad de detección del habla (SAD): El propósito de esta etapa es clasificar el audio en regiones de habla y no habla. Es importante identificar y descartar regiones de no habla como la música y el ruido al principio del proceso diarización, para evitar obstaculizar los procesos posteriores de segmentación y agrupamiento del locutor. El objetivo de esta etapa es eliminar solamente períodos prolongados de silencio, música o ruido, en lugar de las pausas cortas en el habla del locutor de turno, porque podrían romper los segmentos homogéneos de un mismo locutor. Todas las regiones de audio clasificados como no-habla se excluyen de su posterior procesamiento. La SAD se utiliza comúnmente como un paso de pre-procesamiento en muchas aplicaciones sobre la voz, incluyendo la diarización del locutor, la verificación del locutor y el reconocimiento de voz. En los últimos años, las técnicas basadas en GMM (Modelos de Mezclas Gaussianas) [10,11,8,12,13,14] se han convertido en el enfoque dominante a la tarea SAD. Por otra parte, la SAD ya se encuentra desarrollada y utilizada en la tesis "Métodos de representación y verificación del locutor con independencia del texto" desarrollada en nuestro centro, por lo que esta etapa no se abordará en este trabajo.

En este trabajo se abordaran las dos ultimas etapas, exponiendo las ventajas y desventajas de los algoritmos de referencias en la actualidad, además se mostrará la poca investigación realizada o existente sobre la dinámica (Diarización en tiempo real) en estos sistemas. En los sistemas del estado del arte estas dos etapas, por lo general, se abordan de forma simultánea. En estos casos la noción de módulos de segmentación y agrupamiento estrictamente independientes es menos relevante. Sin embargo, los dos módulos son fundamentales para la tarea de diarización del locutor [13,14], en este trabajo serán abordados de forma separada para un mejor entendimiento.

2 Segmentación de locutores

El objetivo de esta etapa es encontrar los probables puntos, en el flujo de audio, que sean puntos de cambio entre las fuentes de voz. El método empleado para encontrar los puntos de cambio es lo que distingue a los diferentes sistemas de segmentación de locutores. En la bibliografía ([15,16,17,18]) por lo general son separados en varios grupos:

- Segmentación Basada en la energía: El flujo de entrada es segmentado en partes a partir del silencio, para esto se utiliza un detector de energía y se establecen los puntos de cambio en los mínimos que superan un umbral determinado [16,19,20]. En [21] se utiliza MAD (del inglés Mean absolute deviation statistic) para medir la variabilidad de los segmentos sin energía para encontrar los puntos de cambios.
- Segmentación basada en Modelos: En la década del 1990 se crean modelos iniciales (por ejemplo, Modelos de Mezclas Gaussianas) para un conjunto cerrado de clases acústica (teléfono-banda ancha,

hombres-mujeres, música-voz-silencio y combinaciones de ellos) usando datos de entrenamiento. Entonces el flujo de audio se clasifica a partir de la ML (máxima verosimilitud) con estos modelos ([16,22,23]). Los límites entre los modelos se vuelven los puntos de cambio para la segmentación. En años posteriores, algunos sistemas de agrupamiento hacen uso de una decodificación por ML con modelos evolutivos que buscan los puntos de cambio con una acústica óptima. En [24] una segmentación previa en modelos entrenados se combina con una segmentación evolutiva. Por último, en [25] una SVM (del inglés Support Vector Machines) se utiliza como un clasificador en lugar de la ML de los modelos entrenados.

- Segmentación basada en métricas: El flujo de audio es segmentado evaluando una métrica entre dos segmentos vecinos de audio (de la misma o diferente longitud). Los puntos de cambio son los máximos locales de los segmentos que lo rodean. Hay muchos indicadores posibles de usar, se propone separar en dos grupos:
  - Algoritmos BIC [26]: Criterio de Información Bayesiana, el cual es el algoritmo más utilizado en la literatura, junto a su variante ΔBIC [27,28]. BIC utiliza información Bayesiana para medir la similitud entre dos segmentos adyacentes de voz y decidir si puede existir entre ellos un punto de cambio de locutor. La similitud se mide calculando las probabilidades cruzadas entre dos modelos entrenados con cada uno de los dos segmentos y su evaluación en el segmento contrario. Esto permite la creación de sistemas de segmentación en tiempo real.
  - Otras medidas: Muchas otras distancias se han propuesto con el fin de localizar los puntos de cambio en una secuencia de audio, en [29] la GLR (del inglés Generalized Likelihood Ratio) se utiliza para segmentar la señal por locutores haciendo una verificación del locutor, en [30] utilizan como medida K-L2 (Kullback-Leibler) y en [31] utilizan una medida llamada NLLR (Normalized Log Likelihood Ratio).
- Algoritmos híbridos: Estos algoritmos combinan la segmentación basada en métricas con la segmentación basada en modelos. Usualmente la segmentación basada en la métrica es empleada inicialmente para pre-segmentar el flujo de audio, luego los segmentos resultantes son utilizados para crear un conjunto de modelos de locutores y por último se realiza una re-segmentación basada en modelos [32,33].

De estos diferentes enfoques los más robustos y eficaces son la segmentación basada en Modelos, las Métricas y los Híbridos, aunque el más popular y eficiente de todos es la métrica basada en el algoritmo BIC [13,14]. Por este motivo, nos concentraremos en esos enfoques.

2.1 Segmentación basada en modelos

En este tipo de segmentación, un conjunto de modelos se deriva y se entrena, para diferentes clases de locutores, a partir de una base con voces de entrenamiento. El flujo de habla es clasificado utilizando estos modelos, como resultado se convierte en un requisito el conocimiento previo para inicializar los modelos de los locutores. Además, se entrena un modelo universal de fondo (UBM) previo a todo, para crear un modelo genérico del locutor [10,34,35]. Durante la segmentación, este modelo discrimina entre segmentos de voz y no voz. Obsérvese en este caso, que al tener precalculado los modelos de los locutores, el algoritmo se puede utilizar en tiempo real.

Este enfoque permite crear de ante mano varios tipos de modelos, por ejemplo un modelo universal de genero (UGM), permitiendo reconocer el género del locutor de turno en el audio. Una técnica más complicada es el modelo de anclaje, donde una expresión del locutor se proyecta a un espacio de locutores de referencias [36]. Otros modelos pueden ser creados por las Cadenas Ocultas de Markov (HMM) [37,33] o por Maquinas de Vectores de Soportes [38,39].

2.2 Segmentación basada en métricas

La segmentación basada en una métrica evalúa la diferencia entre ventanas vecinas, usando una función de distancia que se desplaza sobre un flujo de audio, estas ventanas pueden o no estar solapas en dependencia de la aplicación. La máximos locales de la función de distancia que superan un umbral determinado se consideran como puntos de cambio. Estos métodos no requieren ningún conocimiento previo sobre el número de hablantes, sus identidades, o las características de la señal.

Una amplia variedad de métricas ha sido utilizada. Una de las métricas comúnmente empleada es la divergencia de Kullback-Leibler [31] o la divergencia Gaussiana (también conocida como divergencia simétrica Kullback-Leibler-2) [10]. Otras han sido, la razón de verosimilitud generalizada (GLR) [33,40], la perdida de Entropía [16], estadísticas de segundo orden [41,42], Estadística T^2 Hotelling [43,44], etc. El criterio más popular es el BIC [45,46,26,47,41,48,43,49,50], introducido inicialmente por Chen y Gopalakrishnan en [26].

2.2.1 Algoritmo de segmentación BIC

Es un criterio de selección de modelos Bayesianos asintóticamente óptimo, utilizado para decidir cuál de los modelos paramétricos representa las mejores muestras N de los datos $x_i \in \mathbb{R}^d, i = 1,2,\ldots N$. Cada una de estas muestras x_i son simples vectores de dimensión d, teniendo como elementos los rasgos acústicos (Coeficientes Cepstrales de Predicción Lineal (LPCC) [51], Coeficientes Cepstrales en escala Mel (MFCC) [51] u otros).

Para la segmentación del locutor, solo dos modelos diferentes son utilizados, asumiendo dos ventanas de análisis vecinas X y Y alrededor del tiempo t_j. El problema consiste en saber cuando ocurre un punto de cambio en t_j, dado $Z = X \cup Y$, se formula el problema como una prueba estadística entre dos hipótesis, H_0: no existe un punto de cambio en el tiempo t_j. Las muestras pertenecientes a Z son modeladas por la función de densidad de probabilidad Gaussiana, cuyos parámetros son el vector de medias, y la matriz de covarianza, los cuales confeccionan el modelo M_Z. El modelo M_Z puede ser estimado vía la máxima verosimilitud (ML) o emplear otros estimadores robustos, tales como $M - estimador$ [52]. La verosimilitud L_0 es calculada por:

$$L_0 = \sum_{i=1}^{N_X} log\ p(x_i|M_Z) + \sum_{i=1}^{N_Y} log\ p(y_i|M_Z), \tag{1}$$

donde N_X y N_Y son el número de muestras en la ventana correspondiente X y Y.

Bajo la hipótesis H_1 ocurre un punto de cambio del locutor en el tiempo t_j, para ello las ventanas de análisis X y Y son modeladas por distintas densidades Gaussianas, cuyos modelos son M_X y M_Y respectivamente. Entonces la verosimilitud L_1 se obtiene por:

$$L_1 = \sum_{i=1}^{N_X} log\ p(x_i|M_X) + \sum_{i=1}^{N_Y} log\ p(y_i|M_Y). \tag{2}$$

Luego la disimilitud entre las dos ventanas vecinas X y Y se estima por el criterio BIC, definido como:

$$D(X;Y) = H_1(X) + H_1(Y) - H_0(X \cup Y), \tag{3}$$

de forma análoga,

$$\delta = L_1 - L_0 - \frac{\gamma}{2}\left(d + \frac{d(d+1)}{2}\right) log\ N_Z, \tag{4}$$

donde $N_Z = N_X + N_Y$ es la cantidad de muestra de la ventana Z, d es la dimensión de los rasgos y γ es un factor de penalidad dependiente de los datos (se asume 1,0). Si $\delta > 0$, se encontró un máximo local y t_j se considera un punto de cambio de locutor. Si $\delta < 0$ no existe un punto de cambio en el tiempo t_j.

