Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 109-116 recibido 05-05-12 revisado 04-06-12 aceptado 07-06-12

Técnicas de post-procesado de resultados en un sistema de diarización de locutores

Post-processing techniques for a speaker diarization system

David Tavarez UPV/EHU Alda. Urquijo s/n 48013 Bilbao david@aholab.ehu.es

Eva Navas UPV/EHU Alda. Urquijo s/n 48013 Bilbao eva@aholab.ehu.es

Daniel Erro UPV/EHU Alda. Urquijo s/n 48013 Bilbao derro@aholab.ehu.es

Ibon Saratxaga UPV/EHU Alda. Urquijo s/n 48013 Bilbao ibon@aholab.ehu.es

Inma Hernaez UPV/EHU Alda. Urquijo s/n 48013 Bilbao inma@aholab.ehu.es

Resumen: Este artículo presenta las técnicas de postprocesado diseñadas para mejorar los resultados de un sistema de diarización de locutores. Se han propuesto tres técnicas de mejora: el refinado de la segmentación voz/no voz, la asimilación de los segmentos cortos y la fusión de los clusters del mismo locutor. Las técnicas se han implementado en un módulo que se aplica como etapa de postprocesado y que ha mejorado un 22.3 % el resultado del sistema base. El módulo se ha aplicado sin realizar ningún ajuste sobre otro sistema de diarización de arquitectura similar al sistema base con una mejora del 21 % y sobre uno con arquitectura muy diferente sin conseguirse mejoras. Asimismo se ha utilizado con otra base de datos y se ha conseguido mejorar el DER un 17 %. Esto demuestra la validez de las técnicas desarrolladas para la mejora de los resultados de la diarización.
Palabras clave: Diarización de locutores, segmentación, transcripción enriquecida

Abstract: This paper presents the post-processing techniques designed to improve the results of a speaker diarization system. Three different techniques are proposed: refinement of speech vs. non speech segmentation, assimilation of short speech segments and fusion of clusters from the same speaker. These techniques have been implemented in a post-processing module that improves the result of the baseline system by 22.3 %. The same module has been applied to another speaker diarization system with a similar architecture to that of the baseline system with a DER improvement of 21 % and to another one with a very different architecture where no improvement has been achieved. It has also been used with another database with an improvement of 17 %. These experiments prove the validity of the techniques developed.
Keywords: Speaker diarization, segmentation, rich transcription

1. Introducción

La diarización de audio consiste en dividir un flujo de audio en regiones homogéneas de acuerdo a sus fuentes de audio específicas (Cettolo, Vescovi, y Rizzi, 2005). Estas fuentes pueden incluir tanto el tipo de audio (voz, música, ruido de fondo, etc.), como la identidad del locutor y las características del canal (Reynolds y Torres-Carrasquillo, 2005). La diarización de locutor se puede considerar un subtipo de diarización de audio que consiste en segmentar de forma automática una grabación de audio en regiones homogéneas relativas a la identidad del locutor, sin información a priori sobre la identidad y número de locutores presentes (Tranter y Reynolds, 2006). Adicionalmente, es posible llegar a identificar cada uno de ellos si se dispone de la información necesaria. Para cumplir este cometido deben combinarse varios algoritmos con diferentes finalidades que, en la mayoría de los sistemas, suelen ejecutarse de forma secuencial, es decir, cada uno se aplica a la señal completa antes de empezar con la tarea siguiente (Anguera, 2006). Comúnmente estas tareas incluyen la detección de voz, la detección de cambios de turno entre locutores, la agrupación de locutores y la resegmentación de la señal de audio (Anguera et al., 2012).

ISSN 1135-5948

Para determinar de manera objetiva la validez de los algoritmos desarrollados se organizan campañas competitivas de evaluación, como la NIST Rich Transcription[1] y la de Albayzin (Zelenák, Schulz, y Hernando, 2010). En estas campañas distintos grupos de investigación prueban sus algoritmos sobre una base de datos común, lo que permite comparar el rendimiento de los mismos e identificar las técnicas más adecuadas para cada etapa del sistema.