En ec. 4, $L_1 - L_0$ se refiere a la calidad de la comparación entre las ventanas de análisis X y Y, mientras que el término $-\frac{\gamma}{2}\left(d + \frac{d(d+1)}{2}\right) log\, N_Z$ es un factor de penalidad para la complejidad del modelo. En teoría de la codificación, el algoritmo BIC con γ igual a 1 representa la longitud del código más corto con que los datos pueden ser codificados. En la segmentación del locutor, γ sirve como un umbral. Su elección es dependiente de la tarea, además γ nos es robusto ante diferentes condiciones acústicas y ambientales. En consecuencia, γ requiere ser ajustado para cada aplicación.

Un criterio que no requiere ajuste (γ) (ver ec. 4), es propuesto por Ajmera y colegas en [53]. Las muestras bajo H_0 son modeladas por un modelo de mezcla Gaussiana (GMM) con dos componentes en lugar de una sola. Dado el nuevo modelo M'_Z que contiene los parámetros del GMM, se modifica 4 a:

$$\delta' = L_1 - \sum_{i=1}^{N_X} log\ p(x_i|M'_Z) + \sum_{i=1}^{N_Y} log\ p(y_i|M'_Z). \tag{5}$$

El número de parámetros que se utilizan para modelar los datos en las dos hipótesis se ve obligado a ser el mismo, de modo que las probabilidades son directamente comparables. Como resultado, los autores en [53] afirman que no es necesario ningún ajuste y se espera que el criterio sea robusto a las condiciones cambiantes de los datos [13].

2.3 Algoritmos híbridos

Estos algoritmos combinan técnicas basadas en métricas y modelos. Por lo general, la segmentación basada en métricas se utiliza inicialmente para pre-segmentar la señal de audio de entrada. A continuación los segmentos resultantes son utilizados para crear un conjunto de modelos de locutores y realizar una re-segmentación basada en modelos, resultando en una segmentación más refinada. En [32], un HMM se combina con BIC, en [54] después de realizar una segmentación inicial con BIC, los cambios acústicos no encontrados son detectados a través de una técnica de divide y vencerás. Otro sistema híbrido interesante se introduce en [33], donde se combinan dos sistemas, ALIZE [55] que se basa en HMM y el sistema CLIPS [33] (Communication Langagiere et Interaction Personne-Systeme), que realiza la segmentación del locutor basada en BIC seguido por una agrupación jerárquica.

En [56,14], los mejores componentes de dos sistemas diferentes de diarización del locutor presentes en el estado del arte e implementados por dos laboratorios Franceses (LIUM[2] [57] y IRIT[3] [58]) se fusionaron o se utilizaron de forma secuencial, obteniendo una significativa mejora en el desempeño en comparación con los sistemas individuales.

2.4 Conclusiones: ventajas y desventajas

En esta etapa de la diarización de locutores, los algoritmos basados en Modelos no son de interés para la investigación que se realiza. Los mismos necesitan información previa para realizar la segmentación, lo cual no suele suceder en el dominio de las señales telefónicas. Por otra parte los algoritmos híbridos implican la necesidad de una pre-segmentación en busca de la información previa de los modelos, lo

[2] Laboratorio de Informática de la Universidad de Maine, Francia
[3] Laboratorio de la Universidad de Toulouse

que implica que estos algoritmos trabajen necesariamente sobre la señal completa, conllevando a su poca utilidad en aplicaciones que necesitan analizar el flujo de voz en línea. Por tales motivos nos centraremos en las ventajas y desventajas de la segmentación basada en métricas que se expresan a continuación.

Ventajas de la *segmentación basada en métricas* (algoritmo BIC):

- No asumen ningún conocimiento previo del número de los locutores, sus identidades, o características de la señal. Esto implica la posibilidad de *segmentar en tiempo real*.
- Se puede aplicar a ventanas de análisis de diversas duraciones, que exhiban diferentes superposiciones. La elección del tamaño de la ventana de análisis N_Z es de gran importancia. Por un lado, si N_Z es demasiado grande, puede contener más de un cambio de locutor y en consecuencia producir un elevado número de omisiones en la detección. Por otro lado, si N_Z es demasiado pequeña, la falta de muestras va a causar una mala estimación de las componentes Gaussianas, especialmente de la matriz de covarianza y como resultado, una pobre precisión en la segmentación [43]. Una duración típica de la ventana de análisis inicial es de 2 s, que en la mayoría de los casos aumenta de forma incremental [47,34,53,43]. Esto se debe a que los investigadores coinciden en que el rendimiento del algoritmo BIC es pobre, cuando dos cambios sucesivos de locutores se separan menos de 2 s [45,46,43]. Sobre la base de esta observación, varios investigadores adoptan la actualización continua de los modelos de locutores [45,46,48,54].
- La segmentación puede ser realizada con diferentes rasgos acústicos para la fusión posterior de los resultados, ya que los diferentes rasgos pueden complementarse en diferentes contextos [45,59,60]. Además, se podría realizar la segmentación con varias métricas y/o varios clasificadores, y luego fusionar los resultados individuales. En general, la fusión ofrece muchas ventajas, como el aumento del rendimiento y robustez de un sistema [61].

Desventajas de la *segmentación basada en métricas* (algoritmo BIC):

- La mayoría de las variantes de los algoritmos BIC utilizan un término de penalización que depende del contexto de la aplicación, este término, en general, es difícil de estimar, de tal manera que presente un rendimiento estable frente a diferentes contextos. Por esto, existen varias modificaciones para adaptar o eliminar el mismo [31,53,13,14]
- En cuanto al costo computacional, la aplicación completa del BIC es computacionalmente costosa, alcanzando el orden de $O(N^2)$ [13]. El alto costo computacional ha motivado a los investigadores en el área a emplear heurísticas para resolver el problema.
- El rendimiento o eficacia del algoritmo BIC para la segmentación es menor que cuando se utiliza la segmentación basada en Modelos [37,33] precedida de una pre-segmentación. Esto ha implicado que los algoritmos híbridos utilicen la segmentación basada en métricas BIC en una etapa de pre-segmentación y luego se refine la misma con la segmentación basada en modelos [54,32] o con otro pase del algoritmo BIC[4].
- Se debe tener en cuenta que el supuesto Gaussiano de las muestras acústicas no siempre es correcto [62].

[4] La razón principal para llevar a cabo una pre-segmentación es para asegurar que las ventanas de análisis sean más grandes que 2 s, lo que permite al algoritmo BIC producir resultados más precisos.

3 Agrupación de locutores

El agrupamiento de objetos se encarga de organizar una colección de objetos en clases o grupos, de forma tal que los objetos pertenecientes a un mismo grupo sean lo suficientemente similares como para poder inferir que son del mismo tipo y los objetos pertenecientes a grupos distintos sean lo suficientemente diferentes como para poder afirmar que son de tipos diferentes [63].

Mientras que la etapa de segmentación del locutor opera en ventanas adyacentes a fin de determinar si se corresponden o no con el locutor, la agrupación tiene como objetivo identificar y agrupar los segmentos del mismo locutor que pueden ser localizados en cualquier parte del flujo de audio. Idealmente, habrá un grupo de segmentos para cada locutor. Sin embargo, los algoritmos de agrupación se enfrentan a un problema fundamental, no existe provisión para los segmentos que contienen más de un locutor, por lo que estos algoritmos solo pueden funcionar bien si la segmentación inicial presenta una alta calidad. Para solucionar el problema, los enfoques del estado del arte combinan la agrupación con una re-segmentación de forma iterativa, provocando que la mayoría de estos enfoques realicen la segmentación y la agrupación de manera simultánea [14].

Vale la pena mencionar que la complejidad de un problema de agrupación del locutor depende del tamaño de la población, la duración del segmento de voz, el ancho de banda de la señal, el ruido ambiental y si la tarea tiene que ser realizada en tiempo real o no.

3.1 Cinco criterios de agrupamiento

Para la tarea del agrupamiento en la diarización del locutor los trabajos en el estado del arte se centran en 5 criterios:

1. Agrupación utilizando cuantificación vectorial.
2. Agrupación de manera jerárquica utilizando métricas o medidas de divergencia.
3. Agrupación basada en modelos.
4. Agrupación utilizando un decodificador HMM.
5. Agrupación utilizando tiempo de retardo de llegada.

Los enfoques (1,2) y (3) realizan la agrupación de locutores sobre una serie de segmentos de habla no identificados. Estos enfoques tienen sus raíces en la identificación del locutor y de hecho el problema subyacente de la agrupación de locutores se puede reducir a decidir, si dos expresiones del habla dados fueron hechas por un mismo locutor.

Los enfoques (4) y (5) combinan la etapa de la segmentación y el agrupamiento. Estos algoritmos realizan al mismo tiempo la segmentación y la agrupación, es decir, no distinguen la segmentación y el agrupamiento como pasos separados y distintos dentro de la diarización.

3.1.1 Agrupación utilizando cuantificación vectorial

La Cuantificación Vectorial (Vector Quantization, VQ) se ha explorado en la agrupación de locutores en diferentes trabajos [64,65,66,67] y consiste en que cada vector N-dimensional de entrada (un punto en el espacio N-dimensional) se representa por el más cercano “codebook” o centroide de un pequeño grupo de centroides o “codebooks” altamente representativos de la distribución de vectores de entrada en el espacio N-dimensional. Este “codebook” se selecciona con los mejores representantes de los diferentes “clusters” o grupos no solapados en los que se hayan dividido los datos de entrenamiento, por lo tanto no es adecuado para las aplicaciones que requieran la agrupación sin supervisión de los locutores o información previa.

3.1.2 Agrupación de manera jerárquica utilizando métricas o medidas de divergencia

La agrupación jerárquica, en esencia, es un algoritmo de "divide y vencerás" [68], se divide la tarea de agrupar los locutores dentro de una grabación en una serie de sub-tareas de agrupamiento. Cada sub-tarea trabaja en un sub-conjunto de los segmentos obtenidos. El propósito final de la etapa de agrupamiento consiste en asociar o agrupar los segmentos de cada locutor, produciendo idealmente un grupo para cada locutor en el audio. La técnica de agrupamiento aglomerativo jerárquico consta de los siguientes pasos:

- Cada segmento de voz se asume como la semilla de un grupo, por lo tanto se inicializa cada hoja del árbol con un segmento.
- Calcular las distancias (BIC, GLR u otra) por pares entre cada grupo u hoja.
- Combinar los grupos más cercanos.
- Actualizar las distancias restantes de los grupos a los nuevos grupos.
- Iterar por los pasos 1 al 3 hasta que se cumpla un criterio de parada.