El sistema de diarización de locutores desarrollado en el grupo fue presentado a la campaña de evaluación Albayzin 2010 (Luengo et al., 2010). El algoritmo aplicado se basa en una implementación eficiente de un detector de cambio de turno basado en BIC (Chen y Gopalakrishnan, 1998) que utiliza únicamente los segmentos sonoros de voz y una agrupación de locutores off-line mediante proceso de agrupación o clustering jerárquico acumulativo de abajo arriba (Tavarez et al., 2012). Este sistema obtuvo buenos resultados en la evaluación, aunque carecía de etapa de resegmentación de la señal de audio. Para diseñar un módulo de post-procesado de los resultados que mejorara el rendimiento global del sistema base se ha llevado a cabo un detallado estudio de los errores cometidos y se han planteado distintas estrategias para paliar cada uno de ellos. En este artículo se presenta este módulo de post-procesado, describiendo las estrategias implementadas y los resultados obtenidos.

La sección 2 del artículo presenta brevemente la base de datos con la que se ha desarrollado el sistema. En la sección 3 se describe el análisis de los errores cometidos y en la sección 4 las técnicas de post-procesado propuestas para mejorar los resultados. La sección 5 detalla los experimentos realizados y los resultados obtenidos y finalmente en la sección 6 se exponen las conclusiones del trabajo.

2. Base de datos

En la campaña de evaluación Albayzin 2010 se proporcionó una base de datos de voz en catalán de programas de noticias emitidos por el canal de televisión 3/24 (Zelenák, Schulz, y Hernando, 2010). Fue grabada por el grupo de investigación TALP de la UPC y etiquetada por Verbio Technologies. Consta de un total de 24 grabaciones o sesiones en las que el número de locutores que interviene varía desde 30 hasta 250. La base de datos contiene unas 87 horas de audio con la siguiente distribución: 37 % de voz limpia, 5 % de música, 15 % de voz con música de fondo, 40 % de voz con ruido de fondo y 3 % de otros, donde se engloba todo el material que no pertenece a las cuatro clases anteriores, incluyendo el ruido.

Para la campaña de evaluación la base de datos se dividió en dos partes: 16 sesiones para entrenamiento y desarrollo y las restantes 8 sesiones para pruebas.

3. Análisis de los errores

Los resultados obtenidos por el sistema base se han calculado de acuerdo con los criterios definidos por el NIST. La principal medida de error es el error total de diarización (DER, overall Diarization Error Rate) que está formado por la suma de los siguientes errores: voz de locutores no detectada (MST, Missed Speaker Time), segmentos de "no voz"marcados como voz de locutores (FAST, False Alarm Speaker Time) y errores de etiquetado incorrecto de locutores (SET, Speaker Error Time).

Con el fin de reducir el DER final, se ha llevado a cabo un análisis de los errores cometidos por el sistema base. Para ello, se han comparado las marcas de tiempo obtenidas con las marcas de referencia porporcionadas por la organización de Albayzin, para identificar la naturaleza de los distintos errores y diseñar estrategias que permitan tratar cada caso adecuadamente.

El DER obtenido por el sistema base en las señales de prueba es del 30.11 %, donde 2.8 % corresponde al MST, 2.2 % al FAST y 25.1 % al SET. Debido a la importancia del SET en los resultados obtenidos, se ha estudiado en detalle la aparición de errores que contribuyen en este aspecto, como son:

- segmentos cortos de un locutor que el proceso de agrupación de locutores asigna a otro diferente cuando la segmentación detecta un cambio que en realidad no existe.
- segmentos correspondientes a locutores con intervenciones de corta duración que el proceso de agrupación de locutores asigna a locutores ya existentes en lugar de crear un nuevo cluster.

[1] http://www.itl.nist.gov/iad/mig/tests/rt/

- distintas intervenciones de un mismo locutor que el proceso de clustering asigna a diferentes clusters, aumentando de forma errónea el número total de locutores identificados.

Tanto el MST como el FAST se deben a un mal funcionamiento del módulo de detección de voz. Debido a este incorrecto funcionamiento, segmentos de música que deberían ser eliminados pasan al módulo de detección de cambio de locutor y de agrupación de locutores y generalmente se les asigna una etiqueta de locutor nueva.

4. Módulo de postprocesado

Una vez analizados los errores cometidos por el sistema de diarización base se ha desarrollado un módulo de postprocesado con el fin de tratar cada uno de forma adecuada y reducir el DER final. A continuación se describe cada una de las partes que componen dicho módulo y que como puede verse en la Figura 1 consisten en un refinado de la segmentación voz/no voz, seguida de una etapa de asimilación de los segmentos de corta duración y por último de una fusión de los clusters correspondientes a un mismo locutor.