Cada grupo en general es representado por una mono-Gaussiana con covarianza completa [69,8,12,70], aunque también se han utilizado las GMM [33,71,72]. Las métricas de referencia en el estado del arte para calcular las distancias entre los grupos son: el criterio de información Bayesiano BIC (ec. 8) y la razón de verosimilitud generalizada GLR (ec. 6), aunque sea posible utilizar otras métricas. El criterio de parada compara el valor de la estadística de los dos grupos que se consideran, en el caso del BIC utilizando mono-Gaussiana con covarianza completa la ec. 10. Si los dos grupos se describen mejor por una sola Gaussiana con covarianza completa, entonces el valor de ΔBIC (ec. 9) será bajo, mientras que si hay dos distribuciones diferentes, lo que implica dos locutores, el valor del ΔBIC será alto. Por cada paso, el par de grupos con el valor del ΔBIC más bajo se fusiona y las estadísticas se vuelven a calcular. El proceso se detiene generalmente cuando el ΔBIC más bajo es mayor que un umbral especificado, por lo general 0. También se han empleado ligeras variaciones de esta técnica, por ejemplo, el sistema descrito en [22,73,70,74], donde se elimina la necesidad de sintonizar el término de penalidad.

Dos enfoques son utilizados generalmente para la agrupación de manera jerárquica (fig. 3) de los segmentos de locutores, conocidos por: el enfoque de *abajo hacia arriba* (del inglés bottom-up) que es el método más popular reportado en la literatura y el enfoque de *arriba hacia abajo* (del inglés top-down), donde todo el flujo de audio se modela primero como un único modelo del locutor, para luego ir dividiendo sucesivamente en subgrupos obteniendo nuevos modelos hasta que se logre el número total de hablantes a tenerse en cuenta. Como este enfoque, en general es mucho menos popular y es superado por los mejores enfoques de su contraparte de abajo hacia arriba [14] en el estado del arte, no será tratado en este trabajo. Ejemplos de enfoques de arriba hacia abajo en la diarización de locutores fueron desarrollados por LIA[5] y se pueden encontrar descritos en [75,76].

Agrupación Bottom-Up

En los enfoques bottom-up, la etapa de agrupación de los locutores se realiza a través de un agrupamiento aglomerativo jerárquico, siguiendo los pasos descritos al inicio se la subsección 2.1.2.

Al igual que con los algoritmos de segmentación, la similitud entre grupos se determina mediante la evaluación de una medida de distancia elegida, siendo esta elección un paso importante para el éxito del algoritmo de agrupamiento. Comúnmente, las métricas de distancia utilizadas para la segmentación también se pueden emplear para la agrupación y aunque el BIC es el enfoque predominante, a menudo se utiliza el modelado Gaussiano [69,8] utilizando Razón de Verosimilitud Generalizada (GLR) [41] como criterio de decisión para el agrupamiento. Obsérvese en la sección 2.2 en la ec.3, que el BIC se puede

[5] Laboratorio de Informática de Avignon, Francia

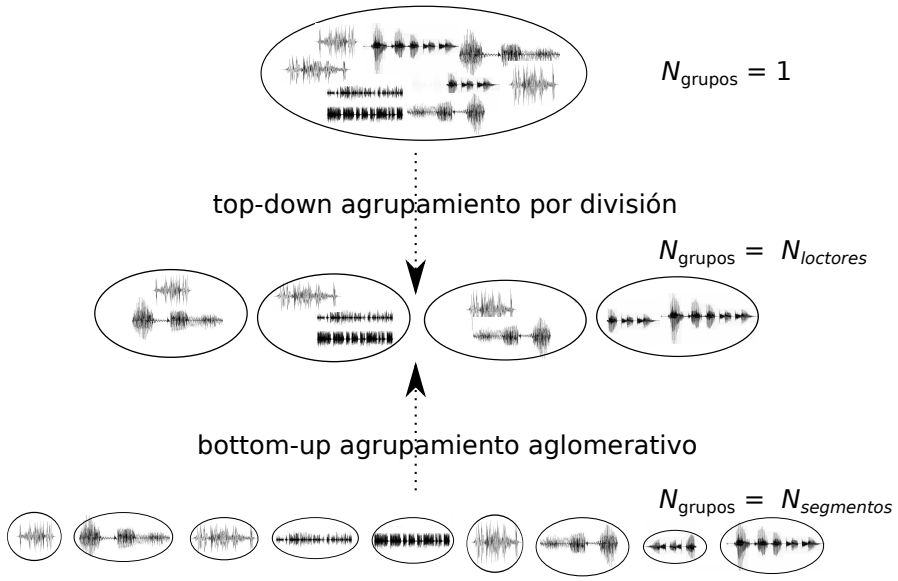

Fig. 3. Agrupación jerárquica. El objetivo consiste en obtener un número de grupos, N_{grupos}, correspondiente al número de locutores, $N_{locutores}$.

utilizar como criterio de selección del modelo, para determinar si los datos contenidos en los dos grupos de interés están representados más apropiadamente por un modelo combinado, o dos modelos separados. Para el algoritmo de agrupamiento, si el modelo combinado se ve favorecido (mayor probabilidad), esto indica que los dos grupos deberían fusionarse, de lo contrario los grupos deben mantenerse separadas. También se han propuesto variaciones de las técnicas BIC para la tarea de agrupamiento [73,22,74], donde la mayoría de las variaciones consisten en eliminar la necesidad de sintonizar el factor de penalidad. Sin embargo el método clásico BIC ha superado generalmente esos enfoques respecto a su eficacia [77].

La adición de una etapa re-segmentación utilizando Viterbi, que tiene como objetivo mejorar el rendimiento de la diarización mediante el refinado de las fronteras de los locutores, también se ha reportado en la literatura. La re-segmentación con el método Viterbi, o bien se puede realizar entre varias iteraciones del agrupamiento, o dentro de una sola iteración [10].

En esta sección se describen dos de las medidas de distancia más utilizados reportadas en la literatura de diarización del locutor, GLR y BIC, en la tarea de agrupamiento.

Razón de verosimilitud generalizada (GLR)

En el marco de las hipótesis presentadas en la sección 2.2, el GLR se formula como una prueba estadística de razón de verosimilitud que compara directamente la probabilidad de las dos hipótesis en competencia. Dado el segmento de voz $X_0 = x_i,\ i = 1,\ldots,N$, el cual esta compuesto por dos segmentos de locutores $X_1 = x_i,\ i = 1,\ldots,m$ y $X_2 = x_i,\ i = m,\ldots,N$. Además, M_0, M_1 y M_2 denotan los modelos correspondientes a cada segmento de voz, cuyos parámetros están dados por las estimaciones de máxima verosimilitud calculadas utilizando los respectivos segmentos X_0, X_1 y X_2. Entonces utilizamos GLR para determinar si existe un límite de segmento entre X_1 y X_2. En este caso el GLR se plantea como:

$$GLR = log\frac{p(X_1|M_1)\ p(X_2|M_2)}{p(X_0|M_0)}. \tag{6}$$

El denominador representa la probabilidad de los datos combinados X_0 dado el modelo combinado M_0, y el numerador representa sus homólogos independientes. El valor de GLR determina cuánto se favorece un límite de segmento entre X_1 y X_2, cuanto mayor sea el valor, más evidencia de que los dos segmentos se modelan mejor por dos distribuciones distintas por lo que deben pertenecer a locutores diferentes, y viceversa.

En el caso de un modelado mono-Gaussiana multivariable, el GLR se convierte en:

$$GLR = log\frac{p(X_1|\mu_{X_1},\Sigma_{X_1})\ p(X_2|\mu_{X_2},\Sigma_{X_2})}{p(X_0|\mu_{X_0},\Sigma_{X_0})}. \tag{7}$$

La estimación por máxima verosimilitud supone que hay un valor "correcto" para cada parámetro del modelo, siendo este valor el que maximiza la probabilidad de los datos.

Criterio de información bayesiano (BIC)

El concepto detrás del algoritmo BIC para la agrupación del locutor es muy similar a la del GLR. En la expresión del GLR (ec. 6), cada una de las probabilidades, $p(X|M)$, está dada por la estimación de máxima verosimilitud para cada modelo. En el BIC la estimación de cada probabilidad es matemáticamente equivalente a las estimaciones de máxima verosimilitud utilizadas en el GLR, excepto por el hecho de que las probabilidades son penalizadas por las complejidades del modelo, es decir, el número de parámetros utilizados en los modelos. Como un criterio de selección de modelos, el BIC se utiliza para seleccionar el modelo óptimo que mejor represente un conjunto de datos dado, a partir de una serie de candidatos de modelos paramétricos [78]. Para cualquier segmento dado, de forma general, el valor BIC está dado por:

$$BIC = p(X|M) - \frac{\lambda}{2}\cdot k\cdot log\ N, \tag{8}$$

donde $p(X|M)$ representa la máxima verosimilitud de los datos X dado el modelo M, λ denota el factor de penalidad del algoritmo, k es el número de parámetros en el modelo M, y N es el número de muestras de los datos. La decisión de si dos segmentos pertenecen al mismo locutor se determina por la variación del valor BIC entre las dos hipótesis en competencia y esta dada por:

$$\Delta BIC = GLR - \frac{\lambda}{2}\cdot \Delta k\cdot log\ N, \tag{9}$$

donde Δk es la diferencia entre el número de parámetros de las dos hipótesis. De acuerdo con [79], ΔBIC tiene la ventaja de no requerir ningún umbral. Idealmente, las decisiones de agrupamiento de los locutores deben tomarse en función de si el valor ΔBIC es mayor o menor que 0. Sin embargo, esto solo es verdadero si $\lambda = 1$ o si se cuenta con una manera sistemática para encontrar el valor óptimo de λ. De no existir esta posibilidad, λ es un umbral implícito integrado en el término de penalización [53].

En el caso de un modelado mono-Gaussiana multivariable, la puntuación de la medida BIC entre los segmentos del locutor X_1 y el X_2 puede formularse como en [79],

$$\Delta BIC = N\ log|\Sigma_0| - m\ log|\Sigma_1| - (N-m)log|\Sigma_2| - \frac{\lambda}{2}(d + \frac{d(d+1)}{2}\ log\ N, \tag{10}$$

donde d, como en la ec. 4, es la dimensión de los vectores rasgos.

Al igual que el GLR, el criterio de información Bayesiano es también un criterio de máxima verosimilitud, pero no es estrictamente un criterio de selección de modelos Bayesianos, en el sentido de que no requiere, ni tiene en cuenta ninguna información previa.

3.1.3 Agrupación basada en modelos

Las mejoras en la eficacia de la agrupación de los locutores a través del uso del factor de Bayes [69,8], sugieren que la incorporación de algún conocimiento previo acerca de todo el audio (población, canales y otras variaciones) son beneficiosas para la tarea de agrupamiento. Un criterio de decisión presente en el estado del arte que incorpora el conocimiento de todo el audio "posible" es la Razón de Verosimilitud Cruzada (CLR del inglés Cross Likelihood Ratio), que combina la información presente en ambos grupos de interés con el conocimiento del modelo de universal de fondo (UBM). El empleo del CLR para la agrupación del locutor utilizando modelos de mezcla de Gaussianas (GMM) es ampliamente reportado en la literatura de diarización, como en [10,80,12].

La capacidad del las GMMs para modelar la identidad de locutor es lo que subyace en los algoritmos de esta clase. El GMM ha demostrado previamente ser eficaz en el reconocimiento del locutor independiente del texto [81] y junto con técnicas de Análisis Conjunto de Factores (JFA)[82] se ha convertido en el enfoque más popular en el estado del arte actual de reconocimiento del locutor [83,84,85,86,87]. La continua permanencia en el estado del arte de este enfoque, gracias a la alta eficacia demostrada en las competencias internacionales [88,89,4,90,9] conllevo a su empleo en los sistemas de diarización.

Modelado del locutor con voces-propias

El modelado del locutor con voces-propias (del inglés Eigenvoice[6]), es el enfoque más reciente en el campo del procesamiento de la voz, con base en las técnicas de JFA, la idea consiste en representar locutores utilizando modelos de voces-propias [91,92,93,3]. Al igual que en los enfoques tradicionales, la técnicas de modelado utilizando voces-propias se basan en el uso de las GMM para obtener el modelo de un locutor.

Hay un acuerdo general, en el área de la voz, que tanto en la diarización del locutor como en el reconocimiento del locutor independiente del texto el tipo de modelo generativo más eficaz para distinguir entre locutores es el GMM, derivado del UBM mediante la adaptación de los vectores de medias Gaussianos, pero no las matrices de covarianza o los pesos de las mezclas [91,92]. Para desarrollar esta técnica se utiliza la formulación de los super-vectores: un super-vector del locutor es un vector de altas dimensiones obtenido mediante la concatenación de todos los vectores de medias en un GMM perteneciente al locutor.

Los sistemas actuales del estado del arte para el reconocimiento y la diarización del locutor, trabajan en el espacio de variabilidad total[7] (T) (tambien referido como i-vector) basados en un pre-entrenamiento del modelo universal UBM con un análisis de factor previo en el espacio de los super-vectores [94,95,96]. Sea F y K la dimensión de los rasgos acústicos y el número total de componentes de la mezcla Gaussiana, implicando FK dimensiones para el espacio de los super-vectores. Entonces a partir de una expresión de voz de un locutor dado se formula el super-vector dependiente M como:

$$M = m + Tw, \tag{11}$$

donde m es un super-vector independiente del locutor y del canal, que se deriva del modelo UBM, para representar el centro del espacio de parámetros de todos los locutores o de la población. T es una matriz rectangular de bajo rango con dimensión $FK \times W$ (W rango de la matriz) y w nombrado vector intermedio o i-vector, es un vector aleatorio que sigue una distribución normal estándar $\mathcal{N}(0,I)$, donde sus componentes son los factores del locutor. La eq. 11 impone severas restricciones a los super-vectores de los locutores que suelen tener decenas de miles de dimensiones y son restringidos a yacer en un subespacio afín cuya dimensión es a lo sumo unos pocos cientos.

[6] Vectores propios correspondientes a los mayores valores propios de una población de locutores

[7] Modela simultáneamente el locutor y la variabilidad del canal en un solo espacio.

El factor w es una variable oculta, la cual se puede definir por su distribución Gaussiana utilizando las estadísticas de Baum-Welch de un extracto de voz dado, resultando que la media de la distribución se corresponde exactamente con el i-vector.

Las estadísticas de Baum-Welch utilizadas para obtener el i-vector son extraídas utilizando el UBM. Dado un extracto de voz $X = \{x_1, ..., x_L\}$ y un UBM λ_{UBM} compuesto por K componentes definidos en un espacio de rasgos de dimensión F, las estadísticas de Baum-Welch son obtenidas a través de:

$$N_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM}), \tag{12}$$

$$F_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM})\ x_l, \tag{13}$$

donde N_k es la estadística de cero orden y F_k la de primer orden, con $k = \{1, ..., K\}$ índice de las Gaussianas y $P(k|x_l, \lambda_{UBM})$ corresponde con la probabilidad a posterior de la componente Gaussiana k modelando el vector de rasgos x_l. Además, con el fin de estimar los i-vectores, también se necesita calcular las estadísticas de primer orden centralizadas de Baum-Welch, basadas en los vectores medios de los componentes del UBM.

$$\tilde{F}_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM})\ (x_l - m_k), \tag{14}$$

donde m_k es el vector medio de la componente Gaussiana k del UBM. Tanto la estadística de cero orden como la de primer orden presentan una complejidad computacional de $O(KL)$ [97].

Luego para obtener el i-vector de la expresión de voz X se utiliza la siguiente ecuación:

$$w = H^{-1}\ T'\Sigma^{-1}\tilde{F}, \tag{15}$$

con H definido como:

$$H = I + T'\Sigma^{-1}NT. \tag{16}$$

Se define $N(X)$ como una matriz diagonal de dimensión $FK \times FK$ cuyos bloques diagonales son $N_k I(k = \{1, ..., K\})$, $\tilde{F}(X)$ es un super-vector de dimensión FK obtenido por la concatenación de todas las estadísticas de primer orden de Baum-Welch $\tilde{F}_k$ de X. La covarianza diagonal Σ es una matriz de dimensión $FK \times FK$ estimada durante el entrenamiento del factor y la misma modela la variabilidad residual no capturada por la matriz de variabilidad total T.

La complejidad computacional al calcular un i-vector w es $O(W^3 + W^2K + WFK)$ [97,**?**,**?**], donde el término W^3 proviene del cálculo de la inversa de la matriz H mientras que el término W^2K es provocado por el cálculo de $I + T'\Sigma^{-1}NT$. Obsérvese que cuando K es grande $(K > W)$ el término W^2K causa un costo computacional enorme.

La medida de similitud para comparar con éxito dos i-vectores en el espacio de variabilidad total que mejor rendimiento y menor complejidad computacional reporta, es el coseno [95]. Dados dos i-vectores w_1 y w_2 la puntuación obtenida por la similitud por coseno se define como:

$$S(w_1, w_2) = \frac{(w_1)^t\ (w_2)}{||w_1|| \cdot ||w_2||} \underset{<}{\geq} \theta, \tag{17}$$

donde θ es el umbral de decisión que permite aceptar o rechazar que dos i-vectores w_1 y w_2 sean o no del mismo locutor. Al trabajar dentro del espacio T y no tener la necesidad de proyectar de nuevo al espacio

de los super-vectores, este criterio de similitud es considerablemente menos complejo que las operaciones para calcular el logaritmo de la probabilidad [98].

El modelado de los segmentos de los locutores utilizando el enfoque i-vector es capaz de explotar el conocimiento previo altamente informativo sobre el espacio de la población (T), para encontrar un vector de baja dimensión[8] de los factores del locutor que resuma las características más destacadas del mismo, al restringir los modelos de los locutores a un subespacio lineal previamente estimado. Estas restricciones permiten estimar, con datos limitados, un modelo del locutor fiable.

Este enfoque permite llevar a cabo una diarización en línea, donde el flujo de audio entrante se trata como una corriente de segmentos con tiempo de duración fija (1 segundo [99]). La segmentación y la agrupación se realiza entonces de manera causal o dinámica, es decir una porción de audio entrante se procesa sobre la marcha sin requerir los siguientes segmentos. Es importante tener en cuenta que la estimación de los modelos de los locutores y la detección del locutor de turno requieren baja complejidad, con el fin de hacer frente a la transmisión del audio. Entonces la cantidad de componentes Gaussianos tiene que ser menor que el rango de la matriz T, $(K < W)$. Un esquema de ello es el propuesto por Castaldo en [99], el cual utiliza un conjunto vectores de voz de baja dimensión y con un alto solapamiento. De esta forma se puede crear un nuevo espacio de locutores aplicable a la diarización con excelentes resultados. Por otra parte, Najim y Kenny [84] mejora el esquema clásico de análisis de factores mediante el modelado de la voz y la variabilidad del canal.

Un sistema de diarización no dinámico basado en las voces-propias es el sistema Variacional de Bayes reportado en [92,93]. Inspirado en el trabajo pionero de Valente [100], que utiliza métodos probabilísticos para la agrupación de locutores e invoca técnicas Variacionales Bayesianas como un método de inferencia aproximada. Esto condujo a resultados superiores en la diarización de locutores en conversaciones telefónicas, en comparación con el enfoque dinámico [92]. Las formulaciones matemáticas completas detrás de este enfoque se pueden encontrar en [91].

3.1.4 Enfoques utilizando un decodificador HMM y el tiempo de retardo de llegada (iv, v)

La segmentación y la agrupación utilizando un decodificador HMM [76,37,73] ha demostrado ser un enfoque eficaz, especialmente cuando la aplicación final es un sistema de *reconocimiento de habla*. Un HMM se utiliza típicamente para decodificar el flujo de audio en clases acústicas, permitiendo que el reconocimiento pueda llevarse a cabo usando modelos específicos a esa clase acústica en particular.