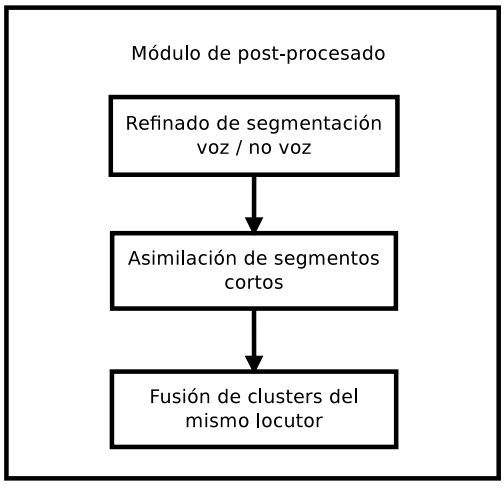

Figura 1: Diagrama del módulo de post-procesado

4.1. Etapa 1: Refinado de la segmentación voz/no voz

La primera etapa del módulo de postprocesado tiene como objetivo el refinado de los errores cometidos en el bloque de detección de voz. Para ello, en primer lugar, se entrena un modelo GMM (Modelo de Mezclas Gaussianas) para cada uno de los clusters obtenidos a la salida del sistema base y un modelo GMM para el silencio a partir de las marcas de referencia de las sesiones de entrenamiento designadas como "otros". A continuación, para cada segmento de voz marcado por el sistema base, se realiza una segmentación de Viterbi que incluye únicamente dos modelos GMM, el entrenado para el silencio y el del locutor marcado originalmente en dicho segmento. Por último, los silencios con una duración inferior a 750ms son eliminados.

De esta forma silencios, música, ruidos y el resto de posibles eventos acústicos detectados se marcan como "no voz", por lo que se consigue una reducción del FAST. Del mismo modo, intervenciones de otros locutores que han sido incluidas de forma errónea en cada uno de los diferentes clusters pueden ser marcadas en esta etapa como "no voz". Esto puede provocar un aumento del MST, sin embargo, también consigue aumentar la pureza de los clusters y con ello, mejorar el funcionamiento de las etapas posteriores.

4.2. Etapa 2: Asimilación de segmentos cortos

La segunda etapa del módulo de postprocesado tiene como objetivo eliminar los segmentos cortos marcados de forma errónea cuando se produce una intervención de larga duración de uno de los locutores. Para ello, en primer lugar, se localizan los segmentos sospechosos de estar erróneamente marcados en función de su duración y la del segmento que le precede. Dichas duraciones se han establecido de forma manual para optimizar el funcionamiento de esta etapa en la parte de desarrollo. A continuación se entrena un modelo GMM (G_x) usando todos los datos disponibles para el locutor marcado originalmente en el segmento sospechoso excepto los correspondientes a ese segmento. Además se entrena el modelo GMM del locutor anterior al segmento sospechoso (G_a), usando todos los datos disponibles en la grabación para él. Finalmente, si el segmento sospechoso queda mejor modelado por G_a que por G_x, es asimilado al cluster del locutor adyacente.

4.3. Etapa 3: Fusión de clusters

La tercera etapa tiene como objetivo la fusión de clusters pertenecientes al mismo locutor. Para ello, en primer lugar se generan modelos

GMM para cada uno de los clusters identificados por el sistema base con 60 segundos del material contenido en ellos. A continuación, para cada uno de los diferentes clusters, se extrae un segmento (con 60 segundos de la información que no ha sido utilizado para realizar el modelo GMM) y se calcula la diferencia de verosimilitudes en dicho segmento para el modelo marcado originalmente y cada uno de los restantes. Valores bajos obtenidos en la diferencia indican que los clusters contienen información similar, por lo que en función de dichos valores y las diferencias relativas obtenidas para cada uno de los diferentes clusters podemos tomar la decisión de combinar dos o más clusters en uno solo. El umbral de decisión para determinar si dos clusters deben fusionarse o no se ha establecido empíricamente optimizando los resultados de esta etapa en la parte de desarrollo de la base de datos.

En la Figura 2 se pueden observar las diferencias obtenidas para el cluster 2, en escala logarítmica. En este caso, los clusters 25 y 41, obtienen valores muy bajos de diferencia en relación al resto de clusters, por lo que parece lógico asumir que los clusters 2, 25 y 41 contienen información del mismo locutor, y por lo tanto podemos fusionarlos en uno solo.

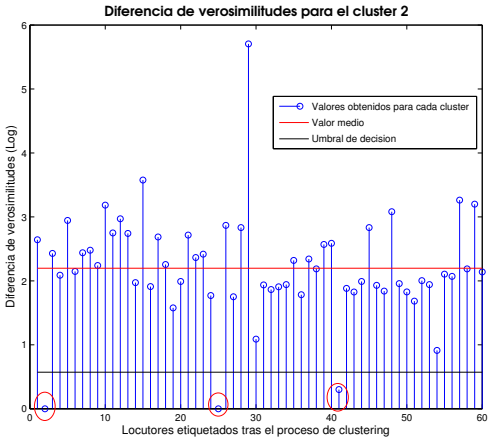

Figura 2: Diferencias de verosimilitudes obtenidas para el cluster 2

5. *Experimentos realizados*

Con el fin de probar el módulo de post-procesado diseñado se han realizado diversos experimentos. Primeramente se han aplicado las tres etapas a los resultados del sistema base, tanto en las sesiones de entrenamiento con las que se optimizaron los parámetros de configuración como a las sesiones de prueba. Posteriormente, y para comprobar la capacidad de generalización del módulo desarrollado se han hecho experimentos para aplicarlo a otros sistemas de diarización con diferentes arquitecturas y al mismo sistema base aplicado a una base de datos diferente.

5.1. Experimentos sobre el sistema base

En las tablas 1 y 2 se muestra el resultado obtenido después de aplicar el módulo de post-procesado a las marcas proporcionadas por el sistema de diarización base. En la Tabla 1 se recoge el DER obtenido para cada una de las sesiones de entrenamiento, tanto para las marcas originales del sistema base como a la salida de cada una de las etapas. Además se muestra en la última línea el valor de DER obtenido en la parte de entrenamiento de la base de datos.

S	DER	E1	E2	E3
1	22.17 %	21.83 %	21.54 %	29.49 %
2	24.58 %	24.46 %	24.38 %	13.10 %
3	23.10 %	23.01 %	22.92 %	18.11 %
4	27.47 %	27.67 %	27.50 %	27.50 %
5	14.15 %	12.94 %	12.93 %	9.89 %
6	21.22 %	21.40 %	21.32 %	16.21 %
7	24.84 %	24.86 %	24.89 %	27.72 %
8	27.26 %	27.38 %	27.38 %	19.90 %
9	28.92 %	28.28 %	28.60 %	26.80 %
10	34.75 %	34.54 %	35.26 %	26.80 %
11	27.94 %	27.70 %	27.90 %	15.91 %
12	27.42 %	27.22 %	27.22 %	25.54 %
13	31.92 %	32.13 %	31.86 %	32.34 %
14	41.16 %	40.87 %	41.00 %	25.84 %
15	32.50 %	32.73 %	32.62 %	27.25 %
16	32.06 %	32.09 %	32.02 %	24.18 %
All	28.25 %	28.14 %	28.16 %	23.33 %

Tabla 1: Resultado de las etapas de post-procesado en las sesiones de desarrollo

En la Tabla 2 se muestra el resultado obtenido para las sesiones de prueba. Al igual que en el caso anterior, se recogen los valores del DER tanto para las marcas originales del sistema base como a la salida de cada una de las etapas, así como el valor del DER obtenido en el conjunto de las sesiones de prueba.

Podemos observar en las tablas anteriores cómo se ha conseguido reducir el error en prácticamente la totalidad de las sesiones, tanto en la parte de entrenamiento como en

S	DER	E1	E2	E3
17	34.92 %	34.69 %	34.62 %	33.97 %
18	31.35 %	30.77 %	30.82 %	19.36 %
19	27.14 %	27.47 %	27.46 %	21.05 %
20	34.72 %	34.57 %	34.76 %	25.52 %
21	34.20 %	34.24 %	34.14 %	18.38 %
22	33.06 %	33.36 %	33.33 %	29.81 %
23	24.92 %	25.05 %	25.18 %	19.48 %
24	22.99 %	22.96 %	23.11 %	17.76 %
All	30.11 %	30.08 %	30.13 %	23.40 %

Tabla 2: Resultado de las etapas de post-procesado en las sesiones de prueba

la de prueba. Al aplicar el módulo de post-procesado se ha conseguido una reducción del DER del 17.4 % en las sesiones de entrenamiento y un 22.3 % en las sesiones de prueba, lo que prueba la validez de las etapas propuestas. En el caso de las dos primeras etapas, apenas se consigue reducción del error, sin embargo, permiten aumentar la pureza de los clusters y contribuyen a mejorar el funcionamiento de la tercera etapa.