Un ejemplo del HMM se presenta en [76], el cual parte de desconocer el número de locutores y utiliza el enfoque top-down, implicando que el HMM comience con un estado que "representa al locutor" y progrese en busca de más estados. El proceso de decodificación se repite hasta que se considere que el número de locutores es óptimo, durante la decodificación en cada iteración, la probabilidad de emisión del HMM por cada trama acústica es almacenada. Luego las tramas que presenten la mayor probabilidad de ocurrencia dentro de cualquier estado son tomadas para crear un nuevo estado del HMM. Este proceso de adicionar nuevos estados (cada estado representa un locutor en el audio) se repite hasta que no se encuentren más tramas con alta probabilidad de ocurrencia.

Otro ejemplo, pero utilizando el enfoque bottom-up, es presentado en [73], en este caso también se desconoce el número de locutores a encontrar por el HMM. Se comienza por asumir un número de locutores $N_{locutores}$ obtenido por el algoritmo K-medias (del inglés K-means), el cual crea una primera agrupación de las clases acústicas que servirán de inicialización al decodificador. Cada grupo inicial (clase acústica) se convierte, mediante un entrenamiento, en un estado del HMM. Luego comienza el proceso de decodificación y las dos clases más similares (GLR como medida de similitud) son mezcladas, repitiendo

[8] Es una técnica utilizada para reducir el número de variables que se encuentran en la voz, a factores más influyentes según describió Najim [95]

el proceso utilizando el algoritmo de Viterbi hasta encontrar el número final de locutores. Este sistema tiende a obtener una cantidad de locutores mayor que la real.

El criterio que crea *la agrupación utilizando tiempo de retardo de llegada*, solamente tiene utilidad en la diarización de locutores en grabaciones microfónicas a partir de arreglos de micrófonos en diferentes lugares de una habitación (MDM del inglés múltiples micrófonos distantes). La idea consiste en tener en cuenta el retardo de cada voz de cada locutor al llegar a cada uno de los micrófonos, esto es posible debido a que la velocidad del sonido es constante y la distancia espacial entre cada locutor y cada micrófono es diferente; provocando diferencia en los tiempos de llegada.

Obsérvese que:

- Ambos ejemplos y en general los enfoques que utilizan las HMM necesita la expresión de audio completa para comenzar el proceso de diarización.
- El criterio que utiliza el tiempo de retardo de llegada para realizar la agrupación solo es lógico en grabaciones microfónicas donde estén presentes múltiples micrófonos espacialmente separados.

3.2 Conclusiones: ventajas y desventajas

En esta etapa de la diarización de locutores, los algoritmos de agrupación dependen principalmente del la cantidad locutores presentes en la señal de audio, la duración de la misma, el ruido ambiental y si la tarea tiene que realizarse en tiempo real o no. En el presente trabajo el objetivo consiste en obtener una respuesta mientras se produce el flujo de audio, manteniendo o mejorando la eficacia de los métodos actuales en el estado del arte. Por tal motivo nos centraremos en las ventajas y desventajas que tienen los algoritmos para satisfacer esta necesidad.

El enfoque *i* utiliza la cuantificación vectorial para realizar el proceso de agrupación, lo que implica la necesidad de tener una información previa de los locutores presentes en el audio para obtener los "codebooks" de cada grupo, esta restricción deja fuera del dominio de trabajo a este enfoque. Los enfoques (iv) y (v) presentados en la subsección 3.1.4 combinan la etapa de la segmentación y el agrupamiento, lo cual podría parecer beneficioso pero en el caso del *iv* necesitan la expresión de audio completa para realizar el proceso y en el caso del *v* queda fuera del dominio del presente trabajo por depender de múltiples micrófonos y ser imposible su utilizad sobre las conversaciones telefónicas. Esto nos conlleva a centrarnos sobre los enfoques *ii* y *iii*.

Ventajas del *agrupamiento aglomerativo jerárquico*:

- Es la técnica, siguiendo el enfoque Bottom-Up, de referencia en el estado del arte en los últimos 10 años con mejor compromiso entre eficacia y eficiencia.
- En una primera impresión, podríamos pensar que estos algoritmos no son utilizables por la necesidad de comenzar el proceso sobre todos los segmentos de audio, esto es cierto, pero estas técnicas permiten modificaciones para crear grupos iniciales y después, de forma incremental ir incorporando nuevos objetos o creando nuevos grupos [101]. Estas modificaciones han sido utilizadas en el área de la minería de datos.
- No asumen ningún conocimiento previo del número de los locutores, sus identidades, o características de la señal. Esto implica la posibilidad de su utilidad para *agrupar en tiempo real* utilizando las variantes en [101].
- El agrupamiento puede ser realizado con diferentes rasgos acústicos, para la fusión posterior de los resultados, ya que los diferentes rasgos pueden complementarse en diferentes contextos [45,59,60].

Además, se podría realizar el agrupamiento utilizando varias métricas y/o varios esquemas de clasificación, y luego fusionar los resultados individuales. En general, la fusión ofrece potencialmente muchas ventajas, como el aumento del rendimiento y robustez de un sistema [61].

Desventajas del *agrupamiento aglomerativo jerárquico*:

- En el caso de la medida BIC, la mayoría de las variantes de los algoritmos utilizan un término de penalización que depende del contexto de la aplicación, este término, en general, es difícil de estimar, de tal manera que proporcione un rendimiento estable frente a diferentes contextos. Existen varias modificaciones para adaptar o eliminar el mismo [31,53,13,14]
- En cuanto al costo computacional, la aplicación completa del algoritmo es computacionalmente costosa, alcanzando el orden de $O(N^2)$ [13]. El alto costo computacional ha motivado a los investigadores en el área a emplear heurísticas para resolver el problema.
- El rendimiento o eficacia del algoritmo para el agrupamiento es menor que cuando se utiliza el agrupamiento basado en Modelos [92,93], tanto para el caso del BIC como para el GLR.
- Se debe tener en cuenta que el supuesto Gaussiano de las muestras acústicas no siempre es correcto [62].

Ventajas del *agrupación basada en modelos*:

- Los sistemas actuales del estado del arte más eficaces para el reconocimiento y la diarización del locutor están basado en el enfoque de los Modelos de Mezclas Gaussiano GMM, trabajando sobre el espacio de variabilidad total (i-vectores) [92,93,84], descrito en la subsección (3.1.3).
- Este nuevo paradigma permite llevar a cabo una diarización en línea, donde el flujo de audio entrante se trata como una corriente de segmentos, realizando tanto la segmentación como la agrupación de **manera causal o dinámica**.
- Este enfoque es capaz de utilizar conocimiento previo altamente informativo sobre el espacio de la población. Notar que esta información previa no pertenece a los locutores que interactúan en la señal de audio.
- Con una cantidad de componentes Gaussianos menor que el rango de la matriz T , $(K < W)$ se logra un bajo costo computacional en la estimación de los modelos de los locutores, esto permite hacerle frente a la transmisión del audio.
- Los i-vectores pueden ser creados con diferentes tipos de rasgos acústicos, para la fusión posterior de los resultados. Los diferentes tipos de rasgos presentan informaciones diferentes que pueden complementarse en diferentes contextos [95].

Desventajas de la *agrupación basada en modelos*:

- Para obtener conocimiento previo sobre la población de locutores se necesitan grandes bases de datos.
- El enfoque tiene sus bases en un marco estadístico, donde la influencia de una información específica se recopila principalmente por la frecuencia de esta información. Todos los enfoques parten del UBM, que contiene la distribución de las clases acústicas de una población, o sea las características particulares de un locutor que no sean frecuentes dentro de la población tendrán una pobre representación dentro del UBM.
- Se asume que los i-vectores obtenidos del espacio de variabilidad total T sigan una distribución Normal, lo cual no se logra, si se observa que la matriz de covarianza total está lejos de ser la matriz identidad.

4 Protocolo evaluación y sistemas de referencia en el estado del arte

Con el fin de evaluar y comparar el rendimiento de diferentes algoritmos que se utilizan en los sistemas de diarización del locutor, son requeridas adecuadas métricas de evaluación del desempeño y protocolos de prueba. Además es esencial contar con bases de datos apropiadas que contengan señales de audio etiquetadas.

En este capítulo se describen los indicadores estandarizados para la evaluación del desempeño y protocolos de pruebas, tal como se utiliza en la tarea de diarización del locutor en el Instituto Nacional de Estándares y Tecnología (NIST) [102]. Además se resumen las principales plataformas de código abierto de referencias en la actualidad y las bases de datos utilizadas. Como se indica en la introducción, el alcance de este trabajo de investigación se limita a la segmentación y la agrupación del locutor en el dominio de las comunicaciones telefónicas con múltiples locutores. Por tanto los sistemas y las bases de datos también se restringen a este género.

4.1 Métrica para la evaluación del desempeño

A partir del estado del arte descrito en este trabajo, la futura investigación se centrará en mejorar el rendimiento de la diarización del locutor. Para lograr esta meta, todos los algoritmos que se desarrollen a lo largo de la investigación serán evaluados utilizando la medida Tasa de Error de Diarización (DER del inglés Diarization Error Rate), como se define en [102].

4.1.1 Tasa de error de diarización

La respuesta de un sistema de diarización del locutor es un conjunto de segmentos por cada locutor hipotético, cada uno de los cuales contiene una etiqueta de la identidad (ID) del locutor y los correspondientes tiempos de inicio y de fin de cada segmento. Este resultado es comparado con la información real de cada flujo de audio dando origen al error cometido.

El DER es una medida basada en el tiempo, que se puede interpretar como la proporción de la cantidad total de tiempo de voz útil que no esté atribuida al locutor correcto, teniendo en cuenta los errores de detección del habla. Se calcula a través de una óptima asignación uno a uno de los ID de los locutores de referencia a los ID de los locutores hipótesis, a fin de maximizar la superposición total entre la referencia y los locutores de hipótesis. El tiempo total de error en la diarización del locutor viene dado por la suma de los tiempos de las detecciones omitidas (locutor de referencia, pero no en hipótesis), las falsas alarmas (locutor en la hipótesis, pero no en referencia) y los errores del locutor (el locutor de referencia asignado no es el mismo que el locutor hipotético).