5.2. Experimentos sobre otros sistemas de diarización

Una vez comprobado su buen funcionamiento para el sistema base y con el fin de comprobar si el módulo desarrollado puede ser de utilidad en otros sistemas de diarización, se han realizado experimentos para aplicarlo a sistemas de diarización con diferentes arquitecturas. Para ello se han utilizado las marcas proporcionadas por dos sistemas diferentes para la misma base de datos usada en Albayzin 2010. En todos los casos se han mantenido los parámetros de configuración obtenidos en las sesiones de entrenamiento para el sistema base. En primer lugar, se han utilizado las marcas proporcionadas por un sistema de arquitectura similar al sistema base, pero que trabaja de forma online (Luengo et al., 2010). La Tabla 3 recoge los resultados obtenidos al aplicar el módulo de post-procesado propuesto a este sistema.

S	DER	E1	E2	E3
E	26.77 %	26.72 %	26.76 %	21.38 %
P	27.17 %	27.18 %	27.32 %	21.45 %

Tabla 3: Resultado de las etapas de post-procesado sobre sistema online

Podemos observar cómo los resultados obtenidos han mejorado de forma similar a los del sistema de diarización base. El módulo de post-procesado elimina la característica de funcionamiento online de este sistema, pero reduce el DER en un 20.1 % en la parte de entrenamiento y un 21 % en la parte de prueba.

A continuación se ha aplicado el módulo de postprocesado a las marcas proporcionadas por un sistema de diarización desarrollado por el grupo GTM de la Universidad de Vigo (Docio, Lopez, y Garcia, 2010). Las distintas etapas que componen el módulo han sido diseñadas para corregir los errores observados en el sistema de diarización base. El objetivo de este experimento es comprobar el funcionamiento de dichas etapas en un sistema a priori diferente, cuyos errores no han sido analizados. Las tabla 4 recoge los resultados obtenidos.

S	DER	E1	E2	E3
E	25.48 %	25.54 %	25.31 %	25.91 %
T	25.62 %	25.62 %	25.26 %	27.00 %

Tabla 4: Resultado de las etapas de post-procesado sobre sistema GTM

Se puede observar cómo en este caso no se consigue mejora del error. Las dos primeras etapas diseñadas, de carácter menos específico, presentan unos resultados similares a los obtenidos en los casos anteriores, sin embargo la etapa de fusión de clusters alterna sesiones con mejora significativa y sesiones con mayor error. Esta etapa se ha diseñado analizando los errores cometidos por el sistema de diarización base y un sistema con una arquitectura diferente puede no compartir dichos errores, por lo que el resultado obtenido al aplicar esta etapa puede ser contrario al que se busca. Sin embargo, cabe recordar que en ningún momento se han modificado los parámetros de configuración del módulo de postprocesado, por lo que una optimización de los mismos podría conseguir una reducción significativa del error.

5.3. Experimentos sobre otra bases de datos

Por último, se ha propuesto estudiar la independencia del módulo de post-procesado de la base de datos utilizada, por lo que se ha utili-

zado el sistema de diarización base para marcar una pequeña base de datos creada a partir de señales proporcionadas por la Radiotelevisión Vasca (EiTB). Las señales corresponden a una colección de clips de noticias emitidas por EiTB en castellano y euskera durante el año 2010. En los ficheros de audio, además de la voz de los periodistas que narran las noticias y que aparecen repetidos en distintos ficheros, se incluyen también entrevistas y habla doblada sobre el audio original. Parte de estos clips de audio se ha concatenado para formar dos sesiones con diferentes características. La primera es una sesión de 20 minutos de duración en la que aparecen 9 locutores diferentes intercalados con largas intervenciones, en condiciones de bajo ruido. Esto favorece en principio el funcionamiento de las dos primeras etapas de módulo. La segunda es una sesión de 25 minutos en la que 40 locutores alternan intervenciones, que incluye segmentos en entornos con ruido y música de fondo, por lo que el funcionamiento de la tercera etapa debería tener mayor relevancia. Para establecer la referencia se ha llevado a cabo un marcado manual de las sesiones. El resultado obtenido se recoge en la tabla 5.

S	DER	E1	E2	E3
1	35.65 %	34.65 %	32.30 %	32.30 %
2	26.83 %	26.78 %	26.78 %	20.53 %
All	30.26 %	29.84 %	28.93 %	25.11 %

Tabla 5: Resultado de las etapas de postprocesado sobre el sistema base y la base de datos de EiTB

Podemos observar cómo los resultados obtenidos han mejorado de forma similar a los de la base de datos usada en Albayzin 2010. En este caso se ha obtenido una reducción del DER del 17 %. Con estos resultados se comprueba que el módulo desarrollado es válido para otras bases de datos de diferentes caracterśticas.

6. Conclusiones

Se han descrito diferentes técnicas de postprocesado diseñadas para mejorar los resultados de un sistema de diarización de locutores. Se han propuesto tres técnicas para tratar cada tipo de error cometido por dicho sistema: el refinado de la segmentación voz/no voz, la asimilación de los segmentos cortos y la fusión de los clusters del mismo locutor. Estas técnicas se han implementado y se han optimizado los parámetros de configuración utilizando la parte de desarrollo de la base de datos. Se ha dessarrollado un módulo de postprocesado para plicar las distintas técnicas al sistema de diarización base consiguiendo una mejora del 22.3 % en las sesiones de prueba. Con el fin de comprobar si el módulo desarrollado puede ser de utilidad en otros sistemas de diarización, se ha aplicado sin realizar ajustes sobre otro sistema de diarización de arquitectura similar al sistema base con una mejora del 21 % y sobre uno con arquitectura muy diferente sin conseguirse mejoras, aunque se plantea la posibilidad de obtener mejoras optimizado los parámetros de configuración. Por último, se ha comprobado que el módulo desarrollado es válido para otras bases de datos obteniendo una reducción del DER del 17 % utilizando las grabaciones de la Radiotelevisión Vasca (EiTB).

7. Agradecimientos

Los autores quieren agradecer a Iker Luengo el desarrollo del sistema base de diarización de locutores, al grupo GTM de la Universidad de Vigo el acceso a los resultados de su sistema de diarización y a la Radiotelevisión Vasca (EiTB) el uso de sus grabaciones.

Este trabajo ha sido financiado parcialmente por la UPV/EHU (Ayudas para la Formación de Personal Investigador), el Gobierno Vasco (proyecto BerbaTek, IE09-262) y el Ministerio de Ciencia e Innovación (Proyecto Buceador, TEC2009-14094-C04-02).

Bibliografía

Anguera, Xavier. 2006. *Robust Speaker Diarization for meetings.* Ph.D. tesis, Universitat Politecnica de Catalunya.

Anguera, Xavier, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, y Oriol Vinyals. 2012. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):356–370.

Cettolo, Mauro, Michele Vescovi, y Romeo Rizzi. 2005. Evaluation of BIC-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147–170, Abril.

Chen, S. S. y P. S. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. En *DARPA speech recognition workshop*, volumen 6, páginas 127–132.

Docio, L., P. Lopez, y C. Garcia. 2010. The uvigo-gtm speaker diarization system for the albayzin'10 evaluation. En *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, (FALA 2010)*, páginas 401–404, November.

Luengo, I., E. Navas, I. Saratxaga, I. Hernáez, y D. Erro. 2010. AhoLab Speaker Diarisation System for Albayzin 2010. En *FALA 2010*, páginas 393–396, Vigo.

Reynolds, Douglas A y P. Torres-Carrasquillo. 2005. Approaches and applications of audio diarization. En *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, páginas 953–956.

Tavarez, David, Eva Navas, Daniel Erro, y Ibon Saratxaga. 2012. Strategies to Improve a Speaker Diarisation Tool. En *LREC*, páginas 4117–4121, Estambul.

Tranter, S. E. y D. A. Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Trans. on Audio, Speech and Laguage processing*, 14(5):1557–1565.

Zelenák, M., H. Schulz, y J. Hernando. 2010. Albayzin 2010 evaluation campaign: Speaker diarization. En *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, páginas 301–304, Vigo, Spain, November.