- Detecciones omitidas: Existe una referencia del locutor hablando en el segmento, pero el sistema no encontró al locutor o etiqueto al segmento como una región de no voz. Este último error proviene del módulo de pre-procesamiento, donde se declaran las regiones de voz y no voz, en esta etapa se pueden clasificar algunos segmentos de locutores auténticos como segmentos vacíos, implicando que estas regiones se omitan en la salida final.
- Falsas alarmas: No existe una referencia del locutor hablando en el segmento, pero el sistema comete un error asignando un locutor a esta región. Este error puede ser provocado por el algoritmo de agrupamiento o por el módulo de pre-procesamiento, en el ultimo caso el módulo puede mantener algunas regiones de silencio implicando que en las siguientes etapas se asuman como regiones con voz.

- Errores del locutor: Es el caso que tanto la información de referencia como el sistema indican que existe un locutor en la región, sin embargo, el sistema asigna una etiqueta del locutor erróneo al segmento.

Como se indica en [102], el DER puede expresarse formalmente como:

$$Error_{Diari} = \frac{\sum_{todossegs} \left\{ dur(seg) \cdot \left(max\Big(N_{Ref}(seg), N_{Sis}(seg)\Big) \; - \; N_{Correcto}(seg) \right) \right\}}{\sum_{todossegs} \left\{ dur(seg) \cdot N_{Ref}(seg) \right\}}, \qquad (18)$$

donde el flujo de audio se divide en segmentos contiguos por los puntos de cambio del locutor y para cada segmento, *seg*,

$dur(seg) =$ la duración del segmento *seg*,
$N_{Ref}(seg) =$ el número del locutores de referencia que hablan en el *seg*,
$N_{Sis}(seg) =$ el número del locutores reconocidos por el sistema que hablan en el *seg*,
$N_{Correcto}(seg) =$ el número de locutores de referencias que hablan en el *seg* correspondientes con los reconocidos por el sistema que también están hablando en *seg*.

4.2 Sistemas de referencia en el estado del arte

Es de señalar que en los últimos tiempos las evaluaciones de los sistemas de diarización han estado dirigidas, en su mayor parte, a la radiodifusión o la transcripción de audio [7]. Los últimos trabajos encontrados sobre la diarización en conversaciones telefónicas están restringidos a dos locutores por conversaciones [85,3,103], en este caso han estado dirigidos a la mejora de la etapa de segmentación.

Varias plataformas para la diarización del locutor están disponibles en la web, distribuidas bajo licencias de código abierto. Algunas de ellas son:

CMU: Grupo de herramientas para la segmentación, fue publicado en el 1997 [104]. Fue desarrollado durante la primera campaña de evaluación de difusión de noticias NIST, se centro específicamente en la tarea de la diarización para el reconocimiento automático del habla.

AudioSeg: Grupo de herramientas desarrolladas por IRISA [105] durante la campaña ESTER en el 2005. Incluye un detector de actividad de voz, algoritmos de segmentación BIC / GLR o KL2 y algoritmos de agrupación, así como un decodificador Viterbi. Es de señalar que el algoritmo CLR de agrupación no está desarrollado en esta plataforma.

ALIZE: Es una plataforma de reconocimiento del locutor que incluye herramientas de diarización del locutor [55], también se encuentra actualizada. La diarización es basada en los métodos E-HMM [33] (decodificador HMM iv), donde la segmentación y la agrupación se realizan de forma iterativa o conjuntamente. Su eficacia es mejor cuando se trata de diarización en reuniones y conversaciones telefónicas que en la difusión de noticias.

DiarTk: Un grupo de herramientas de código abierto para la diarización de múltiples flujos de voz de locutores y su aplicación a las grabaciones de reuniones. Esta plataforma publicada por IDIAP, bajo la licencia GPL, fue desarrollada para facilitar la investigación en la diarización del locutor en el área de las reuniones grabadas utilizando múltiples micrófonos distantes. Al contrario de otras plataformas o

sistemas de diarización, DiarTk esta diseñado explícitamente para manejar un número arbitrario de rasgos con diferentes estadísticas, manteniendo una alta eficiencia computacional. Esta característica favorece el estudio de nuevos rasgos acústicos en el área de la diarización [106].

DiarTK está programado en C++ y diseñado como se explica a continuación:

- Presenta un código simple y esta encapsulado en módulos.
- Capaz de manejar un número arbitrario de flujos de rasgos con mucha diferencia respecto a la estadísticas.
- Limita la complejidad computacional del sistema de diarización permitiendo realizar un procesamiento eficiente.
- Reproduce los resultados en el estado del arte en bases de datos de referencias, NIST.

LIUM_SPKDIARIZATION: Un grupo de herramientas de código abierto para la diarización de múltiples flujos de voz de locutores. Esta plataforma incluye métodos para el agrupamiento aglomerativo jerárquico utilizando medidas como BIC y CLR, puede ser utilizado tanto en radiodifusión como en conversaciones telefónicas. Su configuración por defecto, se orienta hacia la diarización del locutor en la emisión de noticias. Este grupo de herramientas proporciona algoritmos para la segmentación, el agrupamiento de locutores, la decodificación y entrenamiento de los modelos. Según lo descrito en [107] la plataforma LIUM permite el desarrollo, de manera sencilla, de un sistema para la diarización del locutor en conversaciones telefónicas y otros sistemas de diarización en áreas especificas.

LIUM_SPKDIARIZATION está programando en Java, versión 1.6 y para su compilación se necesitan los siguientes 3 paquetes: paquete “gnu-getopt” para gestionar las líneas de comandos con opciones largas, paquete “Lapack” para manejar operaciones con matrices y el “Sphinx4” para calcular los rasgos acústicos en escala Mel (MFCC). Sus característica generales son:

- Se desarrolló con los métodos de referencia en el estado del arte, año 2010, se mantiene actualizado [56].
- Fácil generalización a distintas dominios en la diarización.
- Alta eficacia frente a los datos de referencia en las evaluaciones NIST [56].

4.3 Bases de datos

Dado que en los sistemas de diarización del locutor las características de los dominios de aplicación son diferentes, cualquier enfoque específico de la aplicación debe ser evaluado en una base de datos correspondiente. Por lo tanto, las bases de datos de evaluación en la diarización del locutor se dividen a tipos específicos de aplicación. Los conjuntos de datos de evaluación más comunes en el dominio de las conversaciones telefónicas son presentados a continuación.

Las competencias NIST son la vanguardia a nivel mundial generando cuerpos de bases de datos, en el caso del procesamiento del habla han organizado múltiples evaluaciones en muchos aspectos a través de los años. La tarea de diarización del locutor se evaluó por primera vez en el año 2000, enfocando las competencias sobre las conversaciones telefónicas (2000, 2001 y 2002), luego la emisión de noticias (2002, 2003 y 2004) y más recientemente las reuniones (2002, 2004, 2005 y 2006). Una característica común de estas evaluaciones es que el único conocimiento a priori a disposición de los participantes es el conocimiento acerca de la fuente o escenario de grabación (por ejemplo, reuniones de conferencias, charlas, o coffee breaks), el idioma (Inglés), y el formatos de los archivos de entrada y de salida. Los participantes en la evaluación podrán utilizar datos externos para crear los modelos de la población y / o

con fines de normalización, pero no existe información a priori en relación con los locutores que participan en las grabaciones (evaluaciones de Ricas Transcripciones (NIST RT), 2006).

Inicialmente, en 2002, la evaluación de segmentación del locutor se llevó a cabo dentro de la evaluación de reconocimiento de locutor (SRE-02). Esta rutina fue cambiado del 2004 al 2006, cuando la diarización del locutor fue una parte de la evaluación de transcripción (RT04s, RT05s y RT06s). Algunos ejemplos de las bases de datos de evaluación NIST RT son los siguientes: NIST 2002 *Rich Transcripción* emisión de noticias (BN) y conversaciones Telefónicas de Voz (CTS) (NIST RT'02). Los datos de la base CTS están compuestos por 60 extractos de voz a partir de 60 conversaciones diferentes: 20 extractos de voz de Switchboard−1, 20 extractos de voz de Switchboard−2 y 20 extractos de voz de Switchboard Celular−2.

Otras bases de voces que pueden ser utilizadas en la diarización del locutor en conversaciones telefónicas son las bases NIST [88,89,4,90,9] dedicadas al reconocimiento automático del locutor, con la restricción que solo intervienen dos locutores por señales.

4.4 Conclusiones

A partir del rendimiento mostrado por el plataforma de diarización del locutor LIUM_SPKDIARIZATION en las evaluaciones NIST [56], su nivel de actualización respecto a los algoritmos en el estado del arte y su posible generalización al área de la diarización en conversaciones telefónicas, se considerara como sistema de referencia para comparar con el rendimiento de la futura investigación.

Se utilizará Switchboard como Bases de Datos de referencia, así como las bases NIST de reconocimiento del locutor para evaluar el rendimiento del sistema a desarrollar.

5 Conclusiones y trabajo futuro

Este trabajo ofrece una visión general del estado del arte en los sistemas diarización de locutores y menciona varios retos que deben abordarse en los próximos años. Por ejemplo, la Diarización del Locutor aún no es lo suficientemente madura para que los métodos desarrollados puedan ser fácilmente portados a través de diferentes dominios. Además, muchos y diversos algoritmos de segmentación y agrupamiento se han desarrollado en la literatura para resolver el problema de la diarización del locutor y su popularidad está dada más por los resultados experimentales que por elementos teóricos y entornos de aplicabilidad. Como se aprecia en el documento, son pocos los desarrollos en el área que permiten solucionar el desafío de dar respuesta en "tiempo real", lo que es uno de los aspectos planteados dentro del objetivo principal.

Después del estudio realizado arribamos a que podemos atacar el problema de la diarización dinámica del locutor de dos forma diferentes. Podemos desarrollar algoritmos para solucionar cada una de las dos etapas por separado o podemos desarrollar algoritmos que unifique las dos etapas.

En el caso de dividir el problema en dos etapas, arribamos a que los algoritmos utilizables para la etapa de segmentación son los basados en métricas como el BIC, (véase sección 2.2). Para luego utilizar en la etapa de agrupación este mismo enfoque, algoritmo BIC o GLR (véase sección 3.1.2), pero en su variante incremental [101].

Para el caso de realizar la diarización como un todo respecto a las etapas, los algoritmos utilizables son los basados en modelos, más preciso, el enfoque i-vectores. El cual tiene la capacidad del espacio variabilidad total para extraer características específica de los locutores en segmentos de corta duración [94,95], causante de la alta eficacia obtenidas en las competencias internacionales, sección 3.1.3.

Como trabajo futuro nos proponemos varias tareas encaminadas a satisfacer el objetivo principal.

- Desarrollar una línea base utilizando los algoritmos de diarización (LIUM) del locutor que se basan en modelos (enfoque i-vectores), con la característica que procesen el flujo de audio completo, o sea una línea base no dinámica. Esta nos permitirá compara los resultados de los desarrollos propios a medida que se procesa el flujo de audio.
- Mantener un seguimiento sobre las Redes Neuronales Profundas debido a su repercusión en el procesamiento de la voz, y sus posibles prestaciones en nuestra área.
- Realizar una estudio para compensar la variabilidad respecto a la duración sobre el enfoque i-vectores, que resulte en el desarrollo de un método que impacte en el rendimiento de la diarización del locutor.
- Desarrollar nuestro sistema de diarización dinámica del locutor, unidos a los resultados de la tesis en el reconocimiento del locutor independiente del texto [108].
- Continuar el estudio de la temática en posibles soluciones que surjan en el estado del arte.

Referencias bibliográficas

1. Gish, H., Siu, M.H., Rohlicek, R.: Segregation of speakers for speech recognition and speaker identification. In: icassp, IEEE (1991) 873–876
2. Deng, J., Zheng, T.F., Wu, W.: UBM based speaker segmentation and clustering for 2-speaker detection. In: Chinese Spoken Language Processing, 5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006, Proceedings. (2006) 116–125
3. Sholokhov, A., Pekhovsky, T., Kudashev, O., Shulipa, A., Kinnunen, T.: Bayesian analysis of similarity matrices for speaker diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014. (2014) 106–110
4. Burget, L., Fapso, M., Hubeika, V., Glembek, O., Karafiát, M., Kockmann, M., Matejka, P., Schwarz, P., Cernocký, J.: BUT system for NIST 2008 speaker recognition evaluation,. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009. (2009) 2335–2338
5. Matsoukas, S., Prasad, R., Laxminarayan, S., Xiang, B., Nguyen, L., Schwartz, R.M.: The 2004 BBN 1xrt recognition systems for english broadcast news and conversational telephone speech. In: INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005. (2005) 1641–1644
6. Matsoukas, S., Gauvain, J., Adda, G., Colthurst, T., Kao, C., Kimball, O., Lamel, L., Lefevre, F., Ma, J.Z., Makhoul, J., Nguyen, L., Prasad, R., Schwartz, R.M., Schwenk, H., Xiang, B.: Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. IEEE Transactions on Audio, Speech & Language Processing **14**(5) (2006) 1541–1556
7. Moattar, M.H., Homayounpour, M.M.: A review on speaker diarization systems and approaches. Speech Communication **54**(10) (2012) 1065–1103
8. Reynolds, D.A., Torres-carrasquillo, P.: The mit lincoln laboratory rt-04f diarization systems: Applications to broadcast audio and telephone conversations. In: NIST Rich Transcription Workshop November 2004.
9. Khoury, E., El Shafey, L., Marcel, S.: The idiap speaker recognition evaluation system at nist sre. In: NIST Speaker Recognition Conference, NIST, 2012 (dec 2012)
10. Barras, C., Zhu, X., Meignier, S., Gauvain, J.: Multistage speaker diarization of broadcast news. IEEE Transactions on Audio, Speech & Language Processing **14**(5) (2006) 1505–1512
11. Tranter, S., Yu, K., Evermann, G., Woodland, P.C.: Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004. (2004) 753–756
12. Sinha, R., Tranter, S.E., Gales, M.J.F., Woodland, P.C.: The cambridge university march 2005 speaker diarization system. In: INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005. (2005) 2437–2440
13. Kotti, M., Moschou, V., Kotropoulos, C.: Speaker segmentation and clustering. Signal Processing **88**(5) (2008) 1091–1124
14. Miró, X.A., Bozonnet, S., Evans, N.W.D., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech & Language Processing **20**(2) (2012) 356–370
15. Ajmera, J., McCowan, I.A., Bourlard, H.: Robust Audio Segmentation. PhD thesis, Acole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland (6 2004) thesis (IDIAP-RR 04-35).

16. Kemp, T., Schmidt, M., Westphal, M., Waibel, A.: Strategies for automatic segmentation of audio data. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey. (2000) 1423–1426
17. Pérez-Freire, L., García-Mateo, C.: A multimedia approach for audio segmentation in TV broadcast news. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004. (2004) 369–372
18. Ohtsuki, K., Nguyen, L.: Incremental language modeling for automatic transcription of broadcast news. IEICE Transactions **90-D**(2) (2007) 526–532
19. Wegmann, S., Zhan, P., Carp, I., Newman, M., Yamron, J., Gillick, L.: Dragon systems' 1998 broadcast news transcription system. In: Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999. (1999)
20. Hauptmann, A.G., Witbrock, M.J.: Story segmentation and detection of commercials in broadcast news video. In: Proceedings of the IEEE Forum on Reasearch and Technology Advances in Digital Libraries, IEEE ADL '98, Santa Barbara, California, USA, April 22-24, 1998. (1998) 168–179
21. Rosenberg, A.E., Magrin-chagnolleau, I., Parthasarathy, S., Huang, Q.: Speaker detection in broadcast speech databases. In: Proceedings of ICSLP 98. (1998) 202–205
22. Gauvain, J., Lamel, L., Adda, G.: Partitioning and transcription of broadcast news data. In: The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998. (1998)
23. Bakis, R., Chen, S.S., Gopalakrishnan, P.S., Gopinath, R., Maes, S.H., Polymenakos, L.: Transcription of broadcast news-system robustness issues and adaptation techniques. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997. (1997) 711–714
24. Meignier, S., Moraru, D., Fredouille, C., Besacier, L., Bonastre, J.: Benefits of prior acoustic segmentation for automatic speaker segmentation. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004. (2004) 397–400
25. Lu, L., Zhang, H., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. Multimedia Syst. **8**(6) (2003) 482–492
26. Chen, S.S., Gopalakrishnan, P.S.: Clustering via the bayesian information criterion with applications in speech recognition. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998. (1998) 645–648
27. Li, R., Schultz, T., Jin, Q.: Improving speaker segmentation via speaker identification and text segmentation. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009. (2009) 904–907
28. van Leeuwen, D.A., Huijbregts, M.: The AMI speaker diarization system for NIST rt06s meeting data. In: Machine Learning for Multimodal Interaction, Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers. (2006) 371–384
29. Bonastre, J., Delacourt, P., Fredouille, C., Merlin, T., Wellekens, C.: A speaker tracking system based on speaker turn detection for NIST evaluation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey. (2000) 1177–1180
30. Lu, L., Zhang, H.: Real-time unsupervised speaker change detection. In: 16th International Conference on Pattern Recognition, ICPR 2002, Quebec, Canada, August 11-15, 2002. (2002) 358–361
31. Vandecatseye, A., Martens, J., Neto, J.P., Meinedo, H., García-Mateo, C., Dieguez-Tirado, J., Mihelic, F., Zibert, J., Nouza, J., David, P., Pleva, M., Cizmar, A., Papageorgiou, H., Alexandris, C.: The COST278 pan-european broadcast news database. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. (2004)
32. Kim, H., Ertelt, D., Sikora, T.: Hybrid speaker-based segmentation system using model-level clustering. In: 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005. (2005) 745–748
33. Meignier, S., Moraru, D., Fredouille, C., Bonastre, J., Besacier, L.: Step-by-step and integrated approaches in broadcast news speaker diarization. Computer Speech & Language **20**(2-3) (2006) 303–330
34. Wu, T., Lu, L., Chen, K., Zhang, H.: Ubm-based real-time speaker segmentation for broadcasting news. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003. (2003) 193–196
35. Wu, T., Lu, L., Chen, K., Zhang, H.: Universal background models for real-time speaker change detection. In: MMM. (2003) 135–149
36. Collet, M., Charlet, D., Bimbot, F.: A correlation metric for speaker tracking using anchor models. In: 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005. (2005) 713–716

37. Ajmera, J., McCowan, I., Bourlard, H.: Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Communication **40**(3) (2003) 351–363
38. Arias, J.A., Pinquier, J., André-Obrecht, R.: Evaluation of classification techniques for audio indexing. In Dutagacy, H., Sankur, B., Akgul, T., eds.: 13th European Conf. on Signal Processing (EUSIPCO'2005), Antalya, Turkey, Suvisoft Oy Ltd (septembre 2005)
39. Mesgarani, N., Shamma, S.A., Slaney, M.: Speech discrimination based on multiscale spectro-temporal modulations. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004. (2004) 601–604
40. Kwon, S., Narayanan, S.: Unsupervised speaker indexing using generic models. IEEE Transactions on Speech and Audio Processing **13**(5-2) (2005) 1004–1013
41. Delacourt, P., Wellekens, C.: DISTBIC: A speaker-based segmentation for audio data indexing. Speech Communication **32**(1-2) (2000) 111–126
42. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: Second-order statistical measures for text-independent speaker identification. Speech Communication **17**(1-2) (1995) 177–192
43. Zhou, B., Hansen, J.H.L.: Efficient audio stream segmentation via the combined t^2 statistic and bayesian information criterion. IEEE Transactions on Speech and Audio Processing **13**(4) (2005) 467–474
44. Huang, R., Hansen, J.H.L.: Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. IEEE Transactions on Audio, Speech & Language Processing **14**(3) (2006) 907–919
45. Kotti, M., Benetos, E., Kotropoulos, C.: Automatic speaker change detection with the bayesian information criterion using MPEG-7 features and a fusion scheme. In: International Symposium on Circuits and Systems (ISCAS 2006), 21-24 May 2006, Island of Kos, Greece. (2006)
46. Kotti, M., Martins, L.P.M., Benetos, E., Cardoso, J.S., Kotropoulos, C.: Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches. In: Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada. (2006) 1101–1104
47. Tritschler, A., Gopinath, R.A.: Improved speaker segmentation and segments clustering using the bayesian information criterion. In: Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999. (1999)
48. Cheng, S., Wang, H.: A sequential metric-based audio segmentation method via the bayesian information criterion. In: 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003. (2003)
49. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the BIC. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003. (2003) 537–540
50. Cettolo, M., Vescovi, M., Rizzi, R.: Evaluation of bic-based algorithms for audio segmentation. Computer Speech & Language **19**(2) (2005) 147–170
51. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing **28**(4) (1980) 357–366
52. Campbell, N.A.: Robust procedures in multivariate analysis i: Robust covariance estimation. Applied Statistics **29**(3) (1980) 231–237
53. Ajmera, J., McCowan, I., Bourlard, H.: Robust speaker change detection. Signal Processing Letters, IEEE **11**(8) (2004) 649–651
54. Wang, H., Cheng, S.: METRIC-SEQDAC: a hybrid approach for audio segmentation. In: INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004. (2004)
55. Bonastre, J., Wils, F., Meignier, S.: Alize, a free toolkit for speaker recognition,. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005. (2005) 737–740
56. Rouvier, M., Dupuy, G., Gay, P., el Khoury, E., Merlin, T., Meignier, S.: An open-source state-of-the-art toolbox for broadcast news diarization. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013. (2013) 1477–1481
57. Gupta, V., Kenny, P., Ouellet, P., Boulianne, G., Dumouchel, P.: Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations. IEEE Signal Process. Lett. **14**(12) (2007) 1040–1043
58. el Khoury, E., Sénac, C., Pinquier, J.: Improved speaker diarization system for meetings. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan. (2009) 4097–4100
59. Lu, L., Zhang, H.: Speaker change detection and tracking in real-time news broadcasting analysis. In: Proceedings of the 10th ACM International Conference on Multimedia 2002, Juan les Pins, France, December 1-6, 2002. (2002) 602–610
60. Lu, L., Zhang, H.: Unsupervised speaker segmentation and tracking in real-time audio content analysis. Multimedia Syst. **10**(4) (2005) 332–343

61. Zhu, Y., Li, X.R.: Unified fusion rules for multisensor multihypothesis network decision systems. IEEE Transactions on Systems, Man, and Cybernetics, Part A **33**(4) (2003) 502–513
62. Almpanidis, G., Kotropoulos, C.: Phonemic segmentation using the generalised gamma distribution and small sample bayesian information criterion. Speech Communication **50**(1) (2008) 38–55
63. Pfitzner, D., Leibbrandt, R., Powers, D.M.W.: Characterization and evaluation of similarity measures for pairs of clusterings. Knowl. Inf. Syst. **19**(3) (2009) 361–394
64. Mori, K., Nakagawa, S.: Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings. (2001) 413–416
65. Akita, Y., Kawahara, T.: Unsupervised speaker indexing using anchor models and automatic transcription of discussions. In: 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003. (2003)
66. Rodríguez, L.J., Torres, I.: A speaker clustering algorithm for fast speaker adaptation in continuous speech recognition. In: Text, Speech and Dialogue, 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings. (2004) 433–440
67. Haubold, A., Kender, J.R.: Accommodating sample size effect on similarity measures in speaker clustering. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, June 23-26 2008, Hannover, Germany. (2008) 1525–1528
68. Andrew, A.M.: *Statistical Pattern Recognition,* by andrew webb, arnold, london (cambridge university press, new york, for usa), 1999, xviii+454 pp., ISBN 0-340-74164-3 (pbk, £29.99). Robotica **18**(2) (2000) 219–223
69. Moh, Y., Nguyen, P., Junqua, J.: Towards domain independent speaker clustering. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003. (2003) 85–88
70. Barras, C., Zhu, X., Meignier, S., Gauvain, J.L.: Improving speaker diarization. In: IN PROC. FALL 2004 RICH TRANSCRIPTION WORKSHOP (RT-04. (2004)
71. Moraru, D., Meignier, S., Besacier, L., Bonastre, J., Magrin-Chagnolleau, I.: The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003. (2003) 89–92
72. Moraru, D., Meignier, S., Fredouille, C., Besacier, L., Bonastre, J.: The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004. (2004) 373–376
73. Ajmera, J., Wooters, C.: A robust speaker clustering algorithm. In: In Proc. IEEE Automatic Speech Recognition Understanding Workshop. (2003) 411–416
74. Tranter, S.E., Gales, M.J.F., Sinha, R., Umesh, S., Woodland, P.C.: The development of the cambridge university rt-04 diarization system. In: Fall 2004 Rich Transcription Workshop (RT-04)
75. Bozonnet, S., Evans, N.W.D., Fredouille, C.: The lia-eurecom rt'09 speaker diarization system: Enhancements in speaker modelling and cluster purification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA. (2010) 4958–4961
76. Meignier, S., Bonastre, J., Igounet, S.: E-HMM approach for learning and adapting sound models for speaker indexing. In: 2001: A Speaker Odyssey - The Speaker Recognition Workshop, Crete, Greece, June 18-22, 2001. (2001) 175–180
77. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech & Language Processing **14**(5) (2006) 1557–1565
78. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2) (1978) 461–464
79. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Broadcast News Transcription and Understanding Workshop. (1998) 127–132
80. Nishida, M., Kawahara, T.: Speaker model selection based on the bayesian information criterion applied to unsupervised speaker indexing. IEEE Transactions on Speech and Audio Processing **13**(4) (2005) 583–592
81. Fauve, B.G.B., Matrouf, D., Scheffer, N., Bonastre, J., Mason, J.S.D.: State-of-the-art performance in text-independent speaker verification through open-source software. IEEE Transactions on Audio, Speech & Language Processing **15**(7) (2007) 1960–1968
82. Kenny, P.: Joint factor analysis of speaker and session variability: Theory and algorithms,. Technical report, Montreal, CRIM (2005)
83. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. IEEE Transactions on Audio, Speech & Language Processing **16**(5) (2008) 980–988
84. Senoussaoui, M., Kenny, P., Dumouchel, P., Dehak, N.: New cosine similarity scorings to implement gender-independent speaker verification. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013. (2013) 2773–2777
85. Ilya, S., Neta, R., Irit, O., Itshak, L.: Clustering short push-to-talk segments. In: Sixteenth Annual Conference of the International Speech Communication Association. (2015)

86. Madikeri, S., Himawan, I., Motlicek, P., Ferras, M.: Integrating online i-vector extractor with information bottleneck based speaker diarization system. In: Proceedings of Interspeech 2015. (2015)
87. Delgado, H., Anguera, X., Fredouille, C., Serrano, J.: Novel clustering selection criterion for fast binary key speaker diarization. In: Proc. INTERSPEECH. (2015)
88. Przybocki, M.A., Martin, A.F.: NIST speaker recognition evaluation chronicles. In: ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3, 2004. (2004) 15–22
89. Przybocki, M., Martin, A., Le, A.: Nist speaker recognition evaluation chronicles - part 2,. In: Odyssey 2006: The Speaker and Language Recognition Workshop, 2006. (June 2006) 1–6
90. Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S.S., Shriberg, E., Stolcke, A.: The SRI NIST 2010 speaker recognition evaluation system,. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic. (2011) 5292–5295
91. Kenny, P.: Bayesian analysis of speaker diarization with eigenvoice priors. (2008)
92. Kenny, P., Reynolds, D.A., Castaldo, F.: Diarization of telephone conversations using factor analysis. J. Sel. Topics Signal Processing **4**(6) (2010) 1059–1070
93. Reynolds, D.A., Kenny, P., Castaldo, F.: A study of new approaches to speaker diarization. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009. (2009) 1047–1050
94. Shum, S., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In: Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010. (2010) 16
95. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech & Language Processing **19**(4) (2011) 788–798
96. Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D.A., Glass, J.R.: Exploiting intra-conversation variability for speaker diarization. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011. (2011) 945–948
97. Glembek, O., Burget, L., Matejka, P., Karafiát, M., Kenny, P.: Simplification and optimization of i-vector extraction,. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic. (2011) 4516–4519
98. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1-3) (2000) 19–41
99. Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., Vair, C.: Stream-based speaker segmentation using speaker factors and eigenvoices. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA. (2008) 4133–4136
100. Valente, F., Wellekens, C.: Variational bayesian methods for audio indexing. In: Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers. (2005) 307–319
101. Suárez, A.P., Pagola, J.E.M., Trinidad, J.F.M., y Jesús A. Carrasco Ochoa: Algoritmos jerárquicos y no jerárquicos para la construcción de grupos con traslape en contextos dinámicos. In: RT 019, Serie Gris, CENATAV. (2012) 356–370
102. Fiscus, J.: Fall 2004 rich transcription rt-04f evaluation plan. In: National Institute of Standards and Technology. (2004)
103. Zheng, R., Zhang, C., Zhang, S., Xu, B.: Variational bayes based i-vector for speaker diarization of telephone conversations. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014. (2014) 91–95
104. Siegler, M.A., Jain, U., Raj, B., Stern, R.M.: Automatic segmentation, classification and clustering of broadcast news audio. Proc. DARPA speech recognition workshop **1997** (1997)
105. Gravier, G., Betser, M., Ben, M.: Automatic segmentation, classification and clustering of broadcast news audio. release 1.2., IRISA, http://www.irisa.fr/metiss/accueil.html (2010)
106. Vijayasenan, D., Valente, F.: Diartk : An open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In: INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012. (2012) 2170–2173
107. Meignier, S., Merlin, T.: Lium spkdiarization: an open source toolkit for diarization. In: in CMU SPUD Workshop. (2010)
108. Hernández-Sierra, G., Calvo, J.R., Bonastre, J.F.: Métodos de representación y verificación del locutor con independencia del texto. CENATAV (2014)

RT_081, febrero 2016
Aprobado por el Consejo Científico CENATAV
Derechos Reservados © CENATAV 2016
Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba