UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN**

**ANALYSIS AND DEVELOPMENT OF
ROBUST SPEAKER DIARIZATION FOR
MEETINGS**

PH.D. THESIS

BEATRIZ MARTÍNEZ GONZÁLEZ

**Máster Universitario en Ingeniería de Sistemas Electrónicos
por la Universidad Politécnica de Madrid**

2017

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

**ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN**

**ANALYSIS AND DEVELOPMENT OF
ROBUST SPEAKER DIARIZATION FOR
MEETINGS**

PH.D. THESIS

BEATRIZ MARTÍNEZ GONZÁLEZ

**Máster Universitario en Ingeniería de Sistemas Electrónicos
por la Universidad Politécnica de Madrid**

**Director: José Manuel Pardo Muñoz
Doctor Ingeniero de Telecomunicación**

2017

UNIVERSIDAD POLITÉCNICA DE MADRID
E.T.S.I de TELECOMUNICACIÓN
DEPTO. DE INGENIERÍA ELECTRÓNICA

TESIS DOCTORAL

TÍTULO: Analysis and development of robust speaker diarization for meetings.

DOCTORANDO: D^a Beatriz Martínez González.

DIRECTOR: D. José Manuel Pardo Muñoz

El día _____ de _____ de _____, el Magnífico y
Excelentísimo Señor Rector de la Universidad Politécnica de Madrid, designó el
siguiente Tribunal para evaluar la Tesis Doctoral arriba mencionada:

Presidente: _____

Secretario: _____

Vocal 1.º: _____

Vocal 2.º: _____

Vocal 3.º: _____

Suplente 1.º: _____

Suplente 2.º: _____

Realizado el acto de Lectura y Defensa de dicha Tesis Doctoral en Madrid, el
día _____ de _____ de _____, en la ETSI de
Telecomunicación, el Tribunal acuerda otorgarle la siguiente calificación:

CALIFICACIÓN: _____

EL PRESIDENTE

EL SECRETARIO

LOS VOCALES

Quiero dedicarle esta tesis a mi esposo, la mejor persona que conozco y que nunca conoceré. El compañero perfecto de mi vida. El mejor amigo que se puede desear, que siempre está allí para apoyarme. Esa persona que me hace sentir afortunada sólo por querer seguir a mi lado, que tiene todo lo que alguien podría desear. Que es divertido, leal, cariñoso, comprensivo...

.....Y que además escribe mis dedicatorias

*En mundo tan singular, que el vivir sólo es soñar; y la experiencia me enseña que el hombre que
vive, sueña lo que es, hasta despertar.*

(La vida es sueño, Pedro Calderon de la barca)

Agradecimientos

Quiero agradecer a todas y a cada una de las personas que me apoyaron en el transcurso de esta Tesis. Primero que todo, a mi Director de Tesis, Profesor José Manuel Pardo, sencillamente esto no habría sido posible sin su ayuda.

A Javier, Ascensión y Luis mi especial agradecimiento por la ayuda prestada en la prelectura. Vuestros comentarios y sugerencias sobre esta tesis hicieron que tomara verdadera forma.

Quiero agradecer también a los demás profesores del Grupo de Tecnología del Habla, por la ayuda prestada pero sobre todo por el apoyo y las muchas cosas que he aprendido de vosotros. No sólo de tecnología o ciencia, aunque han sido las más numerosas, sino de todos los temas imaginables.

Quiero agradecerle especialmente a todos mis compañeros de despacho o de trabajo, a todos los que he conocido en este periodo y que se hicieron mis amigos, que me apoyaron y me convencieron de que de verdad había un final. A los que se han marchado ya pero aún me escriben para saber de mí y a los que siguen aquí. Todos lo que me animabais cuando esto parecía insuperable. Los que me enseñaron a bailar salsa o me invitaron a su boda; los que me acompañaron a museos o a jugar padel; los que quedaron para ver un partido de fútbol o simplemente unas cañas... muchas gracias a todos porque sin vosotros no habría aguantado y llegado a donde estoy.

Y por supuesto gracias a mi familia. Gracias a mis padres por el apoyo que siempre me prestan, por ayudarme en todo lo que pueden y por hacerme sentir que siempre estaréis ahí. Gracias Diego e Isa, por ser como sois, por entender que a veces lo que necesitas es que no te agobien y aun así saber transmitir sin palabras que si os necesito no tendré que preguntar dos veces. Gracias a Clara y a David, por hacer felices

a mis dos hermanos, por las conversaciones de química, aunque sea muy básica, que no tengo con nadie más. Por las miles de recomendaciones de gadgets que nunca hubiera comprado si no fuera por David, y sobre todo por los sobrinos que ya después de tantas veces pidiéndolos espero que lleguen antes de que tenga que escribir otros agradecimientos.

Gracias a mis amigos del grupo de “Cuenca”, por sorprenderme con vuestra simpatía y vuestra calidez. Uno nunca sabe en tan poco tiempo si los nuevos amigos que va conociendo estarán ahí en el futuro para apoyarte pero me gusta pensar que podré leer estos agradecimientos en unos años con vosotros a mi lado. Me habéis animado mucho en un periodo muy difícil de mi vida. Leire, Úrsula, Christian y Javi. Muchas gracias.

A Javi un especial agradecimiento por, además, presentarme a Olga, cuyo carácter imposible la convierte en una de las personas más interesantes que he conocido. Gracias Olga por preocuparte de mantener el contacto y por ser siempre tan directa y natural.

Y como no....gracias a la mini. A todos. Sois los mejores amigos que alguien podría querer. A Alberto y Mori por estar siempre ocupados y a la vez siempre disponibles, Mori eres muy especial para mí, pero Alberto es mi amigo más antiguo así que cuídalo porque no hay nadie más leal. A Moni y Rubén que se casaron hace unas semanas y que ni yo misma me creía cuántísima ilusión me hacía estar en su boda. A Sandra y David, sin cuya ayuda nunca habríamos descubierto la pasión por los rebozados (espero un brindis por ellos cuando celebremos esta tesis). A Natalia, que desde que se compró la moto casi no la veo pero que aun así si se lo propongo me viene a buscar a casa y nos vamos a cotillear el tiempo que haga falta. A Jose...que aunque no sabe hablar (el extremeño no es un lenguaje) siempre me hace reír y me hace ver el mundo de maneras que no había sospechado. Y a Bel... que se fue de Madrid y volvió y el tiempo no había pasado, que sólo me confiesa que no me entiende si le pregunto 3 veces, que me ayudó a elegir mi cocina cuando mi marido se declaró en huelga, que comparte mi afición por el champagne y mi odio por las cucas... que es tan cercana y tan importante en mi vida que

parece que nunca hubiera existido una vida sin ella. Y sobre todo que lo sabe todo de mi ... y aun así me quiere.

Y por último gracias a ti, Julián, por todo lo que me das, por todo lo que me has dado y por todo lo que sé que me darás. Por esa forma de apoyarme cada vez que he sentido que no valía para algo. Por preocuparte por mi incluso en tus peores momentos Por descubrir cosas en mi que ni yo sabía que existían. Por saber crear ilusión por proyectos nuevos que siempre quieres compartir conmigo. Por darme a Rupert y a Josefa. Gracias ...por hacer de mi vida un sueño

Abstract

Unlike human beings, following a real conversation is not as easy for machines. We have been trained our whole life, without even noticing, to recognize voices, to remember them and to automatically disregard the non-relevant noises. Speech technologies try to make machines to work as similar to us as possible. Speaker diarization is the part of these speech technologies which tries to solve the task of recognizing, within a recording, who is speaking and when. This analysis could be applied to any recording with two or more speakers (such as broadcast news or telephone conversations) but this thesis will focus on its application in real meetings. To recognize a set of unknown speakers, which the system has never “heard” before, and to follow their speaking turns along the whole meeting requires many pre-processing steps and a very well trained classifier.

In this thesis some features used in diarization systems for meetings are studied and improved by a selection of the most representative ones. Current diarization systems all make use of MFCC features, usually combined with others. If multiple recordings are available, the delay between microphones are the other most extensively used features. We study these delay features in deep and propose several ways to select among them those with the highest quality or those most representative of the real speaker turns. It is found that methods of selection relying on the cross correlation measure between the signals arriving to the different microphones are the most prone to get a reduction in the diarization error. However, other methods which make use of a Principal Component Analysis (PCA) or a Kmeans classification followed by a selection based on the dynamic margin of the delay values or the Silhouette coefficient, achieve great improvement as well.

If the recording has done with only one microphone the use of TDOA features turns impossible. The second part of this thesis presents a study on some glottal features

as substitute features for TDOA when there is no more than one recording channel of the meeting. The analysis is focused on meetings from the media, whose speaking styles can be heavily affected by the type of program (as reportages vs talk shows or meteorology reports). We found that the inclusion of a music detection stage is very beneficial for this kind of audio recordings and that, among the studied features, the Harmonics merged with lif0 (logarithm of the interpolated f_0) and the harmonics to noise ratio (HNR) are the most promising ones that definitely improve the diarization performance.

In the third part of this thesis, all the previously analysed features are used in a preliminary study to detect overlap. As real meetings are supposed to contain people who talk simultaneously, these last glottal features and many more related to the previous TDOA features and its cross correlation are studied to detect overlap. Though this study was preliminary, our conclusion is that some cross correlation and delay related features did show some relation with the presence of overlap regions and should be studied in deep in the future.

Finally, a modification of the segmentation stage of the speaker diarization system is presented. It was discovered during this thesis that the decision of changing from one speaker to another was ruled not only by the acoustics but by a parameter dependent on the number of active speakers, which varies throughout the diarization process. This dependency is cancelled and a new weight factor is added to make the speaker turns independent on the number of speakers and dependent only on the database and the numerical characteristics of the system. In our experiments, the best results were obtained when this weight factor favoured speaker changes when acoustics between speakers were balanced.

Resumen

A diferencia de los seres humanos, seguir los giros en una conversación real no es tan fácil para las máquinas. Nosotros hemos sido entrenados toda nuestra vida, sin siquiera notarlo, para reconocer voces, recordarlas y, automáticamente, ignorar los ruidos no relevantes. Las tecnologías del habla tratan de hacer que las máquinas puedan trabajar tan parecido a nosotros como sea posible. La diarización de locutores es la parte de estas tecnologías del habla que trata de solventar la tarea de reconocer, dentro de una grabación, quién está hablando y cuándo lo está haciendo. Este análisis puede aplicarse a cualquier grabación con 2 o más locutores (como en las grabaciones de noticias o conversaciones telefónicas) pero esta tesis se enfocará en su aplicación para reuniones. Reconocer un grupo de locutores desconocidos, los cuales nunca habían sido “escuchados” previamente por el sistema, y seguir sus turnos de habla a lo largo de toda la reunión, requiere muchos pasos de preprocesado y un clasificador muy bien entrenado.

En esta tesis algunos de los parámetros típicamente usados en los sistemas de diarización para reuniones se han estudiado y mejorado mediante la selección de los más representativos. Los sistemas actuales de diarización hacen uso de los parámetros MFCC, generalmente combinados con otro tipo de características de la señal. Si se dispone de múltiples grabaciones el retardo entre los diferentes micrófonos constituye la otra característica más usada. Estudiaremos los parámetros de retardo en profundidad y propondremos varias formas de seleccionar entre ellos aquellos con la calidad más alta o aquellos más representativos de los turnos que realmente tienen los locutores en la grabación. Encontramos que los métodos de selección que se basan en la medida de la correlación cruzada entre señales que llegan a diferentes micrófonos son los más propensos a conseguir una reducción en el error de diarización. Sin embargo, otros métodos que hacen uso de un análisis de las componentes principales

(PCA) o una clasificación de Kmeans seguida de una selección basada en el margen dinámico de los valores de retardo o el coeficiente de Silhouette, también logran una gran mejora.

Si la grabación se llevó a cabo con un único micrófono el uso de los retardos se vuelve imposible. La segunda parte de esta tesis presenta un estudio sobre algunos parámetros glotales usados como sustitutos de estos retardos cuando no hay más que un canal de grabación en la reunión. El análisis se centra en grabaciones de reuniones sacadas de los medios de comunicación, cuyos estilos de habla pueden estar fuertemente afectados por el tipo de programa (como ocurriría con un reportaje si lo comparamos con un programa de entrevistas o un informe meteorológico). Hemos encontrado que la inclusión de un módulo de detección de música es muy beneficioso para este tipo de grabaciones de audio, y que, entre los parámetros estudiados, los armónicos concatenados con el logaritmo interpolado de la f_0 (lif_0) y la relación de los armónicos a ruido (HNR) son los valores más prometedores que definitivamente mejoran el funcionamiento del sistema de diarización.

En la tercera parte de esta tesis todos los parámetros analizados previamente se usan en un estudio preliminar para detectar habla solapada. Suponemos que una reunión real contendrá personas que hablan simultáneamente. Estos últimos parámetros glotales y otros muchos relacionados con los retardos anteriores y su correlación cruzada se han estudiado con el objetivo de detectar segmentos de habla solapada. Aunque este estudio es preliminar, nuestra conclusión es que algunos parámetros relacionados con la correlación cruzada y los retardos mostraron cierta relación con la presencia de regiones de solapamiento de habla y deberían ser estudiados en profundidad en el futuro.

Para terminar, se presenta una modificación de la etapa de segmentación del sistema de diarización. Durante el desarrollo de esta tesis se ha descubierto que la decisión de introducir un cambio de locutor se regía no sólo por la información acústica, sino también por un parámetro dependiente del número de locutores hipotéticos en cada momento, que varía a lo largo de todo el proceso de diarización. Se ha cancelado esta dependencia y un nuevo factor de peso se ha añadido para que el sistema considere

los turnos de locutor independientes del número de locutores y que tan sólo tenga en cuenta la base de datos y las características numéricas del sistema. En nuestros experimentos, los mejores resultados se obtuvieron cuando este factor de peso favorecía los cambios entre locutores si la información acústica estaba muy equilibrada.

Contents

Agradecimientos	I
Abstract	V
Resumen.....	VII
Contents	XI
Table of figures.....	XV
Table of tables	XXI
1 Introduction and objectives	1
2 State of the art	5
2.1 VAD.....	6
2.2 Segmentation and clustering	7
2.2.1 Online diarization	8
2.2.2 Offline diarization.....	9
2.2.2.1 Bottom-up clustering	11
2.2.2.2 Top-down	14
2.3 Feature extraction.....	15
2.4 TDOAs in deep.....	18
2.4.1 TDOA estimation	20
2.4.2 Use of TDOAs.....	22
2.4.2.1 Creating an enhanced audio signal: Beamforming	23
2.4.2.2 Locate/discriminate Speakers. TDOAs in segmentation and clustering stage	25
2.5 Overlapped speech.....	33
3 Baseline Diarization System	37

4	Databases and evaluation measures.....	43
4.1	Databases	43
4.2	Evaluation measures	48
5	Selection of delay features.....	51
5.1	TDOA selection methods developed.....	53
5.1.1	Cross Correlation Pairs (XCorr-Pairs).....	54
5.1.2	Cross Correlation Channels (XCorr-Channels).....	56
5.1.3	Dynamic Margin (DM)	60
5.1.4	Extreme Regions (ER)	64
5.1.5	Most Common (MC).....	67
5.1.6	Principal Component Analysis.....	69
5.1.7	XCorr threshold	71
5.1.7.1	XCorrThreshold+Baseline	74
5.1.7.2	XCorrThreshold+DM.....	75
5.1.7.3	XCorrThreshold+PCA	77
5.1.7.4	XCorrThreshold+XCorr Methods	78
5.1.7.5	XCorrThreshold+Kmeans methods	79
5.2	Comparison between TDOA selection methods and evaluation	84
5.2.1	Evaluation with the test set	91
6	Alternative speaker diarization features.....	103
6.1	Glottal features	104
6.2	Music detection experiments for speech synthesis.....	111
6.2.1	Evaluation with the test set	118
6.3	Adding glottal features to system with music detection	121
6.4	Features for overlap detection.....	129
6.4.1	Analysis of proposed features.....	130
6.4.2	Experiments and results.....	134
7	Segmentation algorithm	143

7.1	Statement of the problem.....	144
7.2	Experiments, results and discussion	148
8	Conclusions and future work	153
8.1	Conclusions	153
8.2	Future work.....	158
	References.....	163

Table of figures

Figure 1: Diagram of general modules of a diarization system	6
Figure 2: Graphic interpretation of the most common clustering techniques.....	10
Figure 3: Block diagram of a segmentation and clustering stage	11
Figure 4: Sum of every channel of a meeting recording to conform the enhanced signal.	25
Figure 5: Baseline system architecture	37
Figure 6: Diagram of the training and segmentation process of the baseline system	41
Figure 7: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorr-Pairs selection method	55
Figure 8: DER results for the XCorr-Pairs method when the number of pairs selected is 2, 3, 4 and 5, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight	56
Figure 9: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorr-Channels selection method	59
Figure 10: DER results for the XCorr-Channels method when the number of pairs selected is 3 and 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight.....	60
Figure 11: Example of the histogram of delays of a pair of channels and its dynamic range. Delays calculated with this pair show clear separation between peaks which should make it more	

discriminative to locate speakers. The red horizontal line represents the maximum span of the dynamic margin..... 62

Figure 12: Example of the histogram of delays of a second pair of channels and its dynamic range. Bins with very few samples are not taken into account when calculating the dynamic margin. Delays calculated with this pair are mostly concentrated around the same value making it less discriminative to changes in speakers. The red horizontal line represents the span of the dynamic margin..... 62

Figure 13: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the DM selection method..... 63

Figure 14: DER results for the DM method when the number of pairs selected is 2 and 3, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 63

Figure 15: Hypothetical example of the regions that would be defined in the histogram of delays in the ER method. Region A is discarded (containing 0.5% of the total number of delays). Region B is target region (containing the 40% of the remaining number of delay values). 65

Figure 16: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the ER selection method 66

Figure 17: DER results for the ER method when the number of pairs selected is 2, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 66

Figure 18: Hypothetical example of the bins of the histogram of delays that would be chosen in the MC method..... 68

Figure 19: DER obtained for the RT-DEVELSET across the weight applied to the TDOA vector and the final dimension of the TDOA vector when using the MC selection method..... 68

Figure 20: DER results for the MC method when the number of pairs selected is 6 and 8, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 69

Figure 21: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the PCA selection method 70

Figure 22: DER results for the PCA method when the final dimension is 6, 7, 8, 9 and 10, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 71

Figure 23. XCorr of one meeting with two arrays. The order of the combinations of microphones are represented according to its cross correlation measure in descendent order. 72

Figure 24: DER for the baseline system using MFCCs and Delays, weight to MFCC vector is 0.9 and 0.1 for the delays. K is the factor applied to the maximum xcorr to determine the minimum threshold. In columns is the DER for the RT-DEVELSET..... 74

Figure 25: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorrThreshold+DM selection method 76

Figure 26: DER results for the XCorrThreshold+DM method when the number of pairs selected is 5 and 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 76

Figure 27: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorrThreshold+PCA selection method 77

Figure 28: DER results for the XCorrThreshold+PCA method when the final dimension is 5, 6, 7 and 9, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 78

Figure 29: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the Kmeans+DM selection method 82

Figure 30: DER results for the Kmeans+DM method when the number of pairs selected range from 2 to 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight 82

Figure 31: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the Kmeans+Silhouette selection method.....	83
Figure 32: DER results for the Kmeans+Silhouette method when the number of pairs selected range from 4 to 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight.	84
Figure 33. DER of new methods for the development set with their optimal dimension of the TDOA vector, using MFCC and TDOA features. Results are shown across the weight applied to the TDOA stream.	85
Figure 34: Detail in the area of minimum error of the DER of new methods for the development set with their optimal dimension of the TDOA vector, using MFCC and TDOA features. Results are shown across the weight applied to the TDOA stream.	86
Figure 35: Optimal DER for the 6 methods with the development set when selecting a specified number of pairs.....	89
Figure 36: Experiments with the whole range of glottal and F0 features across the weight applied to the new feature and for #initialgaussians to train the model ranging from 1 to 5. Experiments were carried out in combination with MFCC stream. Weights applied to the two streams sum1.	109
Figure 37: Results for the C-ORAL-ROM-DEVELSET when applied and not the noise reduction module and with 16 and 32 clusters for the initial segmentation. Systems with noise reduction are labelled with NR, and k is the number of initial clusters. Experiments were carried out with MFCC and lif0 features, which weights added equals 1.	113
Figure 38: Block diagram of the system with Music detection.....	114
Figure 39: Precision, recall and merit figure versus insertion penalty of the music recognizer for the C-ORAL-ROM DEVELSET. F0_weight=0.05 for system with noise reduction (NR) and 0.15 for system without it. MD stands for Music detection applied.	116
Figure 40: Merit figure for the C-ORAL-ROM DEVELSET through various weights applied to the glottalFeature, and using 1 and 2 gaussians for the initial model created with the glottal features. Music detection was applied to discard all intervals of music. All experiments have	

been done with MFCC features too. The weight applied to MFCC features is 1-weight_glottalFeature..... 124

Figure 41: Merit figure for the C-ORAL-ROM DEVELSET through various weights applied to the glottalFeature, and using 3, 4 and 5 gaussians for the initial model created with the glottal features. Music detection was applied to discard all intervals of music. All experiments have been done with MFCC features too. The weight applied to MFCC features is 1-weight_glottalFeature..... 125

Figure 42: Distribution of the training multispeaker (red dots) vs single speaker (blue dots) for the training AMI set processed to have equal number of overlapped segments than non-overlap segments. X axis in all figures is XCORR2_ave_vars . Y axis is the following: Figure 42.a:XCORR2_ave_means; Figure 42.b: DEL_ave_entropies; Figure 42.c: DEL_ave; Figure 42.d : RATIODEL1DEL2_ave_entropies. 137

Figure 43: Merit figure of the development AMI dataset across the number of attributes used to train the classifier..... 141

Figure 44: Cluster models with minimum duration (from (Anguera, Wooters, & Hernando, 2006)) 145

Figure 45: Diagram of the dependency of turn speaker changes with the number of active speakers. 145

Figure 46: SER of the RT-DEVELSET for minimum duration of speech of 200, 250 and 300 frames against transition weight 150

Table of tables

Table 1: Summary of pros and cons of different channels selection methods	32
Table 2: List of meetings for the RT database used in the experiments.....	44
Table 3: List of meetings for the C-ORAL-ROM database used in the experiments.	46
Table 4: Characteristics of the C-ORAL-ROM database per type of recording	47
Table 5: List of meetings for the AMI database used in the experiments.....	47
Table 6: number of possible pair of channels combinations per number of channels selected.	57
Table 7: Comparison table between the two XCorr related methods for a hypothetical meeting when selecting 4 pairs or 4 channels.....	58
Table 8: The DER and SER for the development and the test sets (DEVELSET and RT09), using MFCC and TDOA streams. Weight of MFCC stream for the baseline and MC methods is 0.9. The Weight of MFCC stream for XCorr-Pairs, DM, ER, PCA, XCorrThreshold+baseline, XCorrThreshold+DM, XCorrThreshold+Kmeans+DM and XCorrThreshold+Kmeans+Silhouette methods is 0.85. Weight of MFCC stream for the XCorr-Channels and XCorrThreshold+pca methods is 0.8. TDOA stream weight is 1-MFCC weight. The percentage error improvement on the baseline is shown in brackets.	92
Table 9: Results of each meeting for set RT09 and methods XCorr-Pairs, and PCA. Weight of MFCC stream for the baseline is 0.9. Weight of MFCC stream for PCA and xcorr-pairs methods is 0.85. Weight of MFCC stream for XCorr-Channels method is 0.8. TDOA stream weight is 1-MFCC weight.	94
Table 10: Results of each meeting for set RT09 and methods Baseline, DM, PCA, Kmeans+DM and Kmeans+Silhouette combined with xcorrThreshold. Weight of MFCC stream	

for the baseline is 0.9, for DM, Kmeans+DM and Kmeans+Silhouette methods is 0.85 and 0.80 for PCA method. TDOA stream weight is 1-MFCC weight. 95

Table 11: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods XCorr-Pairs, XCorr-Channels, PCA and baseline. 97

Table 12: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods Baseline, DM and PCA combined with XCorrThreshold. . 98

Table 13: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods Kmeans+DM and Kmeans+Silhouette combined with XCorrThreshold. 99

Table 14: Computational time for the set RT09 using MFCC and TDOA streams relative to computational time of baseline system..... 100

Table 15: Results for the C-ORAL-ROM DEVELOPSET . Relative precision error improvement is calculated over the baseline. K stands for the initial number of hypothetical speakers, K=16 if nothing indicated. NR stands for noise reduction algorithm and MD stands for music recognition algorithm. 118

Table 16: Results for the C-ORAL-ROM TESTSET. Relative precision error improvement is calculated over the baseline. NR stands for noise reduction algorithm and MD stands for music recognition algorithm..... 120

Table 17: Precision, recall and merit figure for the different styles in the C-ORAL-ROM TESTSET. 121

Table 18: Results for the C-ORAL-ROM DEVELOPSET for the best combinations of features. MD stands for music detection, 126

Table 19: Results for the C-ORAL-ROM TESTSET for the best combinations of features. 127

Table 20: Precision for the C-ORAL-ROM TESTSET meeting per meeting 128

Table 21: precision for the C-ORAL-ROM TESTSET per style of the meetings 129

Table 22: Pearson correlation coefficient of the features used to detect overlap segments	136
Table 23: Precision and recall using a 10-fold validation for the AMI training set processed to have similar number of single speaker frames than multiple speaker. The features used are the best # Features ranked in Table 22.....	139
Table 24: Precision and Recall using a 10-fold validation for the AMI training set when classifying non-overlapped speech and overlapped speech regions. Meetings have many more samples of non-overlapped speech than overlapped. The features used are the best # Features used best ranked in Table 22.	140
Table 25: features ranked with Pearson coefficient from 6 th to 11 th position when analysing its correlation with overlap segments.....	142
Table 26: SER for every point evaluated varying the transition weight of the segmentation stage for the RT-DEVELSET and the RT09 set. M is the number of active clusters at each iteration. Weight applied to MFCC and TDOA features is 0.85 and 0.15 respectively.	150

1 Introduction and objectives

Speech is the most natural way of communication among human beings, and, nowadays, could be even with technologies. Mobile phones, computers or tickets machines can recognise voices but will have a hard time recognising speech in a very noisy environment or differentiating if two similar voices come from different speakers. Technology is everywhere and we have got used to deal with machines in alternative ways to what would be more natural to us: speaking.

The development of technologies has favoured also the quantity of information available, and much of that information is recorded in a video or audio file. But the only way of knowing in advance if that file contains what a specific person is looking for is entrusted to the name of the file or in the few keywords selected by the owner. The search for a word in a text document is nowadays very easy but not in a video or audio file.

Speech technologies try to solve this and other problems related to human voices. Recognition of the speaker for security issues, detection of topic in a video file to be able of index any file in an unsupervised manner, voice synthesis to make the voice of any machine more natural or the transcription of audio files would be among the tasks studied in speech technologies.

In this thesis we will work in a specific area of speech technologies that focuses on differentiating which speaker was talking at any second. It is usually said that diarization aims to answer the question “who spoke when?”. National Institute of Standards and Technology defines the speaker diarization as the task of identifying the number of participants in a recording and creating a list of the speech time intervals of each speaker. This task should be carried out with no previous knowledge of the identity

of the speakers, characteristics of the room where the recording took place, position of the speakers in that room or the quality of the microphones used to record the audio. In this work we will work with meeting recordings.

Speaker diarization can be applied as a first step to many systems as it would help to focus on what is said by only one person even if there are many more in the recording or we do not know who that person is. If it is applied to transcription of audio files the result would be much more readable than an unorganised text where no attention is set to who said what.

The diarization task includes many areas of speech research as speaker recognition, speech/non speech segmentation, location of speakers, etc. In this thesis we will focus mainly on the use of a specific feature, the Time Delay of Arrival (TDOAs), and other parameters related with them, and study how they can be improved to solve some of the main problems affecting the performance of the diarization.

The TDOAs are used nowadays in most of the state of the art systems. A first objective is to modify the system that calculates the delays between microphones to be able to extract more robust and useful TDOA values in order to improve the diarization performance. We intend to implement different techniques to take advantage of all the information concealed in the TDOA values that are not currently used in the baseline system. Among the methods to be tested are the extraction of all the possible values coming from all possible pairs followed by a dimensionality reduction and the selection of the pairs of microphones more suitable to detect the present speakers. The study of different criteria for this selection is a main objective in this thesis.

A second objective is to explore the use of alternative features like some glottal features typical of speech synthesis in the case when TDOAs cannot be used, which is quite common for media data, for instance. If the recording has been done with only one microphone TDOAs are not available. There are many databases that use a single channel or a single microphone and features used in other areas of speech research could help the diarization in such cases. Our intention is to find some features that could substitute TDOAs if necessary and achieve similar performance.

Finally, as the third main objective, we want to take advantage of the delay features and some parameters related with the calculus of them (as the cross correlation measure) in sections of the diarization system in which they are not currently used. This parameter is used to determine the TDOA between microphones; it is therefore calculated by the baseline system but not used furthermore. As it is accessible and a possible measure of the quality of the audio the objective is to find its possible use for other purposes, as the detection of overlapped regions.

As an extra objective, while studying the system we discovered some characteristics of the segmentation stage that could be modified to improve the overall system. There is a transition weight, inherited from the previous system from which our diarization program was developed, that has some impact in the duration of a speaker turn. As we will explain, this parameter was not actually adjusted for the diarization system, is independent of the acoustic data and varies throughout the whole process with the number of hypothesized speakers in each moment. We want to analyse this transition weight in conjunction with other variables that modify the duration of the speaker turn, and see how much diarization can be improved by it.

Summarizing, the objectives of this thesis can be expressed in the following items:

- Development of techniques to improve the TDOAs used as input features for the segmentation and clustering module. This problem will be addressed by two ways:
 - Selection of the best pairs of microphones to compute TDOAs.
 - Extraction of every possible TDOA value followed by a dimensionality reduction.
- Study of the use of alternative features, as glottal features used for speech synthesis, for recordings in which TDOA features are not available.

- Study of the possible uses of TDOAs and other related parameters for the detection of regions suspect of containing impure or overlapped speech and discard them in speaker modelling.
- Modification of the segmentation stage of the diarization system to eliminate the dependency of the duration of speaker turns with the number of speakers hypothesized at each moment. Also to study the effects on diarization of turning this parameter a fixed one established experimentally.

From now on the organization of this thesis will be the following: Chapter 2 will introduce the state of the art of diarization, including not only the most extended approaches to solve the diarization problem, but explanation of the most commonly used features used, focusing on the explanation of the delay features as they will be extensively used in this thesis. Chapter 3 includes the explanation of the baseline diarization system used for the experiments. In chapter 4 the databases and evaluation measures that will be used for the different experiments are described. Chapter 5 begins with the contributions of this thesis, with the explanation of the methods developed for the selection of delay features and the presentation of the error results obtained for each method developed. In chapter 6 we introduce new features (glottal features) and analyse its performance for different tasks and datasets. One of the datasets is compound of meetings with different speaking styles and music and other noises. The use of glottal features and the application of a music detection algorithm will be one of the objectives of this chapter. The other objective will be to perform an initial study about the expected utility of the new features used until now in this work for the detection of overlap. Finally, chapter 8 ends with the general conclusions and some proposals of future lines of work.

2 *State of the art*

Speaker diarization is the task of identifying the number of participants in a recording and creating a list of speech time intervals for each participant. Speaker diarization is useful as a first step in the speech transcription of recordings in which each spoken sentence has to be assigned to a defined speaker. It can also be used for speaker adaptation in speech recognition. An overview of automatic speaker diarization systems is given in (Moattar & Homayounpour, 2012), (Tranter & Reynolds, 2006), (Anguera, et al., 2012).

Speaker diarization systems consists usually of three main blocks: the Voice Activity Detection module (VAD), which filters out non-speech events, the feature extraction module, where all the necessary information is extracted from the recording, and the segmentation and clustering module, which uses the previous features to segment the whole recording into clusters, i.e. speakers. Figure 1 shows a diagram of this structure.

Speaker diarization was first applied to broadcast news recordings (BN). One of the best published systems can be seen in (Barras, Zhu, Meignier, & Gauvain, 2004). Subsequently speaker diarization was applied to the meeting domain under the Single Distant Microphone (SDM) condition. The meeting domain differs from BN as the topics are highly diverse, the participants have idiosyncratic relationships and vocabularies, the meetings are highly interactive, and there can be simultaneous speech from multiple speakers. Furthermore, distant microphones are susceptible to reverberation and background noise. On the other hand, the number of speakers in the BN domain tend to be higher. However, since 2002 most efforts have been addressed to the multiple distant microphone (MDM) domain, extending the methods applied to SDM or previously to BN and adding new features.

In speaker diarization with MDM, redundant information is available (one signal per microphone) in comparison to SDM diarization. All speech signals are usually combined into one (Anguera, Wooters, & Hernando, 2007), from which some acoustic features are extracted. The other source of information used in MDM scenarios is the information related to speaker location (Ellis & Liu, 2004) such as the time delay of arrival (TDOA)

Figure 1: Diagram of general modules of a diarization system

2.1 VAD

The VAD module would discard for posterior stages of the diarization any segment which does not contain speech. Any error in the classification in this part of the process would affect the rest of the diarization. This is a crucial part because most systems cannot recover from a misclassification in this stage. If some part of the recording is actually speech but the VAD module misclassifies it as non-speech, the system neither would use that segment to model any hypothetical speaker nor would associate it to any of the speakers of the final hypothesis. In the same way, if a non-speech segment is detected as speech, the posterior stage could use it to model the speakers and group it with other speech segments of any speaker.

Voice activity detection is, by itself, a large area of study, which is used as a first step in many other areas of speech research. Therefore, there are many options in the literature to carry out this task. One of the most accepted techniques consists in classifying every segment of the recording according to its similarity to two Gaussian Mixture Models (called GMM from now on). One of the models should have been trained for speech and the other for silence and noises (Anguera X. , Wooters, Pesking,

& Aguiló, 2006). Though the use of other types of statistical models like complex Gamma and Laplacian pdf are also studied in the literature (Chang, Kim, & Mitra, 2006), GMMs are still the most widely used in diarization systems.

The methods that classify the signal only in two groups (speech and non-speech) are the simplest. We can find other works with more classes as in (Zhu, Barras, Meignier, & Gauvain, 2005), with 5 models (music, speech, speech over music, silence and noise) or in (Liu & Kubala, 1999), with 8 (including music, laughter, breath, lip-smack, silence, and three phonetic groups). In (Wooters & Huijbregts, 2008), the system classifies segments into 3 groups: speech, silence and noises. The “noises” group could include sounds like door slamming, paper shuffling, laughing, music or any other non-speech sounds.

The main drawback of VAD systems with many models is that they should be trained in advance with labelled external data which makes them less robust to changes in acoustic conditions. To make the system more robust the models should be trained with data of the same recording. One common solution is to perform an initial energy based segmentation, labelling a limited amount of segments as speech or silence according to its energy. Then this labelled data is used to train specific models of speech and silence, which are used to resegment the whole recording.

2.2 Segmentation and clustering

The ultimate goal of speaker clustering is to associate segments of speech to its corresponding speaker. In ideal conditions the number of clusters created should be equal to the number of speakers present in the recording. The very definition of the diarization problem set this number as unknown, thus, the final number of clusters, i.e. speakers, will depend on the process followed by the segmentation and clustering technique.

From a general point of view one can differentiate between online and offline systems. The offline systems have access to the whole recording before the beginning of the diarization process, while the online methods can access only to data recorded up to that moment.

2.2.1 Online diarization

In online diarization data is gradually fed to the systems and decisions should be made using only the current and previously seen speech data. This entails some challenges in the initial models, as for some time, there will not be sufficient data for any training. The key issue will be then, finding the way of alleviating the model initialization problem. One possible solution is to use a generic model trained on external data (from different speakers) to help the initialization, adapting it with the real data, as in (Kwon & Narayanan, 2005). Other works use initial training data extracted from the conversation in order to produce the initial speakers and non-speech models, as in (Ben-Harush, Lapidot, & Guterman, 2010), and working with low latency from that point.

Some generic models used in online diarization are Universal Background Models (UBM) or Universal Gender Model (UGM). In (Kwon & Narayanan, 2005) these options are compared to an alternative, the so called Sample Speaker Models (SSM). This technique creates a set of generic speaker models using samples of external data. These samples are picked up from a pool of generic speaker models using the Markov Chain Monte Carlo method (MCMC) (Andrieu, de Freitas, Doucet, & Jordan, 2003). In this work a speaker change detector based on Generalized Likelihood Ratio detects change points and the generic models are adapted to speaker models with the newly indexed data. Other distance measures classically proposed in the literature could be the Bayesian Information Criterion (BIC), which is very accurate but usually considered very costly for online diarization, or the Kullback-Leibler (KL) divergence (Rougui, Anoutajdine, Gelgon, & Martinez, 2006).

Offline diarization is generally more accurate, as the system can use all the information available from the beginning of the process, but it is more time costly. In

(Vaquero, Vinyals, & Friedland, 2010) both strategies are combined to benefit from the accuracy of the offline systems but working online. The offline system performs regular diarization with data up to some time; when it finishes, the final segmentation is updated to the online diarization system to train new speaker models. In the meanwhile, the online diarization has been working with the previous models, which are removed once new more accurate models are trained. The offline system will then start again, using this time the audio data previously analysed plus the new audio arrived while it was performing the previous diarization. With this technique there is a delay at the beginning of the process for the output of the first label, while the offline system recovers enough data, process it and the online system trains the first models. Every iteration the offline system has more audio data to work with, and the time required to finish the diarization grows. There are two drawbacks of this system. The first one is that the models will be more accurate but will reach the online system a bit later every iteration allowing for more segmentation errors in the meantime. The second one is that the online system is not designed to detect new speakers so every time a new person participates in the conversation the output does not recognize it until the offline system process its speech and decides that there are more participants than before. The speed of the offline system turns to be the key part of the problem, which is assessed in (Friedland G. , 2012), where the offline system is modified to perform the critical parts of the process in a GPU, speeding up the output and achieving error rates comparable to those of offline diarization without GPU.

2.2.2 Offline diarization

In the bibliography, offline methods are the most common with the entire speech file available at the starting point. As mentioned before, this should work better than online clustering because it has access to more information.

Most present state-of-the art systems fit into one of two hierarchical schemes, the top-down (Fredouille, Bozonnet, & Evans, 2009) and the bottom-up techniques (Wooters & Huijbregts, 2008), (Sun, Nwe, Ma, & Li, 2009). The initial segmentation is opposite for the two approaches, as one begins segmenting in only one cluster while the

other starts with many. In top-down approaches one or very few initial clusters are splitted iteratively until obtaining an optimum number of clusters. In bottom-up (also known as Agglomerative Hierarchical Clustering (AHC)) many initial clusters are merged iteratively until reach this optimum number of speakers.

The Figure 2 shows a graphical representation of these two techniques.

Figure 2: Graphic interpretation of the most common clustering techniques

For both techniques the distance function used to determine the similarity of two clusters is important. The stopping criterion applied that would determine when the system finishes the merging/splitting, and therefore, how many clusters are the optimal is also crucial. Figure 3 represents a diagram of the general segmentation/clustering stage of a diarization system. “Audio Input Data” are the speech files and “Data Preprocessing” stages includes any process prior to the segmentation module, such as noise filters, VAD or feature extraction.

Both systems perform an initial segmentation (many segments for bottom-up methods and a single one for top-down methods) and training of the models, then check if the stopping criterion has been met. If it has not been reached, a distance between clusters and a split/merging mechanism is used to iteratively merge clusters or introduce

new ones. When the stopping criterion is met the system finishes delivering the final segmentation.

Figure 3: Block diagram of a segmentation and clustering stage

2.2.2.1 Bottom-up clustering

In the bottom-up approach, the system usually segments the recording in many more clusters than the expected number of speakers. A model for each of the clusters is then trained and the most similar are merged. A single step of segmentation followed by one merging step is rarely enough to have a fine final segmentation. The system could not recover from any error in the segmentation step. In most systems the process works iteratively, new models are trained after the merging step and the recording is resegmented according to the active clusters. This process is repeated until a stopping criterion is reached. One drawback of this technique is that there is the risk of losing some speakers from the very beginning of the process if the system segments, in the first iteration, in a lower number of speakers. Most systems just define an initial number of clusters high enough to avoid this problem.

The initial number of clusters is usually set empirically. Some works have proved that this initial segmentation can affect drastically the rest of the process. In (Sinclair & King, 2013) the authors proved that pure speaker models are crucial to achieve better performance in speaker diarization. Training models based on impure clusters degrade heavily the results, and the models do not recover in the following iterations. The initial segmentation is typically carried out uniformly, with the audio stream divided into a set of equally long segments. These uniform segments are used to train the first models. As the number of initial speakers is set empirically, all the meetings evaluated will have the same number of initial models.

Some works as (Inseng & Friedland, 2010) show that slight variations of the values related to initialization have a large impact on the accuracy of the result, and that the two most important parameters are the number of initial clusters and the initial number of gaussians per cluster. They use an initialization method that takes advantage of a set of prosodic features, as pitch or harmonics, to perform the initial segmentation. In contrast to the uniform segmentation the initial number of clusters and the speech associated to each one can be different. They also propose a technique to determine the initial number of gaussians per cluster based on the number of initial clusters and the length of the seconds of speech.

Other works take advantage of other features to perform the non-uniform initialization, as in (Luque, Segura, & Hernando, 2008) or (Nwe T. L., Sun, Ma, & Li, 2012), where the authors study the application of TDOA for the initial clustering of meetings.

In (Ben-Harush, Ben-Harush, Lapidot, & Guterman, 2012) six initialization techniques are applied to the diarization of telephone conversations. The best results are obtained with a variation of k-means based initialization method called Weighted Segmental K-means (WSKMA). Though experiments are conducted in telephone conversations recordings, the authors posit that the proposed initialization techniques could be applied to any iterative diarization system operating on conversations between more than two speakers.

Once the initial segmentation is ended, the system usually creates a model for each of the initial speakers modelling a GMM with data from each of the previous

segments. Standard distance metrics are then used to select the closest clusters to be merged. The new cluster is then trained on the data that was previously assigned to the two individual clusters.

Many different distance metrics have appeared in the literature being the Bayesian Information Criterion (BIC) and its associated ΔBIC (Chen & Gopalakrishnan, 1998) the most popular. According to (Ajmera & Wooters, 2003), using BIC as a merging criterion, two clusters would become a candidate for merging if the following is true:

$$\log p(D|\theta) - \frac{1}{2}\lambda K \log N \geq \log p(D_a|\theta_a) + \log p(D_b|\theta_b) \quad (2-1)$$

Where D_a and D_b represent the data in two clusters and θ_a and θ_b represent the parameters of the probability density functions (pdfs) of these two clusters respectively. D is the data from $D_a \cup D_b$. and θ represents the parameters of pdf of D . N is the number of data points in D and K is the difference in the number of parameters between θ_a and θ_b . λ is a free term that changes with data conditions and therefore needs to be tuned to get the best results (as in (Chen, et al., 2002) or (Cettolo & Vescovi, 2003)).

To avoid the need of finding an optimal value of this parameter, in (Ajmera & Wooters, 2003) the authors propose to model the pdf of the new cluster with a number of parameters equal to the sum of the number of parameters of the two merged clusters. For equation (2-1) this means that the number of parameters in θ is equal to the sum of the number of parameters in θ_a and θ_b . Given this condition, two clusters would become a candidate for merging if equation (2-2) is true.

$$\log p(D|\theta) \geq \log p(D_a|\theta_a) + \log p(D_b|\theta_b) \quad (2-2)$$

Related to BIC metrics it has being also developed in recent years the so called segmental BIC (Stafylakis, Datsouros, & Carayannis, 2009), with promising results.

Though BIC approaches are more used, other distance metrics could be mentioned as the Generalized Likelihood Ratio (GLR) (Delacourt & Wellekens, 2000), the penalized GLR, the Kullback-Leibler (KL) divergence (Siegles, Jain, Raj, & Stern, 1997), the

KL2 metric (Zhou & Hansen, 2000), or the cross likelihood ratio (CLR) (Fernandez, Otero, & Mateo, 2009). In (de Campos Niero, de Lima Veiga Filho, & Adami, 2014) a comparison between GLR, CLR and a BIC based distance metric for clustering is presented. In this work the authors show that CLR can be as discriminative as BIC but the number of gaussian components and speech duration should be taken into account in order to obtain optimal results.

As mentioned before, in most diarization systems the clustering and segmentation are carried out iteratively. After the merging of two clusters, and the training of the new model, a reassignment of frames to clusters is performed via Viterbi realignment for example. The new clusters are used then to train new models, which will be merged if two of them fulfil the requirement of equation (2-2).

The stopping criterion determines when this process ends. Criteria based on BIC are the most widely used. According to this, the stopping criterion is met when ΔBIC is lower than 0 for every pair of clusters, being ΔBIC defined as shown in equation (2-3)

$$\Delta BIC = \log p(D|\theta) - \log p(D_a|\theta_a) - \log p(D_b|\theta_b) \quad (2-3)$$

Where θ_a is the model created from cluster D_a , θ_b is the model created from cluster D_b and θ from D , where D is the combination of D_a and D_b .

2.2.2.2 Top-down

The second approach starts creating only one model for the whole meeting, but leaving segments unlabelled. Then the system uses some selection procedure to identify suitable training data and creates a new model. Various approaches may be used to select appropriate data but the single largest segment identified from the speech activity detection output has proved to give consistent performance (Bozonnet, Evans, & Fredouille, 2010). As with the bottom-up approach a Viterbi realignment is then applied and the frames are labelled accordingly. New speaker models are iteratively added one-by-one, repeating the process until no more speakers can be added. The system could make use of some stopping criteria similar to those employed in bottom-up approaches or can continue until no more relevant unlabelled segments remain. Top-

down approaches are far less popular than the bottom-up counterparts and are usually outperformed by these. Still, they perform reasonably well and have the distinct advantage of not having to start with a possibly too low or too high number of speakers. Top-down systems are also extremely computationally efficient. A comparison between the two approaches is presented in (Evans, Bozonnet, Wang, Fredouille, & Troncy, 2012).

2.3 Feature extraction

Feature extraction is a key part of the signal processing applications. In diarization the features utilized are intended to provide information about the speakers in the conversation in order to be able to separate them optimally. According to (Kinnunen & Li, 2010) ideal features for speaker discrimination should include the following characteristics:

- Have large between-speaker variability and small within-speaker variability
- Be robust against noise and distortion
- Occur frequently and naturally in speech
- Be easy to measure from speech signal
- Not be affected by the speaker's health or long term variations in voice
- Be difficult to impersonate/mimic
- Number of features relatively low.

Previous works have studied different features for the diarization task. The most widely used are MFCCs, sometimes with the first or second derivatives. However, there is no agreement about the optimal order of the MFCCs. While some systems work with a 24-order MFCCs as in (Ajmera, McCowan, & Boulard, 2004) or (Cheng & min, 2004), other works with 23-order MFCCs as in (Kim, Ertelt, & Sikora, 2005). In (Wu & Hsieh, 2006) several MFCCs orders and their derivatives are investigated before choosing the 12-order MFCCs and their first derivatives.

Other features that we can find in literature are: short-time energy (STE) (Meignier, Moraru, Fredouille, Bonastre, & Besacier, 2006), zero-crossing rate (ZCR) (Lu & Zhang, 2002), intensity channel contribution (Barra-Chicote, Pardo, Ferreiros, & Montero, 2011), pitch or line spectrum pairs (LSP) (Lu & Zhang, 2005), spectrum magnitude (Boehm & Pernkopf, 2009) and perceptual linear prediction (PLP) cepstral coefficients, (Tranter, Yu, Evermann, & Woodland, 2004) and (Chu, Tang, & Huang, 2009).

As mentioned in (Anguera, et al., 2012), MFCCs are traditionally used both for speaker recognition, that requires invariance against words, and for speech recognition which requires invariance against speakers. This theoretical inconsistency has favoured the use of prosodic and other long-term features which, as already suggested in (Shriberg, 2007), can benefit the diarization task revealing individual characteristics of the speakers' voices as well as their speaking behaviour. This information cannot be captured using a short-term frame-based cepstral analysis. In (Friedland, Vinyals, Huang, & Muller, 2009), the authors study the speaker discriminability of 70 long-term features, most of them prosodic features, then, combine the best of them with short-term features. They proved that despite the dominance of short-term features, their combination with long-term features can increase the accuracy of speaker diarization by a 30% relative decrease of diarization error rate (DER). The features utilized are separated into 5 categories:

1. Pitch Features: Pitch is influenced by the length and mass of the vocal folds in the larynx (Dellwo, Huckvale, & Ashby, 2007). Features studied include mean, median, min and max, the difference between them, the standard deviation and the slope of the pitch curve.
2. Energy Features. Changes in energy are not as much related with the speaker identity as pitch but are related with emotions and with stress marks. Still some works have proven their utility in the discrimination of speakers (Barra-Chicote, Pardo, Ferreiros, & Montero, 2011) and the authors hypothesize that related features (as min, max, differences between the former two, mean and standard deviation) could be of use.

3. **Formant Features:** Formants occur around frequencies that correspond to the speaker-specific resonances of the vocal tract, which, theoretically, makes them suitable for recognition of the speaker. However, formants also depend on phonetic content. They occur only in voiced segments and even within one phonetic class the shape of the formants and the relation between them varies. Nevertheless, the variation happens for the most part in the first two formants while the higher ones are generally assumed to capture speaker specific information. Candidate features, calculated for formants 1-5 are mean, median, min, max and standard deviation
4. **Harmonics-to-Noise-Ratio (HNR).** These features quantify the relative amount of additive noise in the voice signal. Spectral noise can be caused by aperiodic vocal fold vibration and turbulent airflow generated by inadequate closure of the vocal folds during phonation (Ferrand, 2002). The resulting friction noise is reflected in higher noise level in the spectrum. The features calculated include mean, min, max, difference between these two and standard deviation.
5. **Long-Term Average Spectrum (LTAS):** To obtain this, the spectral energy in 100-Hz-wide frequency bands is measured over a large portion of speech. The standard deviation, the slope of the curve, the frequency associated with both the lowest energy and the highest and the peak heights were calculated.

The authors of (Friedland, Vinyals, Huang, & Muller, 2009) compare all the features mentioned using a pre-experiment with a different database and combine the 10 best with MFCC in the final diarization experiments. The top ten features that yield 30% improvement are the median, mean and min of f_0 , the standard deviation, min and mean of 4th formant, the mean of the HNR, the mean and standard deviation of the 5th formant and the standard deviation of the LTAS.

The combination of multiple feature streams requires usually the execution of multiple experiments to decide the weight to be applied to such features. The decision of the working point, even with all the experiments finished, may not be straightforward. In (Vijayaseenan, Valente, & Boulard, 2010) the authors addressed the combination of MFCC and TDOA features with two more feature streams: Modulation

Spectrum (MS) Features (Vinyals & Fiedland, 2008) and Frequency Domain Linear Prediction (FDLP) Features (Thomas, Ganapathy, & Hermansky, 2008). MS features represent the slowly varying components of the short term spectrum and FDLP provides a smoothed temporal envelope of the speech signal. The authors discuss the issues and problems related to the combination of 4 feature streams and propose a solution based on a smoothed version of the speaker error.

In MDM diarization, we can also take advantage of the extra information provided by the use of multiple microphones. In comparison to SDM diarization, in MDM diarization there are multiple recordings of the same meeting. As mentioned before, having redundant information we can delay-sum the recordings in an improved channel. The Time Delay of Arrival (TDOA) obtained to perform this sum can also be used to differentiate between the speakers in the room, even though there is no prior information about the location of the microphones in the meeting room. The use of these particular features is the main objective of this thesis, therefore, a thorough analysis of the use of TDOAs will be presented in a separated section.

2.4 TDOAs in deep

The first work using TDOA alone and without any knowledge about the locations of the microphones was carried out in (Ellis & Liu, 2004). Although error rates were considerably worse than with acoustic features the results were promising. In (Pardo, Anguera, & Wooters, 2006) the TDOAs were first successfully applied in combination with MFCCs, to a diarization system. Since then they have become very popular, being one of the most commonly used feature in MDM diarization.

The TDOA features aim to differentiate the speaker in the room using the time difference of arrival of the voice to one microphone against any other. TDOA features do not provide any speaker identity information but locate the source of the voice, therefore, any movement of the speakers around the room would degrade the effectiveness of the features, unless an effective tracking algorithm was applied. If there

is no movement of the speaker there are still other main shortcomings of the TDOAs derived from the use of distant microphones. Noises and reverberations in the recordings can be high enough to degrade the estimations of the delays. In (Evans, Fredouille, & Jean-François, 2009) a method was presented to improve the inaccurate estimates of delays and increase speaker separation.

TDOAs are used in combination with MFCCs. This combination requires a weighting between the two kinds of features, in order to take decisions in the segmentation and clustering level. These weights are tuned usually by hand and set at a fixed level based on the results of a set of meetings used for development. Although nowadays a large number of systems work with a weighting at the log-likelihood level of the GMM models (Wooters & Huijbregts, 2008), (Pardo, Anguera, & Wooters, 2006), (van Leeuwen, 2008), there are some studies exploring more robust solutions against variations across data, as the Information Bottleneck (IB) based Speaker diarization presented in (Vijayasenan, Valente, & Boulard, 2011). In this work the authors assess the problem of having features with a fixed dimension (for example 24th dimension MFCCs) versus features with various dimensions (TDOA features can have as many dimensions as pairs of microphones in the room, and this number varies largely from one meeting to another). Applying a linear log-likelihood combination to decide the clustering would mean to treat equally meetings with different levels of information. They propose a combination made at the level of relevant variables instead of log-likelihoods, making the system less sensitive to the dimension of the TDOA features.

In addition, some works have explored the possibility to automatically set the weights applied to the features (usually MFCCs and TDOA) on the basis of some metrics taken from the current recording. One of these methods is the entropy-based metric used in (Anguera X. , Wooters, Pardo, & Hernando, 2007), and a variation studied in (Vijayasenan, Valente, & Boulard, 2011), but results have not improved the fixed weights set empirically from an independent dataset.

Instead of using TDOA and MFCCs for all the process, some systems use the TDOA only for initialization, as the work presented in (Sun, Nwe, Ma, & Li, 2009), where the TDOA features are used only for the initial segmentation of the speakers. This use of the

TDOA along with a cluster purification technique yielded good results in the NIST Rich Transcription evaluation of 2009.

2.4.1 TDOA estimation

Delays between microphones will differ from one pair to any other, but also from one moment to any other in the recording. The voice will come from different points of the room in different times of the meeting, either because the sound source change (different speaker located in a different place) or because the same source has moved (same speaker that walks to other part of the room). The propagation time of the sound to the microphones will vary along the recording, which ultimately means that one channel (microphone) will not be delayed a fixed number of seconds against any other, but will have different delays, with respect to another channel, in different moments of the meeting.

In order to estimate the TDOA several techniques have been proposed to date, like the use of LMS adaptive filters used in sonar (Reed, Feintuch, & Bershad, 1981) or (Lin & Chern, 1998), however, the most popular techniques nowadays are those based on the cross correlation of the signals. The cross correlations of two signals is defined in equation (2-4) but to be able of work with finite signals of length N , it is usually estimated as in equation (2-5).

$$R_{x_1x_2}(d) = E[x_1(n) \cdot x_2^*(n - d)] \quad (2-4)$$

$$\hat{R}_{x_1x_2}(d) = \sum_{n=-N}^N x_1(n) \cdot x_2^*(n - d) \quad (2-5)$$

To easier the calculus, the signals are usually Fourier transformed, then the product is computed and the inverse Fourier transform is applied to the resulting signal.

As already introduced, the use of distant microphones favours the presence of noises or reverberation. Addressing this problem the Generalized Cross Correlation (GCC) was implemented in (Knapp & Carter, 1976). The general expression of the GCC for two signals x_1 and x_2 is:

$$R_{x_1x_2}^{GCC} = \mathcal{F}^{-1}(X_1(\omega) \cdot X_2^*(\omega) \cdot \psi(\omega)) \quad (2-6)$$

Where $\psi(\omega)$ is a frequency domain weighting function included to make the measurement of the cross correlation more robust to disturbing factors. There can be many options for the ψ function, including $\psi = 1$ with which the standard cross correlation formula would be obtained.

In (Knapp & Carter, 1976) five weighting functions are compared: the Roth processor (Roth, 1971), the Smoothed Coherence Transform (SCOT) (Carter, Nuttal, & Cable, 1973), the Phase Transform (PHAT) (Knapp & Carter, 1976), the Eckart Filter (Eckart, 1952) and the HT Processor (Hannan & Thomson, 1973) (similar to the Maximum Likelihood (ML) correlation).

According to (Anguera X. , 2006), in environments with high reverberation, the PHAT weighting is the most appropriate as it normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the cross correlation. The GCC-PHAT achieves very good performance when the SNR is high although it deteriorates when the noise level increases. The Generalized Cross Correlation with Phase Transform (GCC-PHAT) has been broadly used to calculate the TDOAs in meeting diarization, and it is also the method utilized in this thesis. The equation (2-7) shows the PHAT weighting function.

$$\psi_{PHAT}(\omega) = \frac{1}{|X_1(\omega) \cdot X_2^*(\omega)|} \quad (2-7)$$

To consult the ψ function for all the rest of the methods mentioned before see (Knapp & Carter, 1976).

Summarising, the GCC-PHAT of two signals $x_i(n)$ and $x_j(n)$ is:

$$\hat{R}_{PHAT}^{i,j} = \mathcal{F}^{-1} \frac{X_i(f) \cdot [X_j(f)]^*}{|X_i(f) \cdot [X_j(f)]^*|} \quad (2-8)$$

Being $X_i(f)$ and $X_j(f)$ the Fourier transforms of the original signals, $\mathcal{F}^{-1}$ the inverse Fourier transform and $[]^*$ is the complex conjugate.

The TDOA will be the delay associated to the maximum cross correlation, as defined for the following equation.

$$\hat{d}_{PHAT}(i, j) = \underset{d}{argmax} (\hat{R}_{PHAT}(d)) \quad (2-9)$$

However, in some occasions to use the delay associated to the maximum cross correlation could not be the most desirable. This maximum cross correlation could be due to the synchronization of external noises instead of the voice of one speaker. Also, if two speakers are speaking at the same time the delay calculated can alternate very fast from the value of one speaker to the other, and vice versa. For this reason usually the N best relative maximums are estimated, and decide the best option among these, once all the possibilities have been calculated. The post processing includes a filter to discard any noisy delay, i.e. any delay whose value of GCC-PHAT is below a threshold, and a Viterbi decoding to maximize speaker continuity, avoiding the switching between speakers that could take place for instance, if there is overlapped speech. For more detailed information consult (Anguera X. , 2006).

2.4.2 Use of TDOAs

As we said before, TDOAs are nowadays used in most MDM diarization systems. This information improved highly the performance of the diarization systems. We could talk about two main ways of using these new features:

1. **To create an enhanced audio signal.** Having more than one recording of the same meeting the system can combine two or more signals to create a better one, with highest SNR. The signals have to be delayed appropriately before being added. The estimation of delays between channels is unavoidable to be able of create an enhanced audio signal. This process is usually known as beamforming.
2. **To locate speakers.** Different distance from the source of sound (the speakers) to the microphones, mean different times of arrival. If one sound arrives at one microphone some milliseconds before it arrives at another microphone it means that the source is closer to the first microphone than

to the second. Assuming that the speakers do not move from their location, the TDOAs can aid differentiate the speakers, and, therefore, discriminate who is speaking (it could determine the exact location of a speaker if the position of the microphones in the room is known).

2.4.2.1 *Creating an enhanced audio signal: Beamforming*

When multiple microphones are available the system can use only one of them or try to use all of them. Multiple microphones placed in different positions in a room will not record all the voices or sounds with the same quality. One possible solution is to select the most centered microphone or the microphone with the highest quality (i.e. highest SNR). The sum of all the signals to create a new one with better quality is the other approach. A direct sum of the channels would decrease the quality of the overall signal as every microphone would have recorded the same sound in different moments (according to the longer or shorter distance from the source of the sound to each microphone). TDOAs can be used to synchronize the audio channels, being then possible to sum them obtaining an enhanced signal.

A popular approach is the Beamforming system published in (Anguera X. , 2006). A brief explanation of the process is presented here.

As every microphone is placed in a different position, and the source of the sound (voices or other noises) is changing throughout the whole meeting (people speaking from different points of the room), some microphones might record a very useful signal in some points in time and a much less informative signal in others. The system described in (Anguera X. , 2006) implements a system of weights that varies from one segment to the next. The initial weight applied to any segment c in a channel m is:

$$W_m[c] = \begin{cases} \frac{1}{M}, & c = 0 \\ (1 - \alpha) \cdot W_m[c - 1] + \alpha \cdot \overline{xcorr}_m[c], & otherwise \end{cases} \quad (2-10)$$

Being α the adaptation rate that has been set empirically to 0.05 (Anguera, Wooters, & Hernando, 2007), M the number of channels available, and $\overline{xcorr}_m[c]$ the average correlation of channel m to the others.

An improved method will consist of the study of all the segments of every recording to detect if some of them have too low quality to be included in the summed up signal. Note that the system has no information a priori about the quality of the microphones and may happen that some of them would only degrade the resulting signal if they were included in the final sum. To decide if a segment from a particular channel is suitable to be added to the rest of segments from the rest of channels the algorithm calculates the average cross correlation between the channel m and the rest of the channels ($\overline{xcorr}$). If the average $\overline{xcorr}$ for the channel m and segment c is lower than a threshold (set to $1/4M$ in (Anguera, Wooters, & Hernando, 2007)), the segment is discarded, i.e. its weight is set to 0.

$$\overline{xcorr}_m[c] < \frac{1}{4M} \Rightarrow W_m[c] = 0 \quad (2-11)$$

After having set to 0 all the segments which should be discarded the weights are recomputed, making the sum equal to 1.

The remaining segments are then prepared to be used to create the enhanced signal. All the segments are then delayed with respect to one of the signals according to the values of TDOA computed before and summed. A direct sum would cause discontinuities in the limits of the segments. Therefore, the segments are smoothed in the borders by the application of a triangular window as it shown in the Figure 4 (see (Anguera X. , 2006)).

In the standard implementation the analysis window overlaps 50% with the segment window, which agrees with the triangular overlap of 50% overlap done. In (Anguera X. , 2006), the analysis window is set to 500ms and segment is 250ms. After all samples from both overlapping windows are summed an overall weighting factor is applied to ensure that the dynamic range of the summed signal is optimally matched with the available dynamic range of the output file.

This enhanced signal can be then used as a unique audio signal and extract any necessary feature from it. This brief explanation is explained in depth in (Anguera X. , 2006).

Figure 4: Sum of every channel of a meeting recording to conform the enhanced signal.

2.4.2.2 Locate/discriminate Speakers. TDOAs in segmentation and clustering stage

Current diarization systems use TDOA to improve the performance of the segmentation and clustering stage but usually they do not take advantage of the TDOA information between every pair of available microphones. In most cases a reduced set of combinations is computed. Usually a reference channel is chosen and then the TDOA is computed between this channel and all the rest. As in diarization there is no information about the quality or the location of the microphones the selection of the reference channel is blind, and the system could either choose any of the available channels, or follow some criteria for the selection, focusing for instance on selecting the most centered channel or the channel with the best SNR. The target is to point out the best channel, on average, to estimate the delay between it and the rest. The technique most widely used is based on the average value of the cross correlation (GCC-PHAT), that was explained before.

The shortcoming of this technique is that the system will discard any information belonging to two non-reference channels. One specific pair of channels could be more suitable for locating one particular speaker than one pair made up of the reference and

any remainder microphone(s). Selecting appropriately a set of pairs could help to better locate all the speakers. The issue is to decide which pairs are the most representative and reliable for the system to choose.

Some previous works have been published to address this problem. We could classify the different approaches into the following groups:

- 1 **Selection of a single reference channel:** the system chooses the most reliable channel on average and extracts the TDOA features between it and the remaining channels. This method is consistent with the Beamforming procedure, as all the channels are delayed and added to the chosen reference one, which would be the channel that maximizes, on average, a predefined property, and therefore could be regarded as the most reliable. This procedure has the obvious shortcoming that even though most pairs will be of sufficient quality for the task, which implies lower expected errors in the estimation of delays, there could be some pairs with very poor quality that degrade the overall performance. Furthermore, the system will always work with all the channels, which in the case of meetings with a large number of microphones, could increase computational time without any improvement in performance. The other main drawback derives from the fact of choosing a reference channel. Only the delays between the reference channel and the remainder channels are used, eliminating any pair not involving the reference channel, even if this pair is more useful than any of the chosen ones. An example of this group is our baseline system, which will be explained in detail in the following section. The baseline method used in this thesis follows this idea (Pardo, Anguera, & Wooters, 2007).

- 2 **Use of all possible pairs of channels:** these methods do not actually perform any selection; instead, the system uses all the information available in extracting TDOA features for every pair of microphones. They extract TDOA features for every pair of microphones. The advantage of these methods is that no information is lost. Every possible pair of

channels is selected, which means that any relevant information hidden in any pair of channels will be taken into account in later stages. Of course, this also means that any inaccurate information will be also used. Another disadvantage of this method is the increase in the computational time required to calculate delays for every combination of channels, especially if the number of microphones is very large.

As the delays extracted are used in the later stages of the diarization, a very large feature vector would also increase the overall computational time. In this case, the dimension of the TDOA vector is more important than ever. Choosing every combination of channels would mean a TDOA vector dimension of $(M*(M-1))/2$, being M the number of microphones in the meeting room. As some meetings have a very large number of microphones, which would mean a much larger TDOA vector dimension, it has become common to apply a dimensionality reduction to the resulting TDOA vector. The works presented in (Otterson S. , 2007), (Evans, Fredouille, & Jean-François, 2009) and (Vijayaseenan & Valente, 2012a) are examples of this group.

- In (Otterson S. , 2007) the values of single-reference correlations are analysed. Microphones in the room can be widely separated, which could make them be subject to significantly different reverberation effects. All microphone pairs are used for computing correlation features, followed by a principal component analysis (PCA) to reduce dimensionality. However, in this work, experiments related to these new features were not as successful as anticipated, obtaining DER improvements only in one of the two sets of meetings evaluated. This work also includes experiments with features related to the energy ratio of the signal, which, in the end, yielded the lowest DER.
- An alternative approach is presented in (Evans, Fredouille, & Jean-François, 2009). Here, the authors compute delays between all possible pairs of channels and then apply a variant of linear discriminant analysis (LDA), known as unsupervised discriminant

analysis (UDA) (see (Yang, Zhang, Jin, & Yang, 2006)), to seek a projection in which the inter-cluster distance is maximized and the intra-cluster distance minimized. Experiments using PCA are also carried out as a way of comparing it with the UDA method.

- In (Vijayasenan & Valente, 2012a), TDOAs are calculated for all pairs and then another kind of linear dimensional reduction, the Karhunen-Loève Transform (KLT), is applied. The authors compare performance in three situations: when the target dimension of KLT is fixed and equal to 7, when this dimension is variable and equal to the number of channels minus one and when no dimensionality reduction is applied. The KLT, with either the fixed or the variable target dimension, clearly reduces the error rate compared to the system that uses all the TDOAs. In a previous work: (Vijayasenan & Valente, 2012b), the authors proved that using TDOAs from every possible pair (without any dimensionality reduction) was better than using TDOAs from all the channels with the reference one.

3 Selection of a set of pairs of channels following some criteria: the system selects a group of pairs of channels as the most representative and reliable, and then computes only the TDOA for each selected pair. The main advantage of this method is that only good pairs are used and no pair is discarded initially. If chosen properly, the system will work only with the best information available and no relevant information will be discarded. It is also important to mention that the number of pairs selected can be very low, reducing the dimension of the final TDOA vector considerably and speeding up the whole process significantly. On the other hand, the selection requires either a number of pairs to be chosen, or a stopping criterion to be defined. If the number of pairs is too high the system would work slower than desired, even without a significant gain in performance. However, if the number of pairs is too low the resulting TDOA vector could lack crucial information.

- In (Sun, Nwe, Ma, & Li, 2009), a method for selecting the pairs of microphones used to calculate the TDOAs and the use of these TDOAs in the first stage of the segmentation and clustering module is presented. The selection takes into account the dynamic range of the TDOAs for each pair of channels, choosing up to 5 pairs with the highest dynamic range. The highest peaks of the histogram of delays of these 5 pairs supposedly will locate the speakers (provided they do not move). The authors succeed in reducing the error of the system considerably by applying this technique alongside other modifications in different stages of the system. Though no experiments are included solely with the modification related to the selection of pairs, the overall improvement makes this approach very interesting. Previous works using similar techniques were published in (Nwe T. L., Sun, Li, & Rahardja, 2009), (Koh, et al., 2008) and (Nguyen & Sun et al., 2009)

4 Selection of a set of channels: this method is usually applied to improve the result of the acoustic fusion module. When several channels are available we can delay and add some of them together to create an enhanced audio signal (composite signal). Instead of adding every available channel together, these techniques imply the selection of a reduced, more reliable, group of channels, discarding those channels that would only decrease the quality of the composite signal. In this method a group of channels is selected that either individually or grouped together maximize a predefined property. In contrast, methods in the previous category focus on selecting pairs of channels (instead of only channels) that fulfil one property designed to guarantee that its associated delays are either very useful or very reliable. The channel selection method can guarantee that those particular channels are of high quality on average, but not that two particular channels of that selection would form a good pair. Used for beamforming, this method

would suppress the unreliable channels, which is an improvement on the strategy in which every channel is added to the final signal regardless of its quality. The main problem is again to decide the optimal number of channels to work with. As regards the gain in computational time, the beamforming stage is not as costly as feature extraction and the number of microphones available in a meeting is not usually very high (therefore the number of channels discarded are not either). However, if this number were much higher as in (Yang & Jiu, 2014), and the number of microphones finally used were much lower, then there could be some slight improvement in time cost. Still, the signal formed in this part will be used for the extraction of all the acoustic features and a very narrow selection could degrade the performance severely, without significant gain in computational time. As examples of this fourth group we could include the works in (Kumatani, McDonough, Lehman, & Raj, 2011) and (Vijayasenan, Valente, & Boulard, 2009).

- In (Kumatani, McDonough, Lehman, & Raj, 2011) a channel selection based on multichannel cross-correlation coefficients (MCCC) is applied to improve distant speech recognition. In contrast to the usual situation in diarization, the authors consider the situation in which microphones are regularly spaced and whose positions are known *a priori*. The channels with the highest MCCC are used to create an enhanced composite signal.
- Another work, whose objective is to improve the composite signal, is presented in (Vijayasenan, Valente, & Boulard, 2009). In this case the mutual information from N groups of channels is computed, and those with the highest mutual information are used in the beamforming stage.

5 Hybrid systems. Finally, we could talk of systems that combine some of the other approaches, in particular the approaches that select a set of channels or a set of pairs of channels (groups 4 and 3 of this list), with any

of the rest. In the first place these systems would select the best pairs according to their quality, then any of the other methods would be applied. This hybrid technique has been usually applied to mix the approach in group 3, in which a set of pairs is selected, with the systems describes in group 2, in which all pairs are selected followed by a reduction of dimensionality. The shortcomings of these methods were mainly two: the fact that every bad pair in the meeting would be used and the large time required to extract delay values for every pair in meetings with many channels. To reduce these effects some works carry out a very broad pre-selection of channels. Some pairs, and the information associated to them, are then lost, but the time savings could be very high. The computational time will be reduced severely in meetings with a large number of channels, which is quite desirable for the systems in which every possible pair combination is computed. Besides, if the pre-selection is done properly, most of the discarded pairs would be of bad quality. An example of these systems is (Yang & Jiu, 2014).

- In (Yang & Jiu, 2014) the authors design an experiment to analyse diarization in the case of a large number of microphones. The authors used 64 microphones to record the speakers. They select from 8 to 64 of these microphones and then compute TDOA for the whole set of possible combinations, followed by a reduction in dimensionality. Experiments are carried out for 5 different dimensionality reduction approaches. In this case the microphones used are selected randomly. Results show that, in this situation, the error is lower when the number of microphones used is larger.

In Table 1 we summarise the pros and cons of all the approaches for the selection of TDOA features which were fully described before.

System pair selection	Pros	Cons	System Examples
Single reference channel-to all the rest	Consistent with the beamforming procedure Most chosen pairs will be of good quality	Possible chosen pairs of very poor quality. Loss of any information included in any non-reference pair. Large TDOA vector when many microphones available.	(Pardo, Barra-Chicote, San-Segundo, Córdoba, & Martínez-González, 2012) (Friedland, et al., 2012)
All pairs combinations + dimensionality reduction	No information lost.	Very high computational time. Possible chosen pairs of very poor quality. Necessity of dimensionality reduction.	(Otterson S. , 2007) (Evans, Fredouille, & Jean-François, 2009) (Vijayasenan & Valente, 2012a)
Selection of pairs of channels	The system works only with the best data Low dimension vectors Reduced computational time.	Necessity to define the final number of pairs.	(Sun, Nwe, Ma, & Li, 2009)
Selection of channels	The system works only with the best channels	Necessity to define the final number of channels	(Kumatani, McDonough, Lehman, & Raj, 2011) (Vijayasenan, Valente, & Bourslard, 2009)
Hybrid systems: Pre-selection + All pair combinations + dimensionality reduction	Very low loss of information.	High computational time. Necessity of pre-selection method. Possible chosen pairs of very poor quality. Need for dimensionality reduction.	(Yang & Jiu, 2014)

Table 1: Summary of pros and cons of different channels selection methods

In contrast to the approaches described in groups 1, 2 and 3, which are used to improve the quality of the TDOA vector or maximize its discriminative information, the methods grouped in the fourth category aim to create the best possible composite signal. Although the target is different, the techniques applied to the latter type could inspire methods for the first three, since the determination of the best quality channels is desirable for both tasks. Furthermore, the selection of the best channels could be used as a good pre-selection for the systems that compute the delays between all the combinations of channels.

2.5 Overlapped speech

Overlapped speech is an important problem in diarization, whose effects have been studied in many works (Shriberg, Stolcke, & Baron, 2001), (Çetin & Shriberg, 2006). Most systems apply no specific technique to detect if two or more speakers are speaking at the same time. Any segment of the recording is assigned to only one speaker, which means that the time corresponding to the other speakers is always counted as an error.

Missed speech is not the only problem of the overlapped regions. The other drawback is that the segments of overlapped speech are taken into account to model the hypothetical speakers, degrading the purity of the models. Systems which do not detect overlap will assume that any segment of speech corresponds to only one speaker and will utilize all the segments to train the model of the speakers, creating less accurate models. The problem of not dealing with overlap is therefore that the models as well as the final performance can be severely affected if there is much overlap in the meeting.

To reduce the effect of the overlap there are two main approaches to the problem.

1. **Exclusion of overlapped regions.** Exclude the overlapped speech segments from the training of the speaker models. Usually this prevents the problem of the speakers' model accuracy. Afterwards these segments can be assigned

to one speaker according to its similarity to the clean trained models. Although, in theory, we could encounter a meeting where most of the time two or more speakers participate simultaneously (leaving very short or no time with clean usable speech), in actual meetings this situation is not so common and current diarization systems do not focus on solving this particular problem.

2. **Labelling the overlapped regions.** The next step to detection is the decision to assign these segments to the correct speakers. The assignment of these regions to more than one speaker is a difficult task that can degrade severely the overall performance of the system if done wrongly. Even if the regions are correctly detected in the first place, the system should then decide how many speakers participate in that part of the speech and also detect which ones.

There are studies that show that the exclusion of the overlapping segments in the training of the models can highly improve the performance of the diarization (Sinclair & King, 2013), (Huijbregts, van Leeuwen, & Wooters, 2012). The second action, labelling the overlapped regions, can improve it even more, but here the correct detection of the overlapped segments is crucial. If the system excludes for the training of the speaker model, a region that is not actually an overlapped section, the model can still be good enough, and the final segmentation as well. However, if the system assigns to all these false overlapped periods a second, or a third speaker, it would greatly increase the error of the system. In (Otterson & Ostendorf, 2007) it is shown that in a situation of perfect detection of overlapped segments even a simple strategy of labelling with the two nearest speakers (neighbouring speakers of the overlapped region) have the potential of reducing the diarization error.

In recent years many efforts have been done to solve the problem of overlap detection and labelling. In (Huijbregts, van Leeuwen, & de Jong, 2009), an overlapped speech model is trained using the 500 ms before and after every speaker change point (which were previously detected in a first diarization step). With this model the system localises the overlapped regions. Once overlapped regions are detected the system labels them with the two speakers before and after the detected overlapped segment.

In (Boakye, Trueba-Hornero, Vinyals, & Fiedland, 2008) and (Boakye & Vinyals, 2011) the authors analyse and compare a set of features that could be used to detect overlap. Their analysis suggest that the most discriminative features would be the short time root-mean-square (RMS) energy, the Linear Predictive Coding (LPC) residual energy, the spectral flatness, the harmonicity and the 12th order MFCCs.

In (Geiger, et al., 2012) the authors applied some energy, spectral and voicing related features as well as the Convolutional Non-Negative Sparse Coding (CNSC) features to detect and label overlapped regions.

In (Zelenak, Segura, Luque, & Javier, 2012) three cross correlation based spatial features are studied with good results: the coherence value, the dispersion of this coherence value and the delta of TDOA.

Overlap in meetings recorded using lapel microphones is investigated in (Yokoyama, Nasu, Iwano, & Shinoda, 2013). In this work two novel features are presented: speech power after cross-channel spectral subtraction and the so called “amplitude spectral cosine correlation coefficient” which calculates the cosine correlation between the spectrum of two microphones. Despite the improvement of these features in comparison to the previous ones (raw speech power and power spectral Pearson’s correlation coefficient) the authors pose the addition of some other features, as entropy, to reduce the misdetected frames.

In (Charlet, Barras, & Liénard, 2013) two overlap detectors relying on standard cepstral features and multi pitch analysis are designed and compared with recordings of broadcast news and debates. It is also proposed a variation of the two nearest labelling strategy. The detected overlapped segments are assigned to one speaker (the nearest) and then to the second nearest only if its temporal distance to the segment is below a given threshold.

Other side effects of overlapped speech can be used as well in overlap detection. For instance, the possibility of applying the distribution of silence to improve the detection of overlapped speech is addressed in (Yella S. H., 2012) and (Yella & Boulard, 2014). The authors proved that there exists an inversely proportional relation between

the probability of presence of silence segments and the presence of overlapped speech. We could mention also the study carried out in (Geiger, Eyben, Evans, Schuller, & Rigoll, 2013), where the authors state that during overlapped segments, mainly in short interruptions, some words are more likely than others, and this information is used to improve an overlapping detector.

Finally, we could mention the work presented in (Dighe, Ferràs, & Bourlard, 2014). The authors propose a new approach to model the overlapping segments based on the Vector Taylor Series (VTS) framework which has been previously used to model noisy speech for the automatic speech recognition task. VTS system requires single speaker models to create the overlapped model and this initial work uses the oracle segmentation to train these models. Though some improvements are reported, the authors recognise that the system relies heavily on the purity of the single speaker models and that further work will focus on improving the diarization output so that purer models can be obtained for use with VTS.

3 *Baseline Diarization System*

The diarization system used as baseline in this thesis follows a bottom-up approach. This means that first it segments the recording into several hypothetical speakers and then, it reduces this initial number until a stopping criterion is reached. But before beginning the actual diarization several steps are needed as the cleaning of the audio signals or the extraction of features. In Figure 5 a general architecture of the diarization system used is included.

Figure 5: Baseline system architecture

As could be seen, at the beginning of the process, the signals are filtered independently applying a Wiener filter to reduce noise (Wiener, 1949). The implementation of Wiener filtering is taken from the ICSI-SRI-UW system used for ASR in (Mirghafori, et al., 2004), and applied to each channel. This implementation performs an internal speech/non-speech and noise power estimation for each channel

independently, using an analysis window of 20ms. The algorithm is based on the wiener filter used in (Adami, et al., 2002), which used a noise over-estimation factor as a function of the local a posteriori SNR, smoothing of the filter response and a spectral floor parameter to avoid negative or very low transfer function components. The single noise spectral estimate for each meeting waveform was calculated over all the frames judged to be non-speech by the voice-activity detection component. The clean speech signal is then obtained by the application of the designed wiener filter to the noisy speech signal. Full details can be found in (Adami, et al., 2002).

Afterwards, an automatic estimation of the reference channel is applied. This is a process which aims to perform an unsupervised selection of the most centered and with best audio quality channel, which would allow a better estimation of the delays between this channel and the rest of the microphones.

We usually work with more than one recording for the same meeting. In this situation the first step of the system will be to create an enhanced signal using all the recordings available. This process selects a reference channel according to its average cross correlation (see equation (3-1)) and then, as explained in section 2.4.2.1, calculate the TDOAs of the rest of the channels with respect to this reference and delay and sum all the audio recordings.

$$\overline{xcorr}_i = \frac{1}{K(M-1)} \sum_{k=1}^K \sum_{j=1, j \neq i}^M xcorr_{i,j}(k) \quad (3-1)$$

This formula shows the average cross correlation ($\overline{xcorr}$) of a channel i against all the rest ($j=1, \dots, M, j \neq i$), where M is the total number of channels, K is the number of blocks used ($K=200$ blocks of one second for the determination of the reference channel in the baseline system) and $xcorr_{i,j}(k)$ is the cross correlation between two channels i and j , for each block k . The channel with the highest average cross correlation will be selected as the reference channel. This method aims to select the channel with best quality or best situated against the others. As every meeting recording could come from a different room with different microphones the selection of the reference channel will be carried out for every new meeting analysed, and kept constant for the whole meeting.

In (Anguera X. , 2006) an alternative method to select the reference channel based on the SNR was also evaluated, but results were not conclusive and the cross correlation selection was preferred.

After the selection of the reference microphone and prior to the delay&sum of the channels, every channel is multiplied by a factor, which will be estimated for each of them, to normalize its amplitude. As the meeting could have been recorded with different types of microphones, it is common to found differences in the characteristics of the signal being recorded. Also when two microphones are far from each other, the speech they record will be affected by noise of a different nature, due to the room's impulse response, and will have different amplitude depending on the position and relative orientation to the microphone of the speaker talking. This issue is addressed by weighting each channel in the delay&sum processing. The weights are adapted continuously during the meeting.

To carry out this, it is necessary to find the maximum of every channel, but avoiding sounds of high energy and short in time, like door slamming, etc. To assure that the normalizing value is not affected by these non-speech sounds of high energy, the system calculates the average of the maximums of several windows along the recording which constitutes the normalizing value used for the channel. More details can be consulted in (Anguera X. , 2006).

TDOAs are also estimated between the reference channel and the rest of the channels. This calculus is carried out with the GCC-PHAT technique. As explained before, the PHAT weighting is the most appropriate as it normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the cross correlation. For each pair of microphones the system will extract a delay value for each frame. The length of the frames used varies in accordance to the task at hand. The selection of the reference channel calculates xcorr for 200 blocks of 1sec of duration (see equation (3-1)). On the other hand, the delays used as input feature for the agglomerative and clustering stage are calculated for every frame of 10ms using window size of 500ms, and delays used in the composition of the enhanced signal are calculated only each 250ms with the same window of 500ms. Remember that, to be able of working

with only one enhanced signal, all the available channels must be summed, but for doing so, they have to be delayed prior to be added to the reference channel (Anguera, Wooters, & Hernando, 2007).

MFCCs are then extracted from this enhanced signal (we calculate 19 components for every frame of 10ms with a window of 30ms, as it was used in previous works (Pardo, Anguera, & Wooters, 2007)),

The composed signal is also processed by the VAD module, which is a hybrid energy-based detector and model-based decoder (Anguera, Aguiló, Wooters, Nadeu, & Hernando, 2006), in order to extract the information about the speech/non-speech regions.

In the baseline system used in this thesis, MFCCs, delays, and the speech/non-speech regions are the only features used though, of course, any other features could be extracted.

The system then initiates the final stage in which the most probable number of speakers and their moments of speech are listed. This step begins with the process of segmenting the speech in K equally long segments, being K the number of hypothetical speakers that have to be higher than the expected final number of participants. Empirically this number has been set to 16. With these K initial clusters the system trains GMM speaker models with diagonal covariance matrixes, based on the MFCCs and the TDOAs separately. These trained models are used to resegment the signal according to its similarity to these models. Every cluster is modelled using a Gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for [mfcc] and 1 for [tdoa] streams as was set in (Pardo, Anguera, & Wooters, 2007)). Also, the algorithm has been designed to force a minimum number of consecutive frames assigned to one cluster to avoid very short unrealistic utterances. This number of frames imposes then a minimum duration in the speaker turns. The kind of meeting can vary heavily the optimal minimum duration, as according to the meeting styles the speakers could tend to interrupt each other or speak for long periods (as family meetings vs lectures). This minimum duration was set empirically to 250 frames of 10ms (2.5 secs) but can be adjusted with the database.

Then, the system starts an iterative process where new models are trained with the recently resegmented clusters. The two most similar according to BIC criterion are joined and a new model is trained for it, then, a resegmentation is performed with the new models. The process continues until the stopping criterion is reached. The general diagram of this process is represented in Figure 6.

Figure 6: Diagram of the training and segmentation process of the baseline system

The process of decoding where the speaker changes will be (in segmentation stage of Figure 6), is the following. After the initial segmentation in a predefined number of clusters and their initial training of the GMM models for each speaker, takes place the resegmentation. The system uses an ergodic Hidden Markov Model (HMM) where each state corresponds to each cluster and then performs a Viterbi search to find out the new best sequence of turns according to the likelihood of each frame to belong to each speaker.

The training of the speaker models necessary for the segmentation and the cluster comparison use the features previously calculated. As we work usually with two features (MFCCs and TDOAs), there will be always two models calculated, one according to the MFCCs and another to TDOAs. Any decision as to join or not two clusters has to be taken using both models. The system will join the loglikelihoods of the two models according to the following equation:

$$\log(\mathcal{L}(x[n], y[n] | \theta_1)) = \alpha \log(\mathcal{L}(x[n] | \theta_{1x})) + (1 - \alpha) \log(\mathcal{L}(y[n] | \theta_{1y})) \quad (3-2)$$

Where θ_1 is the model compound of cluster 1, θ_{1x} is the model trained using only MFCCs and θ_{1y} is the model trained with only TDOAs. The x stream corresponds to the MFCC coefficients and the TDOA features are treated as the y stream. The parameter α in this equation represents the weight applied to every model, and has to be set in advance by the user. In the initial system this parameter is set to 0.9.

The ΔBIC shown in equation (2-3) will be used then to determine which clusters should be joined. The two clusters with the highest positive ΔBIC (equation (2-3)) will be merged. Loglikelihoods are computed having into account MFCCs and TDOA, combining them as in equation (3-2). The number of components for the joined GMM model will be set as the sum of the components of the separated cluster models. For instance, as the initial number of components for the MFCC model is 5, the first merging of clusters will result in a joined cluster trained with 10 components.

The comparison between every pair of active clusters is performed and only the two most similar (i.e. the two with the highest positive ΔBIC value) are joined. If no pair would result in a ΔBIC higher than 0 it would mean that no active cluster is similar enough to any other to be considered the same. Then the process ends.

More detailed information about the baseline diarization system that will be used in this thesis can be found in (Pardo, Anguera, & Wooters, 2007).

4 Databases and evaluation measures

In this section the databases used for experiments are presented with a summary of their characteristics. Also the measures usually utilized to evaluate the results in diarization systems are presented.

4.1 Databases

- NIST Rich Transcription Evaluation databases (RT). The National Institute of Standards and Technology has conducted from year 2002 to 2009 a series of competitions in diarization technologies. In this work a subset of the NIST Rich Transcription of 2002-2005 sets, the RT06 set and the RT-07 has been used as the development set. For the evaluation set, we have used RT-09 (Rich Transcription Evaluation Project, National Institute of Technology (NIST), 2002-2009,, s.f.). A subset of 12 meetings from RT02, RT04 and RT05 (also used previously and named devel06 in (Pardo, Anguera, & Wooters, 2006)) together with RT06 and RT07 (hereinafter called –RT-DEVELSET-) is made up of more than eighteen hours of audio data divided into twenty-eight different meetings, and RT-09 comprises more than five hours of audio data divided into seven different meetings. All the meetings included are in English. The segments defined by NIST for the official evaluations have been used to measure the performance of the systems described in this work. These segments are defined in the UEM file provided for the evaluation. They consist of 27,612.64 seconds (2,761,264

frames of 10ms) for 28 meetings of the –RT-DEVELSET- set and 10,858.49 seconds (1,085,849 frames of 10ms) for the 7 meetings of RT-09, which are taken into account to calculate the statistical significance of the results. NIST database includes meetings with a number of microphones ranging from 2 to 24, and a number of speakers from 3 to 9. The list of meetings used and their characteristics are listed in Table 2.

	Set	Meeting	# of microphones
1	RT-DEVELSET	AMI_20041210-1052	12
2		AMI_20050204-1206	16
3		CMU_20050228-1615	3
4		CMU_20050301-1415	3
5		ICSI_20000807-1000	6
6		ICSI_20010208-1430	6
7		LDC_20011116-1400	8
8		LDC_20011116-1500	8
9		NIST_20030623-1409	7
10		NIST_20030925-1517	7
11		VT_20050304-1300	2
12		VT_20050318-1430	2
13		CMU_20050912-0900	2
14		CMU_20050914-0900	2
15		EDI_20050216-1051	16
16		EDI_20050218-0900	16
17		NIST_20051024-0930	7
18		NIST_20051102-1323	7
19		VT_20050623-1400	4
20		VT_20051027-1400	4
21		CMU_20061115-1030	3
22		CMU_20061115-1530	3
23		EDI_20061113-1500	16
24		EDI_20061114-1500	16
25		NIST_20051104-1515	7
26		NIST_20060216-1347	7
27		VT_20050408-1500	4
28		VT_20050425-1000	7
29	RT09	EDI 20071128-1000	24
30		EDI 20071128-1500	24
31		IDI 20090128-1600	8
32		IDI 20090129-1000	8
33		NIST 20080201-1405	7
34		NIST 20080227-1501	7
35		NIST 20080307-0955	7

Table 2: List of meetings for the RT database used in the experiments

- C-ORAL-ROM (Moreno-Sandoval, et al., 2005) database. This corpus is a multi-language and multi-style database covering a wide spectrum of formal and informal speaking styles, in public and private situations. All the languages included are Romance (French, Italian, Portuguese and Spanish), with styles ranging from formal to informal, extracted either from the media or from private spontaneous natural speaking. These data have been extracted from media broadcasts of different stations, and they present a great deal of variability in the recording environments and a high number of speakers (more than 200). This results in some speakers uttering only a few short sentences, making them almost irrelevant from a statistical parametrical point of view. The number of speakers per session is variable (between 1 and 28 speakers). Table 4 summarises the average characteristics for each speaking style

To evaluate the implemented methods this database has been splitted into two, the development set and the test set. Both sets are composed of sessions from all the styles evaluated. Around of a third part of the database has been reserved to test experiments. The development set is composed of 27 sessions that add up 234.26 minutes, and the test set is composed of 14 sessions that add up 115.17 minutes.

The manual transcriptions of these sessions include speaker turns were we can find the speaker specified, but the segment includes also noises, silences or music (everything from the end of the previous speaker to the beginning of the next). To refine these references labelling speech only segments we have force aligned the speech with the text provided also in the transcriptions, using acoustic models trained from the spanish partition of TC-STAR – EPPS (European Parliament Plenary Sessions) and PARL (Spanish Parliament Plenary Sessions). Although the forced alignment helped highly to this task, it was not free from errors, and we had to correct manually some labels.

- TC-STAR database. This database of Spanish audio is composed of two partitions: the EPPS (European Parliament Plenary Sessions) and PARL (Spanish Parliament Plenary Sessions). The EPPS is made of 61 hours of audio recordings of European Parliament sessions from years 2004 to 2007. Although there are some native Spanish speakers, the speech come mainly from translators that traduce to the

Spanish what the other members of the parliament say in other languages. On the other hand, the PARL is composed of 38 hours of native Spanish speakers recorded from year 2004 to 2006.

	Set	Meeting	Style
1	C-ORAL-ROM-DEVELSET	Emedin02	Interviews
2		Emedin03	
3		Emedin04	
4		Emedmt02	Meteorological reports
5		Emedmt03	
6		Emednw02	News broadcast
7		Emednw03	
8		Emednw05	
9		Emednw06	
10		Emedrp01	Reportages
11		Emedrp02	
12		Emedrp04	
13		Emedrp07	
14		Emedsc01	Scientific press
15		Emedsc02	
16		Emedsc03	
17		Emedsp02	Sports
18		Emedsp04	
19		Emedsp05	
20		Emedsp06	
21		Emedts03	Talk shows
22		Emedts04	
23		Emedts06	
24		Emedts07	
25		Emedts08	
26		Emedts09	
27		Emedts11	
28	C-ORAL-ROM-TESTSET	Emedin01	Interviews
29		Emedin05	
30		Emedmt01	Meteorological reports
31		Emednw01	News broadcast
32		Emednw04	
33		Emedrp03	Reportages
34		Emedrp05	
35		Emedsc04	Scientific press
36		Emedsp01	Sports
37		Emedsp03	
38		Emedts01	Talk shows
39		Emedts02	
40		Emedts05	
41		Emedts10	

Table 3: List of meetings for the C-ORAL-ROM database used in the experiments.

Style	# sessions	#spk/session	Time/session
Interviews	5	2-4	7-9 min
Meteorology	3	1	2-3 min
News	6	5-10	7-9 min
Reportage	6	7-28	9-12 min
Scientific press	4	3-6	8-10 min
Sports	6	1-7	7-14 min
Talk shows	11	2-8	6-11 min

Table 4: Characteristics of the C-ORAL-ROM database per type of recording

- The AMI Meeting Corpus. (Carletta, 2006). This database consists of 100 hours of meeting recordings. The recordings use a range of signals synchronized to a common timeline. These include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic whiteboard. The meetings were recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers. We will only use a subset of the meetings recordings from a single site scenario with far field microphones. The meetings used are listed in Table 5.

AMI training set		AMI development set
IS1000b	IS1005b	IS1000d
IS1000c	IS1005c	IS1002d
IS1001d	IS1006a	IS1004d
IS1002b	IS1006c	
IS1002c	IS1007a	
IS1003a	IS1007b	
IS1003c	IS1007c	
IS1004a	IS1009a	
IS1004b	IS1009b	
IS1004c	IS1009c	
IS1005a	IS1009d	

Table 5: List of meetings for the AMI database used in the experiments.

4.2 Evaluation measures

For the evaluation of the proposed systems we will use the toolkit distributed in NIST Rich transcription evaluations to calculate the DER. This tool takes the hypothesis that returns the system and compare it with the true speaker segmentation. The hypothesis speakers are assigned to the real ones in order to maximize the time correctly assigned to that speaker. The tool counts as errors the time labelled as speech for one speaker that either belongs to another speaker or that is actually no speech. Also any of the time of actual speech not assigned to any speaker in the hypothesis will be counted as error.

Results are presented in terms of Diarization Error Rate (DER) for comparison with other systems but this term includes a fixed error of speech/non speech (SPNSP) that will not change from one experiment to another. The error that we will try to diminish is the Speaker Error Rate (SER), which takes into account only the errors of incorrect assignment of the speech segments to the true speaker. In the evaluation of any meeting we could talk about 5 kinds of errors that will be explained later, but in most experiments the only part of the error that will change is SER.

The DER, is actually $SER + SPNSP$, being $SPNSP = FA + MISS$. All these terms are explained below.

FA: False Alarms. This error occurs when a segment of the meeting is classified as speech being non speech, and therefore assigned to one of the hypothesis speakers. Notice that non speech includes silence but also noises. There could be also another type of FA error if the system detects overlap. If the hypothesis includes overlapped speakers it could happen that two speakers were detected while there is only one. In this case the FA in diarization would be:

$$FA_{SPEAKER} = FA_{SPEECH} + FA_{OVERLAP}$$

MISS: In the literature MISS errors are often defined as Missed Speech but it would be actually MISS speaker because this MISS error includes also the error due to

overlap. When we speak about MISS speech, this value of error includes the segments classified by the VAD as non-speech, being actually speech. If the system does not tackle the overlapped speech this MISS value includes the segments of the overlapped speakers not classified. This could lead to consider the same segment of recording twice or more times (as many as the number of overlapped speakers). The first run the evaluation considers if the system has classified the segment as non-speech belonging actually to one or more speakers. If that has happened the lost speech time would be included in the $MISS_{SPEECH}$ error. Then the evaluation tool would check if there is speech of more than one speaker in that period of time and if the system has detected it and assigned it to any speaker. If there is more than one speaker and the system has not assigned the audio frames to more than one speaker the overlapped speech is included in the $MISS_{OVERLAP}$

$$MISS_{SPEAKER} = MISS_{SPEECH} + MISS_{OVERLAP}$$

SPNSP: Speech/non-Speech error. This is just the sum of FA and MISS. As before, in diarization systems it will include not only the errors about the miss classification of speech or silence/noises, but the $FA_{OVERLAP}$ and $MISS_{OVERLAP}$. Most state of the art diarization systems have not solved the problem of overlap though many are working on it. If the system does not handle overlap, as the diarization program we work with in this thesis, the SPNSP error would remain invariant when experimenting with the last module of the architecture, the segmentation and clustering stage.

SER: Speaker Error Rate. This error rate takes into account only the errors due to the incorrect assignment of the speech segments to its correspondent speaker.

DER: Diarization Error Rate. This is the complete error of diarization, including MISS, FA and SER. In this thesis DER is included for comparison purposes, though SPNSP error does not change and DER is the sum of SPNSP and SER.

5 *Selection of delay features*

In this chapter we present different methods for selecting pairs of channels. The aim of these alternatives is to select the most discriminative or reliable pairs of channels to calculate their TDOA vector. The TDOA vector is composed of a value of delay associated to every frame of the meeting for each pair of channels analysed. The TDOA vectors extracted from the selected channels will be then used in the clustering phase. Not every possible pair of channels is suitable for the estimation of the position of the speakers. The poor quality of some microphones, their position in the room or the noises could lead to miscalculation of the delays associated, and, thus, the wrong location of the speakers. There is also another issue, sometimes the number of recordings available for one meeting is much higher than what we should need and working with too many features will increase the computational time. In these situations we encounter another reason to make a selection of channels: to speed up the process. We have experimented with different approaches to select a reduced number of pairs trying to maximize its utility for diarization.

We want to implement and contrast results with two hypothesis:

1. Use of techniques of selection of pairs of microphones taking into account all the possible combinations and selecting a reduced set of them according to some criteria.
2. Use as much information as possible, what would mean to use all the possible pair combinations, estimate delays between all the possible pairs and use this information for the segmentation stage.

The second technique has the problem of the computational time. The diarization process is very time consuming and this technique would increase heavily

the time required to finish the TDOA calculations. Also, a high increase in the number of parameters could make the system less robust as they may increase the complexity of the trained models. The number of combinations depends on the number of available microphones, and this number is different for each meeting of the database. The delay computed between two signals could be either t or $-t$, depending on which signal we consider x_i and x_j in equation (2-8), which means that we are interested in combinations of two microphones without repetition. The equation for combinations of n channels in groups of r members, without repetition is:

$$C(n, r) = \binom{n}{r} = \frac{n!}{r! \cdot (n-r)!} \quad (5-1)$$

Which translated to our particular situation with n the number of channels of the meetings and r equal to two (combinations of pairs of audio channels), would make the number of combinations equal to:

$$C(n, 2) = \binom{n}{2} = \frac{n \cdot (n-1)}{2} \quad (5-2)$$

In the RT database that we will be using, there are some meetings with only 2 or 3 microphones, which would mean 1 or 3 possible combinations of pairs respectively. With 2 microphones, as there is only one pair available, the selection method can only choose this pair, which is the same pair used by the baseline system. However, with three microphones, the baseline system would have to estimate TDOA's between only two pairs, the reference channel against the other two microphones, while the number of possible pairs combinations is 3. The increase in computational time for this case could be affordable, but it will be growing with the number of microphones. For the meetings with 16 or 24 microphones the baseline calculates delays for 15 or 23 pairs respectively, while the total number of possible combinations is 120 and 276 respectively. The time required to extract delays for all these combinations would be prohibitive. Nevertheless, the planning includes carrying out this experiment to check if there is enough improvement in the error rate to compensate the loss in time.

Selecting pairs of channels requires the prior definition of the number of pairs to be chosen. As mentioned, for meetings with large number of microphones may be compulsory to reduce computational time and memory use. Furthermore, the information that the TDOA values carry is the location of the speakers which is limited to three dimensions in space. It does not make sense to use a large vector to code such a reduced amount of information. To avoid this we have carried out experiments selecting from 1 to 10 pairs of channels. The optimal number of pairs for each method was selected in accordance with the results of these experiments. This work was presented in (Martínez-González B. , Pardo, Echeverry-Correa, Vallejo-Pinto, & Barra-Chicote, 2012) and (Martínez-González B. , Pardo, Echeverry-Correa, & San-Segundo, 2016). The database used for this TDOA selection related study is the RT database.

5.1 TDOA selection methods developed

In this section we will explain all the algorithms developed and the analysis carried out to establish the optimal number of pairs combinations to use. Remember that the selection of pairs is only the first step. After that, the process continues as it did in the baseline system, extracting delay values between each of the selected pairs of channels for every frame of 10ms (this will be the TDOA vector). Every DER or SER result presented in this chapter assume that the selection was performed according to one of the proposed methods and then the delays were extracted from the selected pairs to be used for the last stage of diarization (the Segmentation and Agglomerative clustering of speech regions stage in Figure 5). The xcorr between channels, and therefore the delays, will be calculated using the GCC-PHAT explained in section 2.4.1 which is widely used in diarization systems and whose weighting functions normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the xcorr.

The experiments required to determine the optimal number of pairs include the use of MFCC features because we have checked that TDOAs alone could improve the

results when compared to the baseline, but in combination with MFCC this improvement may not be noticeable. The final goal is not to reduce the error using only TDOAs, but to achieve the lowest possible error with our best combination of features. TDOAs help the diarization but MFCCs are much more informative and including them in the experiments we can assure that the modifications made to TDOAs are useful for the real system. Likelihoods of the models trained with TDOA and MFCC features are combined using weights that sum up 1. The baseline system achieves a DER for the development dataset of 12.93 (SER of 5.49), so this will be the value to beat. As this experiments will be carried out with the development dataset, the final optimal number of pairs will be the best in average for all the meetings of the dataset using a specific technique. The optimal number of pairs per meeting can be slightly different and the study of techniques to select automatically this number should be addressed in future research.

5.1.1 Cross Correlation Pairs (XCorr-Pairs)

This method takes advantage of the technique used by the baseline system to select one of the channels of the recording as the reference one. As explained before, the baseline system computes the average cross-correlation for all possible channel combinations for a block of 1s in duration. This process is repeated for $M=200$ blocks linearly spaced throughout the recording. It then singles out the channel with the highest cross-correlation (Anguera X. , 2006). This channel has proven to be the most reliable channel for computing the delay with the other microphones in order to calculate the composite signal.

Using this proposed method, we can select not only one channel but several pairs of them. Instead of calculating the average cross correlation between one channel and all the rest of them, we calculate the cross correlation between each possible combination of channels, and compare them individually. What we want to test is if using only pairs of channels with high cross correlation the system would avoid some errors in the estimation of delays leading to a better performance of the system. For each pair of channels the cross-correlation is computed as in equation (5-3):

$$\overline{xcorr}_{i,j} = \frac{1}{M} \sum_{m=1, i \neq j}^M xcorr_{i,j}(m) \quad (5-3)$$

Where M is the number of blocks of 1 sec used ($M=200$ blocks), and $xcorr_{i,j}(m)$ indicates the cross-correlation measure between channels i and j for each block m .

We have conducted experiments in order to determine how many pairs of channels would be sufficient to carry out the diarization with the lowest possible error.

In Figure 7 we observe that the best strategy when using the XCorr-Pairs selection technique should be apply a weight to the TDOA vector between 0.1 and 0.25, and selects less than 5 pairs, as the DER obtained for the RT-DEVELSET is minimum in this area.

Figure 7: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorr-Pairs selection method

Figure 8 shows the DER obtained for the development set using this method for a number of pairs ranging from 2 to 5, for different TDOA weights. We realized that, although selecting 2 pairs has a slightly better DER than selecting 4 (11.41 vs 11.55), it was also less reliable, since small changes in the TDOA weight raise the DER. Instead of choosing the point with the minimum DER, we decided to select the working point whose neighbours are less variant with the weight while keeping the DER low. Under this circumstance, the point with lowest numerical DER (pairs=2 and TDOA_weight=0.25) is surpassed by our selected working point (pairs=4 and TDOA_weight=0.15), which has a slightly higher DER (although not significant) but whose performance remains less variant in the vicinity of the working point (MFCC_weight=0.85 and TDOA_weight=0.15). The number of pairs finally selected, and therefore the dimension of the TDOA vector, has been then set to 4.

Figure 8: DER results for the XCorr-Pairs method when the number of pairs selected is 2, 3, 4 and 5, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.2 Cross Correlation Channels (XCorr-Channels)

Regarding XCorr there is another selection strategy that we have experimented with. Coming back to the idea of the baseline system, we have selected a predefined number of channels (not pairs of channels as in the previous method) with the highest

average xcorr and then computed the delays between all the possible combinations of these channels. The average xcorr for each pair is the same measure used to select the references channel, though in this case the number of channels selected will be higher than one.

$$\overline{xcorr}_i = \frac{1}{M(K-1)} \sum_{m=1}^M \sum_{j=1, j \neq i}^K xcorr_{i,j}(m) \quad (5-4)$$

Where M is the number of blocks of 1 sec used (M=200 blocks), K is the number of channels, and $xcorr_{i,j}(m)$ indicates the cross-correlation measure between channels i and j for each block m.

The idea is to discard the channels that, in average, have less reliability, and take advantage of any pair involving two channels with high cross correlation between them and all the rest. The number of combinations of pairs of selected channels will define the dimension of the TDOA vector. In contrast to the other XCorr selection strategy, where the number of pairs selected defined the dimension of the TDOA vector, in this approach if we select 3 channels the total pairs combination will be also 3, and the system will work with TDOA vector of three dimensions. However, if we select 4 channels, the possible number of pair combinations grows to 6. TDOA vector dimension will be then 6, 3 more pairs used and therefore 3 more pairs from which calculate the TDOA values only for the selection of one extra channel. In Table 6 is included the number of pairs combinations per number of channels selected.

#Channels	#Pairs
2	1
3	3
4	6
5	10

Table 6: number of possible pair of channels combinations per number of channels selected.

To illustrate the difference between this method and the XCorr-Pairs, we have included in the Table 7 a hypothetical example of the selected pairs in both situations for the same meeting being 4 the number of pairs or channels selected.

	XCorr-Pairs	XCorr-Channels
#channels available	6	6
#pairs combinations available	$C_{6,2} = \binom{6}{2} = 15$	$C_{6,2} = \binom{6}{2} = 15$
channels selected	No channel selection	Highest average XCorr: 4 Channels selected (For example channels: 1, 2, 4, 6)
Pairs selected	Highest XCorr: 4 Pairs selected (For example: 1-2, 1-3, 2-4, 2-6)	Pairs combination of the selected channels: $C_{4,2} = \binom{4}{2} = 6$ pairs selected (In the example: 1-2, 1-4, 1-6, 2-4, 2-6, 4-6)
Example of a portion of a TDOA vector (delays from same pairs in columns, each row correspond to one frame)	$\begin{array}{cccc} \vdots & & & \\ -5 & -12 & -5 & 7 \\ -5 & -12 & -6 & 7 \\ -5 & -12 & -6 & 7 \\ -4 & -12 & -6 & 7 \\ -4 & -12 & -6 & 7 \\ -4 & -12 & -6 & 7 \\ \vdots & & & \end{array}$	$\begin{array}{cccccc} \vdots & & & & & \\ -5 & -2 & 2 & -5 & 7 & 12 \\ -5 & -2 & 2 & -6 & 7 & 12 \\ -5 & -2 & 2 & -6 & 7 & 12 \\ -4 & -10 & 2 & -6 & 7 & 12 \\ -4 & -10 & 2 & -6 & 7 & 12 \\ -4 & -10 & 2 & -6 & 7 & 12 \\ \vdots & & & & & \end{array}$

Table 7: Comparison table between the two XCorr related methods for a hypothetical meeting when selecting 4 pairs or 4 channels.

The number of pairs if we chose 10 channels would grow up to 45, so, for this method, we have carried out experiments selecting only from 2 to 5 channels, which would mean a maximum number of pairs of 10, equal to the experiments carried out for the other methods. The results of the experiments carried out to determine the best combination of dimension of the TDOA vector and weight applied to it is shown in Figure

9. Note that the axis shows not the final number of channels selected but the number of pairs of channels associated to those channels (the number of pairs used when a specific number of channels is selected can be consulted in Table 6). The lowest error is obtained in this case when the weight applied to the TDOA stream is between 0.1 and 0.2 and a number of pairs equal to 3 or 6, which would mean respectively 3 or 4 channels selected by this method.

Figure 9: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorr-Channels selection method

If we look in detail in the working area (see Figure 10) we see that although selecting 4 channels (i.e. 6 pairs) have very good results, they are not as good as those with only 3 pairs and also there is more variation in the error rate. Results variation when working with 3 pairs are lower and more stable, thus making it a better option as a working point. As before, these experiments were carried out using MFCC and TDOA. The weight applied to TDOA vector could either be 0.15 or 0.2, as the difference in DER between this two is very small and either increase or decrease this weight make the error rate worse. However, the numerical result for TDOA_weight=0.2 is slightly better than the result for TDOA_weight=0.15 (11.5 vs 11.65), so the final selected working point has been 0.2 of TDOA_weight and 3 channels (and therefore 3 pairs) selected.

Figure 10: DER results for the XCorr-Channels method when the number of pairs selected is 3 and 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is 1-TDOA_weight

5.1.3 Dynamic Margin (DM)

This method focuses on finding the best pairs according to the actual TDOA vector that we could extract from them and how useful it could be to the diarization process. As mentioned before, if the speakers do not move, the delay value associated to each one will be the same throughout the whole recording. This value will depend on the position of the speakers and how close they are to the microphones of the pair evaluated. Ideally each speaker will have a different delay value associated and it will be calculated with no errors. We will hardly ever find this ideal scenario. Some pairs of microphones are located in such positions that two or more speakers can share their delay value. A selection method which enlarge the probability of choosing pairs whose delays associated are different for every speaker could solve this problem.

The so called Dynamic Margin (DM from now on) method focus on selecting the pairs of channels with the highest separation between the two more extreme delay values of a particular pair. The aim of this method is to be able to select the more discriminative pairs of channels as we hope that a higher margin allows the presence of

much more easily separated clusters of delays. If any pair have very similar delays for most of the frames calculated the separation of frames from one speaker to another would turn impossible, even if the delay values are correctly estimated. In Figure 11 and Figure 12 we show two examples of histograms of delay values from different pairs of channels of the same meeting. The Figure 11 is what we would desire, a clear separation between the peaks of the histogram which would mean a clear location of the voice sources. In Figure 12 the two location of the voices are much nearer and even one of them has very reduced number of examples. Assuming that there are more than one speaker, one location seems unlikely and, in any case, not useful at all, as the ultimate objective is to be able of separate between these speakers. We hope that choosing pairs with a higher dynamic margin of the TDOA values would help the system to discriminate between speakers.

The implemented method creates a histogram of delays for each possible pair of channels. The calculation of all the possible delays for each frame of 10ms (which is the frame length which we will use to evaluate the system) would take a very long time (increasing severely the computational time) which looks like a waste of resources if we only want to decide which ones have the higher dynamic margin. Therefore, the delays are calculated only every 250ms throughout the whole recording. The histograms of the TDOA values are generated using a bin width of 5ms. Once the histograms are generated, we select the subset of pairs with the highest dynamic range (highest difference between the maximum and the minimum delay, see Figure 11 and Figure 12). Due to miscalculations the histogram of delays may present some very high or very low values, hopefully not in a very large proportion. To prevent the selection of these pairs of channels with only some outliers in the furthest regions the system will ignore any bin of the histogram with less than 25 samples. This method is similar to that presented in (Sun, Nwe, Ma, & Li, 2009) although, in that paper, no details on the performance of the method alone were presented and the TDOAs were used only for the initial segmentation.

Figure 11: Example of the histogram of delays of a pair of channels and its dynamic range. Delays calculated with this pair show clear separation between peaks which should make it more discriminative to locate speakers. The red horizontal line represents the maximum span of the dynamic margin

Figure 12: Example of the histogram of delays of a second pair of channels and its dynamic range. Bins with very few samples are not taken into account when calculating the dynamic margin. Delays calculated with this pair are mostly concentrated around the same value making it less discriminative to changes in speakers. The red horizontal line represents the span of the dynamic margin

As before, the optimum number of pairs needs to be established. The results obtained with the RT-DEVELSET when applying this selection method are shown in Figure 13. As we have done with the other methods we zoom the area of lowest DER in Figure 14. This graph represents the DER values for a final number of pairs selected equal to 2 and 3, which were the only ones with error values lower than the baseline system. In this case, the lowest error is obtained with the final dimension of 3 and a TDOA weight

equal to 0.15. As the next possible working point has significantly higher error rate we have decided to continue with this one. Therefore, the selection method will choose the 3 pairs with the highest dynamic range and then calculate the TDOA values for each frame of 10ms only for these selected pairs.

Figure 13: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the DM selection method

Figure 14: DER results for the DM method when the number of pairs selected is 2 and 3, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.4 Extreme Regions (ER)

In contrast to the previous algorithm, in the Extreme Regions (ER from now on) method, the objective is not to have the largest possible range of possibilities for the delays values, but to have as many examples as possible with high delays. If a pair of microphones are located very near one to the other, or in positions equally separated from the speakers, the delay of the voice signal between the moment it reaches the first microphone and the moment it reaches the second one would be very low. The resulting set of TDOA values would exhibit a very high concentration of values near to 0, which is a situation we have observed for many pairs. If a speaker is represented with a very high value of delay, the probability of sharing this value with many other speakers would be reduced. With this configuration we want to study what would happen if we force the system to work with pairs of microphones that return mainly high delays.

As happened in the DM method, the Extreme Regions method (ER) will begin by calculating delays between all the different channels every 250ms. As it was done in the DM method, one histogram of delays is calculated for each pair of microphones, using a bin size of 5ms and then discarding all the bins with less than 25 appearances. Once this estimation has been carried out, the algorithm will make a histogram joining all the histograms previously calculated for each pair. This new histogram would then be composed of all the delay values calculated from all the pairs of channels (thus at the end there will be only one histogram instead of one per pair of microphones, as in the DM method).

Some positions of the histogram are not taken into account for the decision of the pairs to use. It has been decided to discard the most extreme positions that form 0.5% of the total number of delays calculated. This procedure aims to avoid some very high delays that are considered outliers. Immediately afterwards two margins are set up, one in the positive part of the histogram and one in the negative part. These two margins define two regions which will contain 40% of the total number of delays calculated for the whole meeting and all the pairs. In Figure 15 a hypothetical example

of a histogram with the regions discarded (region A) and the regions with 40% of values (region B) is represented for better understanding.

Figure 15: Hypothetical example of the regions that would be defined in the histogram of delays in the ER method. Region A is discarded (containing 0.5% of the total number of delays). Region B is target region (containing the 40% of the remaining number of delay values).

Once the target region is defined, the system will check which of the pairs have more delay values in that area. As happened in the dynamic range method, we chose only a subset of the total possible combinations. After carrying out experiments selecting a number of pairs from 1 to 10, the only good results founded use only 2 pairs of channels (see Figure 16).

The TDOA vector therefore, will have a dimension of 2. About the weight that should be applied to the TDOA stream shows that the minimum error is obtained with 0.15. Once the 2 pairs are selected, the implementation continues as usual, estimating the delay values for each selected pair and for each frame of 10ms, and then using them for the segmentation stage.

Figure 16: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the ER selection method

Figure 17: DER results for the ER method when the number of pairs selected is 2, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.5 Most Common (MC)

The next method presented in these pages will focus on the opposite delay values than the ER method. The most common delay values (those that are repeated several times) are usually the lowest ones and can be shared by many speakers, still, if these values are so common it could mean that they do correspond to an actual speaker who participates often in the dialogue. Identifying the most repeated delay values and the pairs of channels with which they have been estimated could be useful for the selection of the better pairs. As we will see, the discrimination of speakers will be harder when using this method, but the analysis can still serve as a matter of comparison with the ER. This method will initially calculate the same histogram of delays calculated for the ER method. First delays are estimated for every pair of channels every 250ms. Then a histogram of delays is calculated using a bin width of 5ms. At this moment delay values with less than 25 appearances are discarded. Finally, all the histograms from all the delays are joined in a whole histogram containing all the delay values from all the pairs. The selection of pairs is carried out regarding the most common TDOA values of the histogram.

The highest bins of the histogram correspond to the most common delay values. We have decided to select as many bins as the number of pairs of channels we are looking for. To calculate the TDOA vector we will use the microphone pairs with the highest contribution to those bins, which means that those pairs of microphones that have more delay values among those most common delays will be used in the final stage of clustering and segmentation. In Figure 18 a hypothetical example of a histogram and the bins that would be chosen if we were aiming at 3 pairs is represented.

Figure 18: Hypothetical example of the bins of the histogram of delays that would be chosen in the MC method.

Again, to determine the best number of microphone-pair combinations several experiments have been carried out. In Figure 19 the results are represented. As we can see, this method requires more pairs to reach the lowest possible error. No good results are obtained until we reach 6 or 8 dimensions, probably due to the difficult separation of speakers in the TDOA vector estimated from the selected pairs. In Figure 20 both situations are represented, though results are similar the best performance is finally obtained for a number of pairs combinations equal to 8.

Figure 19: DER obtained for the RT-DEVELSET across the weight applied to the TDOA vector and the final dimension of the TDOA vector when using the MC selection method

Figure 20: DER results for the MC method when the number of pairs selected is 6 and 8, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.6 Principal Component Analysis

With this method, we try to take advantage of any information which can be held in any delay computed between any pair of channels. Until now we have tried either to select the most discriminative pairs or the most reliable. With this method we assume that any pair can hold information about the location of the speakers and we will try to use all of it.

As we intend to get as much information as possible from the delays we would need to calculate a TDOA feature vector including every possible pair of channels. Then, we carry out a PCA to reduce dimensionality. However, we have already mentioned that some meetings have up to 24 microphones, which means 276 combinations and a considerably increase in the computational time. To reduce this effect we have decided to use the cross correlation measure that we presented in section 5.1.1. This measure will allow us to discard those pairs that do not align so well and that would be less reliable. As the objective is to limit the computational time required to end the task of

extracting the delays but keeping as much information as possible, we have decided to choose the 50 pairs of channels that have largest xcorr. With this threshold we assure that the computational time does not raise too much, while keeping most of the information for the majority of the meetings (only meetings with more than 10 microphones would generate more than 50 pairs, and therefore, get some of them discarded). Once the TDOAs for each pair have been extracted for every frame of 10 ms, a PCA is applied reducing the dimensionality of the TDOA vector. Note that this time the final dimension of the TDOA vector will not represent the number of pairs selected, but the final dimension specified to the PCA algorithm. This optimal dimension is what we have to study. We carried out experiments from 1 to 10 final dimensions and we checked that this method, like the previous one, reaches its optimal working point when the dimension is higher than 6. The weight to be applied to the TDOA vector should be between 0.15 and 0.2.

To select a final working point we should focus on the area of minimum DER. Figure 22 represents the numerical values for the combinations that perform well enough to compete with the baseline. Selecting a final dimension for the PCA equal to 8 and a weight of the TDOA vector equal to 0.15 is clearly among the best choices.

Figure 21: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the PCA selection method

Figure 22: DER results for the PCA method when the final dimension is 6, 7, 8, 9 and 10, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.7 XCorr threshold

To estimate the delay between microphones we use the cross correlation between the two signals coming from the two microphones. As we explained before, the cross correlation will be maximum when the two signals are aligned the best they can. During the experiments, we observed that some pairs of microphones obtain an average cross correlation measure much lower than other pairs for the same meeting. In Figure 23 the average cross correlation for one meeting which has two arrays of 8 microphones is shown. Different colours are used to represent that any particular combination has been made with two microphones from the first array (array1-array1), two microphones from the second array (array2-array2), or two microphones from different arrays (array1-array2). Microphones 1 to 8 correspond to the first array and microphones 9 to 16 to the second. Any combination of microphones $i-j$ is the same as $j-i$, thus we do not represent it. In this meeting the total number of possible combinations is 120.

Figure 23. XCorr of one meeting with two arrays. The order of the combinations of microphones are represented according to its cross correlation measure in descendent order.

As it can be seen, there is a big difference between the cross correlation achieved between two microphones of the same array, and two microphones coming from different arrays. It is important to know that this situation occurs for the whole development database every time we deal with meetings recorded with more than one array of microphones. This is the case for the “EDI” meetings or the “AMI” meetings in our database. We conclude that it may represent some particular characteristic of the meeting room. Maybe the two arrays are situated in different physical rooms and communicated by loud speakers, or maybe one of the arrays has a much lower quality. In any case we have decided that, if we cannot rely in the delay measure, it is preferable not to use it at all.

Based on this finding, we proposed a preselection of pairs discarding all the pairs of microphones with very low average cross correlation when compared to the maximum average cross correlation of any other pair at that meeting. For our purpose we have defined the following threshold:

$$Threshold = K \cdot \max(xcorr_{i,j}), K < 1 \quad (5-5)$$

Being $xcorr_{i,j}$ the cross correlation between channel i and j , and K the factor applied to the maximum $xcorr$. Once the pair of channels with the maximum cross correlation is selected, the average $xcorr$ of the other pairs is compared with this threshold. Only pairs with a cross correlation above the threshold are kept to be used in following stages.

The optimal value of K is selected empirically. To set the value of K we will carry out experiments with the baseline system, which selects a reference channel and then computes the TDOA's between this reference channel and all the rest. These experiments are included in section 5.1.7.1. As we will see the best K is set as $2/3$. This threshold was conceived as a minimum quality measure for the pair to be considered in the following selection of channels. Some recent works have focused on the key part of the segmentation and clustering is the correct estimation of the speaker model through data as pure as possible (Sinclair & King, 2013), which is consistent with the idea of using only the reliable delays.

As a minimum quality measure, it can be applied to any of the aforementioned methods, including the baseline method. We propose three new combined methods: XCorrThreshold+baseline, XCorrThreshold+DM, XCorrThreshold+PCA. The XCorrThreshold+XCorr methods, which would mean the application of the threshold to discard the worse pairs, followed by the selection of pairs using the XCorr method, will be also discussed. The ER and MC method were not included in this section because their performance was not as promising as the results for the other mentioned methods. The DM is kept because the idea of selecting pairs with high dynamic range looks promising, as the works in (Sun, Nwe, Ma, & Li, 2009) perform a selection with a similar idea and good results. We also think that some of the errors in performance could be avoided by the application of this XCorrThreshold. Finally, we have developed one final selection technique that we have called XCorrThreshold+Kmeans. This method begins with the extraction of all the delay values for every pair of channels not discarded by the XCorrThreshold (as it is done for the PCA method). After that, it computes a Kmeans of

delay values and selects pairs according to the Dinamic Margin (as in DM method) and the number of clusters found or according to a measure called Silhouette coefficient.

5.1.7.1 XCorrThreshold+Baseline

In the baseline system we select a microphone and compute delays between it and all the others. We intend to apply the aforementioned rule on the minimum cross correlation, and then, continue with the regular running. We extract the same delays as the baseline system and after that the system will remove delays corresponding to any pair of channels whose average cross correlation is below the threshold. The remaining pairs, and their computed delays, are the only ones used for following stages of the diarization process. Note that the baseline system would choose the most reliable channel and then compute delays between it and all the remaining channels, regardless of the quality of these other channels. In this situation if one channel has been previously discarded, the pair formed with the reference channel and this other microphone would not be eligible.

To evaluate the modification with this baseline system we vary K in steps of $1/6$ and discard every pair whose average xcorr is below the threshold in equation (5-5). All the pairs with higher xcorr can then be selected following the baseline procedure. Results are shown in Figure 24.

Figure 24: DER for the baseline system using MFCCs and Delays, weight to MFCC vector is 0.9 and 0.1 for the delays. K is the factor applied to the maximum xcorr to determine the minimum threshold. In columns is the DER for the RT-DEVELSET.

There are three possible K values whose DER is below the baseline ($K=0$). With K equal to $3/6$ and $4/6$, the error value is very similar. Though DER when $K=3/6$ is slightly better, the computational time is higher, as the number of pairs discarded is lower. Considering this and that the difference in DER is very small, we have preferred to work with a $K=4/6$, to keep low DER and, at the same time, we will be able to discard more non-reliable pairs, working only with the best ones, which also makes the system faster.

With this threshold we are sure that all non-reliable pairs are discarded, keeping nevertheless enough of them as usable pairs. The value of $K=2/3$, extracted from experiments with the baseline and the *XCorrThreshold*, will be used for every method from now on.

5.1.7.2 XCorrThreshold+DM

In the analysis of the results for the DEVELSET with the previously proposed DM method, we have noticed that most of the pairs with a low *XCorr* measurement have a high variation in the delays calculated. This high variation will result in pairs with a high dynamic margin of the delays. Therefore, when using the DM method, it is probable to select pairs with low *XCorr*, which, according to what we have explained before, could be less reliable.

To cope with this problem we have included the rule of the minimum cross correlation before the selection. The system will compute all the histograms as before but will not take into account those with average cross correlation lower than the threshold defined.

The number of selected channels must be defined again, as the previously optimal number was extracted from the experiments that do not use the *XCorrThreshold*. The experiments (Figure 25) with the RT-DEVELSET show that when we discard the non-reliable channels the optimal number of pairs of channels to be used grows to 5 or 6. The slightly better result for the dimension 5 (see Figure 26) make us choosing it as the optimal working point.

Figure 25: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorrThreshold+DM selection method

Figure 26: DER results for the XCorrThreshold+DM method when the number of pairs selected is 5 and 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.7.3 XCorrThreshold+PCA

The PCA method already takes advantage of the XCorr quality measurement. The first experiments reduced the number of possible combinations to 50 at most, selecting them with the XCorr method. As a matter of fact, this procedure discarded most of the pairs with very low cross correlation, which, in the end, is what we intend to do. However, this new threshold depends only on the actual meeting, as it is a percentage of the maximum cross correlation at that precise meeting and will discard more pairs initially, thus reducing the computational time.

The number of discarded channels will be higher and the PCA method will then work with a delay vector of lower dimension, therefore, it is necessary to recalculate the optimal final dimension of the PCA. Figure 27 shows the values in which the error rate is minimum. There are some good results with dimension equal to 9 but most of the points with lowest errors are found between dimensions 5 and 7. Observing Figure 28, it becomes clear that the final dimension of 9 does not work as good as the other ones, in particular the best results are obtained for a dimension equal to 6, and, in this case, even increasing the TDOA weight to 0.2.

Figure 27: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the XCorrThreshold+PCA selection method

Figure 28: DER results for the XCorrThreshold+PCA method when the final dimension is 5, 6, 7 and 9, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.7.4 XCorrThreshold+XCorr Methods

Though in principle this combination could be possible, first to assure the minimum quality of the pairs available and then the XCorr methods (either XCorr-Pairs or XCorr-Channels) to choose from among them (which could result in lower number of pairs available than those requested for the XCorr regular method), in practice the cross correlation methods are restrictive enough to discard, by themselves, any pair of microphones with such a low cross correlation that would be discarded for the defined XCorrThreshold. The XCorr-Pairs and the XCorr-Channels method only keep 4 and 3 pairs respectively, those with the highest cross correlation. We have checked that none of these pairs selected for any of the meetings have a cross correlation lower than the defined threshold (this, of course, would mean that the rest of the pairs at the meeting, those that were not selected in the first place, had an even lower cross correlation). In the current situation the system will not modify anything, and the resulting delays and DER will be the same. Of course, we can assure this situation only with our current sets of meetings. We could encounter one meeting with less than 3 or 4 usable pairs

according to the XCorrThreshold strategy, and the system should be able to cope with this situation if presented.

5.1.7.5 XCorrThreshold+Kmeans methods

This new approach was developed inspired by the work in (Sun, Nwe, Ma, & Li, 2009), where the authors perform a clustering of the delays to select the pairs to use. As mentioned before the ultimate task that we want to accomplish with the selection of the TDOA pairs is to be able to distinguish between different speakers by their localization in the room (sort of gps of speakers). As mentioned in other points, the errors in the estimation of the delays can make the system to choose one pair above another, when, in terms of separate location, it may be better this last one. Diarization in absence of overlapped speech, could be carried out perfectly with just TDOAs extracted from one of the pairs, provided perfect conditions for the extraction of delays. These conditions would include that the speakers do not move, that the pairs chosen were placed in such a way that every speaker of the room had only one specific delay associated and that these delays were perfectly estimated by the system. Of course the perfect conditions are never met. In actual situations speakers can move and speak at the same time, errors in the estimation of delays can be made, and, even if nothing of the previous happens, the delays estimated for one speaker can vary slightly and also, due to the relative position of the microphones and the speakers, two particular speakers can share the same range of delay values.

This technique pretends to solve some problems that we have observed using the DM technique. DM selection method selected the pairs with the highest range between the highest and the lowest delay value. With the DM method we intend to select those pairs with high probability of having differentiated clusters of delays but, even if the outliers are correctly removed, we could still encounter the problem of selecting merely the pairs with two groups of delays very separated, as if there would be just two speakers in the room. We want to check if better results could be achieved by selecting pairs whose distribution of delays show more than two clearly separated clusters.

With this idea in mind we decided to perform a Kmeans of the delays extracted and select the pairs according to the number of clusters found and the separation among them. The TDOA values are extracted initially with the XCorrThreshold method and then clustered applying a Kmeans algorithm. To decide how many clusters to look for, the system performs several Kmeans of delay results of each pair of channels varying the final number of desired clusters. We then calculate the silhouette criterion to decide which segmentation is the most accurate.

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the i^{th} point is defined as:

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (5-6)$$

Where a_i is the average distance from the i^{th} point to the other points in the same cluster as i , and b_i is the average distance from the i^{th} point to points in the nearest cluster.

The silhouette value ranges from -1 to 1. A high silhouette value indicates that i is well-matched to its own cluster, and poorly-matched to neighbouring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters.

In this work we have used the average silhouette value of a subset of the dataset to determine whether the segmentation carried out with the Kmeans of delays had an appropriate number of clusters. The algorithm of selection computes Kmeans of delay values for 4 clusters to 7 clusters, and then compare the average silhouette value of the clustering, keeping only the best segmentation.

Once Kmeans of delays has been performed for every pair of channels we have followed two strategies to carry out the final selection, one that takes into account the

dynamic margin of the distribution (Kmeans+DM) and one that takes advantage of the silhouette criterion to decide the best pairs.

5.1.7.5.1 *XCorrThreshold+Kmeans+DM*

It focuses on the dynamic margin of the delays as well as the number of actual differentiated clusters in the segmentation. To do so, initially the system compares the centroids of the clusters, if they are very near one to another then fuse these two clusters in only one and consider it as a whole unique cluster. For the porpoise of these experiments we have set that two clusters are too near to be considered separately if their centroids differentiate in only 1ms. Then compute the dynamic margin of the clusters, remember that we are working with delays from only one pair at a time which means that data has only one dimension, and dynamic margin of the segmentation would be just the difference between the most positive centroid and the most negative.

The measure used to rank the pairs from the first to be chosen to the last one is:

$$DM * \#clusters \quad (5-7)$$

Where DM denotes the dynamic margin of the clusters and #clusters are the number of actual clusters once the system has carried out the fusion of those very near one to each other (sometimes two clusters are created whose centroids are next to each other, and one of them has a very low number of samples in comparison to the other).

As usual, we have carried out experiments to determine the best number of pairs to be chosen, in this case, there is a wide range of possibilities, from 2 to 6 combinations and a TDOA weight between 0.15 and 0.2. In Figure 30 we can see that the best results are obtained with 5 pairs of channels

Figure 29: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the Kmeans+DM selection method

Figure 30: DER results for the Kmeans+DM method when the number of pairs selected range from 2 to 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$

5.1.7.5.2 *XCorrThreshold+Kmeans+Silhouette*

The second strategy is just to compare the average silhouette coefficient of the kmeans clustering of the delays belonging to each pair and then, select those with the highest value. The silhouette coefficient is computed as before for the Kmeans clustering of delays (equation (5-6)) for a group of samples of the clusters and then compute the average. The idea behind the sampling is just to speed up the process of calculating the Silhouette coefficient.

Once the average Silhouette coefficient has been calculated for every pair then we rank them and keep those with the highest average values. The delays associated to them will be then used in the posteriors stages of diarization.

The final optimal dimension has been set empirically, as it was for the other methods. In Figure 31 the results for the different combinations of weights applied to TDOA vector and dimension of it are represented. The best results are located around the TDOA weight equal to 0.15 or 0.2, and dimension between 4 and 6. In Figure 32 the exact numerical values are presented.

Figure 31: DER obtained for the RT-DEVELSET across the TDOA weight applied to the TDOA vector and the final dimension of the TDOA vector when using the Kmeans+Silhouette selection method

A final dimension of 4 and 5 pairs presents minimum error values at 0.15 and 0.2, any of them could represent a good working point. The TDOA weight equal to 0.2 has slightly better result than 0.15 for the dimension of 4, but the neighbour weight=0.25 has more error than the 0.1 so we decided to keep working with the point of weight=0.15, which is also consistent with most of our previous algorithms.

Figure 32: DER results for the Kmeans+Silhouette method when the number of pairs selected range from 4 to 6, across the weight applied to the TDOA stream. The weight applied to the MFCC stream is $1 - \text{TDOA_weight}$.

5.2 Comparison between TDOA selection methods and evaluation

As stated before, we have carried out experiments to decide the optimal dimension of the desired TDOA vector for each technique. We also need to establish the weight to be applied to the TDOA vector, which is also decided empirically. The results obtained for the RT-DEVELSET when using the six initial methods that do not use the XCorrThreshold technique, are shown in Figure 33, where values are given across different weights for the two streams of data: MFCC and TDOA. Also a zoom of the area of minimum DER for all the systems is included in Figure 34.

Figure 33. DER of new methods for the development set with their optimal dimension of the TDOA vector, using MFCC and TDOA features. Results are shown across the weight applied to the TDOA stream.

Figure 34: Detail in the area of minimum error of the DER of new methods for the development set with their optimal dimension of the TDOA vector, using MFCC and TDOA features. Results are shown across the weight applied to the TDOA stream.

In this figures only the results of each method with its optimal dimension of the TDOA vector, i.e., the number of pairs of channels selected, are detailed. Note that these dimensions are included in the figures and the procedure followed to their determination was explained in the previous pages. The baseline method is also included for comparison. Its dimension is not specified because it varies with the meeting, as it will use a number of pairs equal to the number of channels minus 1. For the development set this dimension cannot grow beyond 15, as the meeting with the maximum number of channels has 16 of them.

By analysing the results of the XCorr-Pairs, we can see that the error for this method remains below the baseline for any TDOA weight (see Figure 33). The lowest ones, shown in Table 8, are obtained with a combination of weights for the MFCC and TDOA features of 0.85 and 0.15. In comparison to the baseline system, with 0.9 for the weight of the MFCC stream, and 0.1 for the TDOA stream, the weight applied to TDOA

values increases, which means that the selected TDOAs are more useful, even though the number of pairs used has been reduced to only 4 pairs.

Nearly the same occurs for the XCorr-Channels algorithm, every point analysed has lower error than the baseline, with the exception of the experiment in which MFCCs took no part (TDOA weight equal to 1) and when the weight applied to the TDOA stream is only 0.05. However, the selected working point is far from these two, in the TDOA_weight=0.2. As before, the contribution of the delay values to the diarization has grown, as the error is reduced when the weight applied to the TDOA stream increases, proving that the use of this measure (the cross correlation) to select the most suitable pairs is very useful in any case. In Table 8 the exact numerical value of the selected working point is included alongside the 95% confidence interval. To calculate it, we have used the equation in (5-8).

$$95\% \text{ confidence bands} = 1.96 \cdot \sqrt{\frac{\text{errorRate} \cdot (100 - \text{errorRate})}{\#frames \text{ evaluated}}} \quad (5-8)$$

Where the errorRate refers to the DER or the SER, and the #frames evaluated are the number of frames actually used for evaluation.

Looking at the results for the DM algorithm, it is clear, for most of the weights, that the DER obtained with the DM method is better than the baseline method. It is noticeable that one of the few points with worse performance is when the TDOA features are used on their own (the weight of the MFCC stream equal to 0 in Figure 33). However, we do not intend to improve the results at this point but in the area of the lowest DER (around the baseline working point, MFCC weight=0.9). The best result for the DM method is obtained, as with the XCorr-Pairs method, with the weight 0.85 for the MFCC stream and 0.15 for the TDOA stream. The exact values are shown in Table 8.

ER and MC methods are the only two that have many error values above their baseline counterparts. However, when we reach the surroundings of the working area the baseline system cannot keep this tendency and ER and MC techniques begin to surpass its performance. Finally in the selected working points for both of them the

results are lower than the best one obtained with the baseline technique. Though both techniques improve the result only ER makes it with a higher weight applied to the TDOA, which could be an indicator that the delays used in the MC method are not as good as those selected for other techniques.

Similar to what happened with the XCorr-Pairs or the XCorr-Channels algorithms, the performance of the PCA method with the RT-DEVELSET is generally better than the baseline. The best performance of this method is achieved for a combination of weights for MFCC and TDOA features of 0.85 and 0.15 respectively, which is not only the minimum error for this selection method, but the lowest DER of all the techniques, for the RT-DEVELSET. Admittedly, the difference between the XCorr based techniques and the PCA method is very small so any decisions about the best system to use must wait until we test it with the evaluation set of meetings. In Table 8 the final results can be consulted.

It is worth emphasizing that when using the TDOA features on their own, the results for two of the best methods, XCorr-Pairs and PCA, are better than the baseline (the TDOA weight being equal to 1 in Figure 33), indicating that the TDOA information is more robust since it obtains a lower DER. This fact is also demonstrated at the optimum point in which the MFCC weight is 0.85 instead of 0.9, as in the baseline. A bigger weight for the TDOA feature means experimentally that one can rely more on this feature.

In Figure 35 we present the best DER results obtained for each of these techniques for the RT-DEVELSET across the number of pairs chosen. The TDOA weight applied at each point is the minimum we could find, so, for instance, the DER for the XCorr-Pairs method and number of pairs 2 and 3 are very similar but were obtained by applying a different weight to the TDOA feature stream (0.25 and 0.15 respectively). Therefore, these values represent the lowest possible error that we could achieve with any of the developed systems for any of the final dimensions studied. It is clear in this graph that the best algorithms for any final dimension are the XCorr related techniques and the PCA. XCorr-Channels method is represented in bars because the range of channel pairs that can be chosen is not continuous. The system will work with 1, 3, 6 or 10 pairs if the XCorr-Channels selection method selects 2, 3, 4 or 5 channels respectively.

In column 2 of Table 8 we also include the best speaker error rate obtained for all the systems developed using the optimum parameters for the RT-DEVELSET. This error rate is very important because though the diarization process is greatly affected by the speech non speech segmentation, the fact is that none of the experiments carried out modify in any aspect the speech non speech segmentation of the baseline. Therefore the MISS and the FA error keeps constant throughout the whole process and the only part of the diarization error that can be reduced is the SER. Remember that MISS+FA error rate, constant for every system, is equal to 7.4%.

Summarizing information in the Table 8 until now we must remember that the dimension of the TDOA vector calculated is included beside the name of each method. Weights for the MFCC and TDOA features are 0.9 and 0.1 respectively for the baseline and the MC method, and 0.85 and 0.15 for the XCorr-Pairs, the DM and the ER. The XCorr-Channels is the only one until now whose TDOA weight has increased to 0.2 As it was previously said the results for every system developed decreased the error rate of the baseline system, the improvement of every system with respect to the baseline is specified in brackets in Table 8.

Figure 35: Optimal DER for the 6 methods with the development set when selecting a specified number of pairs.

Let us now begin with the methods that combine XCorrThreshold with other approaches. We have carried out additional experiments to set the best combination of weights for the two feature vectors used, and the new systems: XCorrThreshold+baseline, XCorrThreshold+DM, XCorrThreshold+PCA and XCorrThreshold+Kmeans. The ER and MC methods had not very promising results and we have decided to perform the evaluation of XCorrThreshold only with the best ones. Admittedly, DM had not as good results as expected, but the idea behind the method looked promising for us and results for different values of TDOA weight remains below their baseline counterparts for nearly every point so we decided to study how much the XCorrThreshold could improve this method. As mentioned before, the combination of XCorrThreshold with XCorr-Pairs or XCorr-Channels would not change any selection because XCorr-Pairs and XCorr-Channels are far more restrictive than the XCorrThreshold method, which is why that combination was not included in our results.

When we apply it to the baseline, the best DER is obtained for the same combination of weights, 0.1 for the TDOA vector and 0.9 for the MFCC. As we mentioned in section 5.1.7.2, the working point for the XCorrThreshold+DM is still 0.15 for the weight applied to the delay feature vector, though this time the best combination uses 5 pairs. The experiments carried out to define the optimal final dimension for the XCorrThreshold+PCA showed that a final dimension of 6 and a light increase in the weight applied to the TDOA vector (now set to 0.2) was the best combination for reducing DER. About the XCorrThreshold+Kmeans methods, the optimal working point for the XCorrThreshold+Kmeans+DM method was set to 5 dimensions and 0.15 of TDOA weight through experimentation, and, for the XCorrThreshold+Kmeans+Silhouette method, we have decided to test the point with TDOA weight equal to 0.15, as most of the previous methods, and number of chosen pairs equal to 4.

The best numerical results obtained for the RT-DEVELSET for every method combined with XCorrThreshold are included in Table 8. If we compare the results for the new methods with those corresponding before applying XCorrThreshold, all of them improve their performance achieving the minimum value of 11.08 DER for the XCorrThreshold+PCA method. XCorrThreshold+Kmeans+DM has not direct counterpart though the final selection method is highly related to XCorrThreshold+DM method so

we compare these two results and check that the Kmeans algorithm has improved greatly the results for the DM method, achieving the second best result, 11.09, which is also statistically indistinguishable of the result obtained for XCorrThreshold+PCA system. The XCorrThreshold+Kmeans+Silhouette method performs also well, though we do not get results as good as those achieved by the XCorrThreshold+Kmeans+DM, the DER is still much better than the XCorrThreshold+DM and comparable to the PCA system.

It is clear at this point that, as expected, the XCorr measure applied as a quality requirement has proved to be very beneficial for the task of selecting pairs of channels, as there is not one system that has not seen the errors reduced only by its use.

5.2.1 Evaluation with the test set

In order to prove that the methods work properly for other sets of meetings, we have evaluated the results with a new set of meetings: the RT09 set (7 meetings), using the parameters optimized for the development set (dimension of the TDOA vector and weight applied to it). DER and SER values for the RT09 set are also presented in Table 8.

It can be seen that the results obtained for the evaluation set improve the baseline system for nearly all the tested methods, though the improvement in some cases is not very large. The MC method is the only whose error rate is increased for the test set. Remember that this is the only selection method included which optimal TDOA weight for the development set was not higher than the weight used for the baseline system (TDOA weight equal to 0.1), indicating that the info supplied by the delay vector might not be much better than the baseline. The rest of the results for the test set are consistent with those obtained with the development set. Among those without the XCorrThreshold application, the XCorr-Pairs, XCorr-Channels and PCA have the lowest errors. At the same time all the combined methods (those using XCorrThreshold) are better than their non-combined counterparts with the exception of the PCA method which does not change. However, it is clear that the XCorr-Pairs, XCorr-Channels and the PCA methods clearly outperform the EM, MC or DM methods, including the system that applies the pre-processing of the XCorrThreshold.

System	DER DEVELSET (% error improvement)	SER DEVELSET (% error improvement)	DER RT09 (% error improvement)	SER RT09 (% error improvement)
Baseline	12.93 $\pm$ 0.05	5.49 $\pm$ 0.04	26.09 $\pm$ 0.11	17.4 $\pm$ 0.1
Xcorr-Pairs (Dim 4)	11.55 $\pm$ 0.05 (10.67)	4.11 $\pm$ 0.03 (25.14)	16.51 $\pm$ 0.09 (36.72)	7.82 $\pm$ 0.07 (55.06)
Xcorr-Channels (Dim 3)	11.5 $\pm$ 0.05 (11.06)	4.06 $\pm$ 0.03 (26.04)	16.41 $\pm$ 0.09 (37.10)	7.71 $\pm$ 0.07 (55.68)
DM (Dim 3)	12.24 $\pm$ 0.05 (5.34)	4.80 $\pm$ 0.03 (12.57)	25.18 $\pm$ 0.11 (3.49)	16.48 $\pm$ 0.09 (5.29)
ER (Dim 2)	12.36 $\pm$ 0.05 (4.41)	4.92 $\pm$ 0.03 (10.38)	23.28 $\pm$ 0.11 (3.49)	14.58 $\pm$ 0.09 (16.21)
MC (Dim 8)	12.65 $\pm$ 0.05 (2.17)	5.21 $\pm$ 0.04 (5.10)	26.65 $\pm$ 0.11 (-2.15)	17.95 $\pm$ 0.1 (-3.16)
PCA (Dim 8)	11.32 $\pm$ 0.05 (12.45)	3.87 $\pm$ 0.03 (29.51)	18.05 $\pm$ 0.1 (30.82)	9.35 $\pm$ 0.07 (46.26)
XCorrThreshold+Baseline	11.37 $\pm$ 0.05 (12.06)	3.92 $\pm$ 0.03 (28.59)	21.13 $\pm$ 0.1 (19.01)	12.44 $\pm$ 0.08 (28.50)
XCorrThreshold+DM (Dim 5)	11.59 $\pm$ 0.05 (10.36)	4.15 $\pm$ 0.03 (24.40)	24.13 $\pm$ 0.11 (7.51)	15.44 $\pm$ 0.09 (11.26)
XCorrThreshold+PCA (Dim 6)	11.08 $\pm$ 0.05 (14.31)	3.64 $\pm$ 0.03 (33.70)	18.08 $\pm$ 0.1 (30.70)	9.38 $\pm$ 0.07 (46.09)
XCorrThreshold+Kmeans+ DM (Dim 5)	11.09 $\pm$ 0.05 (14.31)	3.65 $\pm$ 0.03 (33.70)	22.62 $\pm$ 0.11 (13.30)	13.93 $\pm$ 0.09 (19.94)
XCorrThreshold+Kmeans +Silhouette (Dim4)	11.3 $\pm$ 0.05 (12.61)	3.85 $\pm$ 0.03 (28.87)	20.28 $\pm$ 0.1 (22.27)	11.58 $\pm$ 0.08 (33.45)

Table 8: The DER and SER for the development and the test sets (DEVELSET and RT09), using MFCC and TDOA streams. Weight of MFCC stream for the baseline and MC methods is 0.9. The Weight of MFCC stream for XCorr-Pairs, DM, ER, PCA, XCorrThreshold+baseline, XCorrThreshold+DM, XCorrThreshold+Kmeans+DM and XCorrThreshold+Kmeans+Silhouette methods is 0.85. Weight of MFCC stream for the XCorr-Channels and XCorrThreshold+pca methods is 0.8. TDOA stream weight is 1-MFCC weight. The percentage error improvement on the baseline is shown in brackets.

The Kmeans+DM method had outstanding improvement with respect to the DM alone, reaching the best result obtained for the development set, however, this improvement is not so sound in the test set. The error is reduced, even when compared to the XCorrThreshold+DM, but is far from reaching the error rates of PCA or XCorr methods. The error with the Kmeans+Silhouette was slightly worse for the development set than the error with Kmeans+DM, however, when evaluated with the test set the reduction of the error is much higher.

It is worth noticing that all the methods developed select the pairs of channels independently of the distribution of delays in the last selected pair. For the XCorr methods the goal is to select the best quality pairs, and the PCA selects every possible pair but the EM, MC, DM and kmeans related methods are focusing on selecting discriminative pairs. With this in mind, for these methods whose objective is having a group of delays able to discriminate among speakers, the selection of pairs considering not only their individual characteristics but their relation with the previously selected pairs would be very beneficial. We could then avoid the selection of two very similar pairs, just choosing the best one and, for the selection of the next pair, look for a different one which actually gives different information. Though this idea has not yet been developed, the possibility of selecting pairs according to the discriminative power of the group of pairs instead of the discriminative power of the pair of channels is very interesting and would be developed in the future.

We want now to analyse more in deep what is happening with the recognized speakers in the diarization results. If for instance some method have an increase in the error due to a particular meeting in which there are many speakers or a high percentage of overlap.

In Table 9 we present the results of the XCorr-Pairs, XCorr-Channels and PCA systems, meeting per meeting, of the RT09 set. These methods either improve the SER meeting by meeting or keep it close to the baseline with the exception of the NIST 20080227-1501 using the PCA method. This is probably due to the pairs discarded in the XCorr-Pairs method that are included in the selection in the PCA method. The results are now only in SER because, as mentioned before, the contribution made by the system to

the diarization process is clearer and including both results would enlarge tables without real contribution to the analysis. In this table the Spnsp error (MISS speaker error+FA speaker error) which should be added to the SER to obtain the DER, is also included.

MEETING	Overlap (%)	SER Baseline	SER Xcorr-Pairs (Dim 4)	SER XCorr-Channels (Dim 3)	SER PCA (Dim 8)	SpNsp error
EDI 20071128-1000	3.06	0.28	0.46	0.39	0.50	6.90
EDI 20071128-1500	7.01	42.94	1.64	2.26	3.00	12.10
IDI 20090128-1600	3.66	6.46	1.33	5.71	1.10	4.80
IDI 20090129-1000	3.54	10.47	4.76	3.85	4.56	9.60
NIST 20080201-1405	14.76	43.75	44.68	34.66	28.84	19.30
NIST 20080227-1501	8.41	2.36	2.43	3.35	8.95	8.80
NIST 20080307-0955	3.53	31.03	13.89	13.53	30.29	4.70
ALL	5.58	17.40 ±0.10	7.82 ±0.07	7.71 ±0.07	9.35 ±0.07	8.70
Relative improvement over the baseline			55.06%	55.68%	46.26%	

Table 9: Results of each meeting for set RT09 and methods XCorr-Pairs, and PCA. Weight of MFCC stream for the baseline is 0.9. Weight of MFCC stream for PCA and xcorr-pairs methods is 0.85. Weight of MFCC stream for XCorr-Channels method is 0.8. TDOA stream weight is 1-MFCC weight.

In Table 10 we include the meeting by meeting results of the RT09 set for the Baseline, DM and PCA method combined with the rule of the XCorrThreshold.

By observing the results in Table 10 we can see the interesting case of the XCorrThreshold+PCA compared to the PCA method because the total error of the test set has not changed significantly, though many of the meetings that make up this set have varied their results. The general result is compensated mainly as a result of two

meetings, NIST 20080307-0955 and NIST 20080201-1405. While the first one improves significantly, the second one decreases its performance. If we compare the number of speakers detected by the PCA and the XCorrThreshold+PCA method (see Table 11 and Table 12), we observe that in the second case, the meeting that reduces its error has identified one more speaker whilst the meeting whose error rises has lost nearly all of them. It is important to know that this last meeting has a very high overlap (around double than of any other) which can greatly affect the performance of our system because we do not apply any specific strategy to deal with overlapped speech (although we count every frame of overlapped speech as error and the system would use the overlapped segments to train the single speaker models as if it were speech from only one speaker).

MEETING	Overlap (%)	SER XCorr threshold+ Baseline	SER XCorr Threshold+ DM (Dim 5)	SER XCorr Threshold+ PCA (Dim 6)	SER XCorr Threshold+ Kmeans+ DM (Dim 5)	SER XCorr Threshold+ Kmeans+ Silhouette (Dim 4)
EDI 20071128-1000	3.06	0.37	0.38	0.29	0.49	0.54
EDI 20071128-1500	7.01	24.32	43.12	1.78	36.30	21.59
IDI 20090128-1600	3.66	5.72	5.97	6.23	5.64	6.06
IDI 20090129-1000	3.54	10.47	4.13	3.75	4.14	2.27
NIST 20080201-1405	14.76	41.78	47.31	46.97	46.97	41.33
NIST 20080227-1501	8.41	4.31	11.68	8.27	8.40	10.30
NIST 20080307-0955	3.53	13.25	14.21	12.80	13.02	13.94
ALL	5.58	12.44 ± 0.08	15.44 ± 0.09	9.38 ± 0.07	13.93 ± 0.09	11.58
Relative improvement over the baseline		28.50%	11.26%	46.09%	19.94%	33.45

Table 10: Results of each meeting for set RT09 and methods Baseline, DM, PCA, Kmeans+DM and Kmeans+Silhouette combined with xcorrThreshold. Weight of MFCC stream for the baseline is 0.9, for DM, Kmeans+DM and Kmeans+Silhouette methods is 0.85 and 0.80 for PCA method. TDOA stream weight is 1-MFCC weight.

The XCorrThreshold+Kmeans+DM method has similar results to the XCorrThreshold+DM. However, there are some noticeable exceptions: the two last NIST meetings in the table and the EDI 20071128-1500 meeting. These better clusterings are responsible of the reduction of the error with respect to the XCorrThreshold+DM. If we checked how many speakers have identified each of the systems (Table 12 and Table 13), there are two which maintain the number of identified speakers while the NIST 20080307-0955 even recognize one extra speaker which is consistent with the fact that the error is reduced. Still, not always recognizing one extra speaker leads to an improvement in diarization, the NIST 20080201-1405 has a slightly lower error for the Kmeans+DM system and the number of identified speaker is reduced with respect to the DM system. This meeting has much overlap, and the extra speaker recognized probably only contribute to the diarization increasing the speaker error.

Finally, the Kmeans+Silhouette method has reduced the error with respect to the Kmeans+DM. The analysis meeting per meeting show that the EDI_20071128-1500 meeting has around 10 points lower error than the Kmeans+DM, which is explained by the fact that the Kmeans+Silhouette has recognized one extra speaker for this meeting.

Table 11 shows the number of identified, missed and false alarm speakers for the baseline, the XCorr-Pairs, XCorr-Channels and the PCA systems. We can see that the four systems have reduced the total number of missed speakers and with the exception of XCorr-Channels system, the false-alarm speakers (FA) too. This is consistent with the improvement in the average SER. But when looking at meeting by meeting results it can be observed that only the XCorr methods always keep the number of identified speakers equal to or higher than the baseline. It is worth noting that having more FA speakers does not always lead to an increase in SER because the extra speaker could have been assigned for very short time and the evaluation is made at frame level. That happens for instance at the EDI_20071128-1000 meeting, which has more FA speakers for the PCA than the XCorr-Pairs method but its SER is nearly the same.

Meeting	Baseline			XCorr-Pairs (dim 4)			XCorr-Channels (dim 3)			PCA (dim 8)		
	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA
EDI_20071128-1000	4		1	4		1	4		1	4		2
EDI_20071128-1500	3	1	1	4			4		1	4		1
IDI_20090128-1600	4		2	4		1	4		2	4		
IDI_20090129-1000	4		1	4			4		1	4		1
NIST_20080201-1405	3	2		3	2		4	1	1	4	1	
NIST_20080227-1501	6			6			6			5	1	
NIST_20080307-0955	6	5		7	4		8	3		6	5	
ALL	30	8	5	32	6	2	34	4	6	31	7	4

Table 11: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods XCorr-Pairs, XCorr-Channels, PCA and baseline.

In Table 12 we present the number of identified, missed and false alarm speakers for the XCorrThreshold+baseline, the XCorrThreshold+DM and the XCorrThreshold+PCA methods. In terms of number of speakers detected, the XCorrThreshold+baseline seems to outperform the rest, however, the resulting clusters of the XCorrThreshold+baseline are probably less pure than those of the XCorr-Pairs, XCorr-Channels or the XCorrThreshold+PCA method, because the total error of the system is greater than these three. As mentioned, though usually having matched more speakers leads to a lower error, the DER is computed by taking into account the number of frames of the meeting, not the number of speakers, and impure speakers can increase the DER. If the system has assigned frames from another correctly detected speaker to this new one, the error rate would increase. This happens very clearly for the NIST_20080201-1405 meeting (Table 12, fifth row). This meeting has a very high overlap (see Table 9, fifth row) which probably led the system to the extra speaker but using frames that previously belonged to any of the other detected speakers.

meeting	XCorrThreshold+Baseline			XCorrThreshold+DM (Dim 5)			XCorrThreshold+PCA (Dim 6)		
	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA
EDI_20071128-1000	4		1	4		1	4		1
EDI_20071128-1500	4		1	3	1		4		1
IDI_20090128-1600	4		1	4		1	4		2
IDI_20090129-1000	4		1	4		1	4		1
NIST_20080201-1405	4	1		2	3		1	4	
NIST_20080227-1501	6			5	1		5	1	
NIST_20080307-0955	8	3		7	4		7	4	
ALL	34	4	4	29	9	3	29	9	5

Table 12: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods Baseline, DM and PCA combined with XCorrThreshold.

The number of identified speakers for the Kmeans related methods are listed in Table 13. The Kmeans+Silhouette algorithm obtained the best DER of the two Kmeans methods and also the highest number of identified speakers. Note that the toughest meeting in terms of DER value (NIST_20080201-1405) is also the meeting with lowest number of identified speakers for the Kmeans+DM method. This meeting comprises more overlapped speech than the rest and its performance is usually very bad. The extra speakers recognized in this meeting by the Kmeans+Silhouette, led to the better result. Also, the EDI 20071128-1500 meeting has a perfect matching between the number of speakers recognized and the actual number of speakers in the meeting, and then, the diarization error for that meeting is reduced considerably when compared to the Kmeans+DM (Table 10).

Another dimension of this work is the evaluation of the performance of the algorithms in terms of the cost of computation. Many of the systems proposed reduce the number of pairs of which the delays have to be calculated. This reduces greatly the computational time required. Also the segmentation and agglomerative stage needs to deal with shorter features, though the difference is not always big enough to make any significant time saving.

Meeting	XCorrThreshold+Kmeans +DM (Dim 5)			XCorrThreshold+Kmeans + Silhouette (Dim 4)		
	SPK	MISS	FA			
EDI_20071128-1000	4		1	4		1
EDI_20071128-1500	4		1	4		
IDI_20090128-1600	4		1	4		2
IDI_20090129-1000	4		1	4		1
NIST_20080201-1405	1	4		3	2	
NIST_20080227-1501	5	1		5	1	
NIST_20080307-0955	8	3		8	3	
ALL	30	8	4	32	6	4

Table 13: Number of identified speakers (SPK), missed speakers (MISS) and false-alarm speakers (FA) for RT09 and methods Kmeans+DM and Kmeans+Silhouette combined with XCorrThreshold.

We have calculated the computational cost of all the systems developed. The results are shown in Table 14.

As expected the algorithms with low final dimension of the TDOA vector reduces greatly the time required to complete the whole diarization process. The time spent for the MC channel selection method was expected as the final dimension is 8 so for many cases the new algorithm is calculating more data than the baseline.

The PCA and Kmeans methods require a pre-extraction of all the possible pair combinations. Even when applying the requirement of the XCorrThreshold the total possible number of combinations is very high, much higher than the total number of pairs computed in the Baseline system

At the end, methods with the highest dimension do not obtain any savings in time while those with the lowest dimension reduce the computational time by about 20%.

The execution time depends heavily on the computer used for the execution and the characteristics of the meeting evaluated. Longer meetings would require more time

and meetings with high number of microphones even more. For example, with the baseline system a meeting with seven channels and duration of 44 minutes (NIST_20080227-1501) would need 97 minutes to extract the delay values and perform diarization. On the other hand, a meeting with 24 channels and duration of 44 minutes (EDI_20071128-1000) would need 2 hours and 25 minutes. Of course the process could be speed up using a faster processor but the same would occur with the methods that select a reduced amount of pairs. The savings in computational time using the selection methods are of special importance with meetings with very high number of channels. For instance, for the first meeting using XCorr-Pairs the process requires around 82% of the baseline time, but for the meeting with 24 channels the XCorr-Pairs needs only a 56% of the baseline time.

	Computational time of RT09
Xcorr-Pairs (Dim 4)	0.8 *Baseline time
Xcorr-Channels (Dim 3)	0.78*Baseline time
DM (Dim 3)	0.77 *Baseline time
ER (Dim 2)	0.82 *Baseline time
MC (Dim 8)	1.07 *Baseline time
PCA (Dim 8)	1.24 *Baseline time
XCorrThreshold+Baseline	1*Baseline time
XCorrThreshold+DM (Dim 5)	1*Baseline time
XCorrThreshold+PCA (Dim 6)	1.1*Baseline time
XCorrThreshold+Kmeans+DM (Dim 5)	1.1*Baseline time
XCorrThreshold+Kmeans+Silhouette (Dim 4)	1.12*Baseline time

Table 14: Computational time for the set RT09 using MFCC and TDOA streams relative to computational time of baseline system

The best systems both in performance and computational time are XCorr-Pairs and XCorr-Channels. Though XCorr-Channels may seem better, the actual error rate (SER) is not statistically different than the SER for the XCorr-Pairs. The savings in computational time are also very similar. Either of these two would improve the

diarization. However, the selection of channels would always avoid the selection of a specific pair if this is not part of the combinations of the selected channels, no matter how good this particular pair might be. Also, selecting channels instead of pairs, the number of pairs finally used could be doubled with every extra channel selected and we would lose the achieved time reduction. In this situation we will prefer to work with the XCorr-Pairs, and it will become the new baseline system for future works.

The xcorr methods achieve the best results because the delays calculated with the acoustic signals are more reliable. Maybe the microphones used are of higher quality or maybe the speaker is placed just in front of them so the voice signal arrives to the microphone in perfect conditions to be used for alignment with other channels. Though we will work from now on with the XCorr-Pairs method, because in average its performance is much better than the rest, the other developed methods (as de Kmeans based methods or the DM) actually focus on using the pairs that would place maximum number of voice sources in the room. Which, after all, is what we are looking for. We would then work with the XCorr-Pairs for other studies, but continue with the experiments with this second kind of approaches.

Also, the bin width used for the histograms in DM, ER and MC, as already mentioned, may be too high to actually differentiate clusters and should be reduced. Even if the speakers do not move, if they are placed near to each other the delay associated for each of speaker will be very similar to the others.

6 Alternative speaker diarization features

The TDOA features are very useful for the diarization task. However, there are a lot of speech databases that have not been recorded with several microphones, making it impossible to use delays. There are, for example, lots of data from the media, recorded in several environments and including numerous speakers. Many tasks of speech technologies could benefit from the automatic preprocessing of these recordings by a diarization system, which would inform about the location of every speaker in the recording. The MFCC features can be used independently of having a recording made with one or more microphones but the TDOA, which have proved to make a tremendous contribution, will not be accessible.

Though the main contribution of this thesis is the improvement of a diarization system through the enhancement of the calculated delay features, we want to study the possible contribution of some other features less frequently used in diarization. Initially our intention is not to combine them with the MFCC and TDOA, but work as if we were found in the situation of not being able to obtain delay features. For these experiments we are going to work with the CORALROM database, which is composed of recordings from different media programs. The number of speakers is very variable as well as the style used. Also, unlike the RT database, the recordings can contain (and they will) sections with music or a lot of noise. Moreover, we will only have one recording per meeting. This database, unlike the RT, contains more variations of speaking styles as well as noises or music intervals which are very common in programs from the media. Working with this database would give us a closer approximation of the expected performance of the diarization system if we would like to analyse any of the many

recordings available nowadays online. Also, this database was used in a previous work focusing on diarization of different style voices and for coherence with it we want to continue working with the same meeting recordings.

The possible number of features is very large, however, this is only an exploratory work and for that reason we have decided to experiment only with a limited group of glottal features used in speech synthesis, in particular in expressive speech synthesis (Lorenzo-Trueba, et al., 2012). The objective is not only to perform a good diarization taking advantage of these features but also study how adequate could be our diarization system to extract data useful for other tasks as speech synthesis.

6.1 Glottal features

Trying to improve the DER of the diarization system if the TDOA are not an option, we have decided to test some of the glottal features studied for expressive voice synthesis in (Lorenzo-Trueba, et al., 2012) as features for the diarization system. As the authors explain, the GlottHMM (Raitio, et al., 2011) used for the extraction of the features is a vocoding technique that was developed for parametric speech synthesis. It is based on decomposing speech into the glottal source signal and vocal tract transfer function through iterative adaptative inverse filtering (IAIF).

The signal is initially windowed with a rectangular window of 25ms. Then the glottal inverse filtering is performed, which uses a series of Linear Predictive Coding (LPC) filters (Makhoul, 1975) to iteratively estimate and cancel out the effects of the vocal tract and the lip radiation, delivering the glottal flow of the original speech signal

This IAIF module has then decomposed a voiced speech signal into the glottal source signal and the p^{th} order all-pole model of the vocal tract. This order p is set to 30.

The spectrum of the glottal source signal is estimated with a tenth-order LPC to capture the spectral properties of the excitation, mainly the spectral tilt, but also the more detailed spectral structure of the source.

This two all-pole models are then converted to the line spectral frequencies (LSF) (Soong & Juang, 1984), a parametric representation of LPC information well suited to be used in a statistical HMM system. The vocal tract is represented by the Line Spectral Frequency (LSF) vector (with 30 LSFs). The spectral tilt of the glottal source will be called LFSOURCE and it is represented by the 10 LSFs from the second model.

In addition, GlottHMM also extracts the F0 and harmonics to noise ratio (HNR) of the glottal source. The voiced-unvoiced decision is made based on the energy of the low frequency band (0–1 kHz) and the number of zero-crossings in the frame. The frames determined as unvoiced are marked as zeros.

In order to capture the degree of voicing in the excitation, the harmonic-to-noise ratio (HNR) of the glottal source signal is analysed in four bands (0–2, 2–4, 4–6, 6–8 kHz). The HNR measure indicates the proportion of the periodic vibratory glottal excitation compared to the aperiodic noise excitation of the voice source. The HNR is determined by first evaluating the fast Fourier transform (FFT) of the windowed speech signal, and evaluating the cepstrum separately for each frequency band. The degree of the harmonicity is then indicated by the strength of the cepstral peak, whose location is defined by the fundamental period. Finally, the HNR is defined for each band as the ratio of the maximum value of the cepstral peak to the averaged value of other frequencies of the cepstrum.

Finally, in addition to the standard features utilized in GlottHMM, two additional features has been added: the magnitude differences between the 10 first harmonics of the voice source and the Normalized Amplitude Quotient (NAQ) (Alku, Bäckström, & Vilkman, 2002). The NAQ is defined as a method to parametrize the glottal closing phase using two amplitude-domain measurements from the glottal flow signals estimated by inverse filtering. In this technique, the ratio between the amplitude of the estimated glottal flow and the negative peak amplitude of the flow derivative is first computed using the concept of equivalent rectangular pulse. More detailed information about the estimation process of these parameters is included in (Soong & Juang, 1984) and (Alku, Bäckström, & Vilkman, 2002).

Summarizing, the list of features extracted that we are going to study are:

- IF0 (log f0)
- HARMONICS (magnitude differences between the first 1 harmonics)
- HNR (Harmonics to Noise Ratio)
- H1H2 (magnitude differences between the first 2 harmonics)
- LSF (Line Spectral Frequencies)
- LSFSOURCE (spectral tilt of the glottal source)
- NAQ (Normalized Amplitude Quotient)

The system delivers in fact not the F0 but its logarithm, which is the reason why we refer to it as IF0. This reminds the lif0 (logarithm of the interpolated F0) features used for one of our previous systems in the NIST evaluation of 2009 (Pardo, Barra-Chicote, San-Segundo, Córdoba, & Martínez-González, 2012). Since, in a previous version, it was used already a variation of F0 as input feature, we have decided to carry out also experiments with the F0 and the transformations that were applied in that system to analyse its utility for this database.

For the extraction of the F0 we have set a minimum value for F0 of 60 Hz. The system extracts the F0 following the algorithm of (Talkin, 1995). For each frame we take a window of about 7.5ms and calculate its normalized cross-correlation with the speech signal in windows at various “lags” in the future. Lags range from less than 2ms (for 500Hz) to more than 20ms (for 60Hz)

Once extracted, the logarithm of F0 was calculated (unvoiced regions are set to 0 in the If0 feature vector). For the unvoiced part of the signal a constant value of F0 was used which is the average of the last value of the previous voiced region and the first value of the following voiced region. This feature will be called lif0 from now on. Similarly, a third feature was the first derivative of the logarithm of the interpolated F0 and called from now on dlif0. The F0 features are therefore: If0, lif0 and dlif0.

Finally, we have concatenated, in the same stream, all the rest of glottal features with our Lif0. This was the feature that worked better for the previous system and we expected to have better results with it than with the other transformations of F0.

To analyse how suitable the features proposed are for the diarization task we have performed several experiments with the diarization system. As mentioned, we cannot make use of the TDOA because we have only one single audio channel per meeting. However, the MFCCs features are calculated and used for the experiments in combination with all the other features proposed. It will be then necessary to adjust the weight we apply to every feature as we did with the TDOA. This is important because one feature could give some information by itself but not as much as another (like the MFCC) and although working alone may improve the results, their use with MFCC could not make any difference to the final result. The experiments explore then various weights for the glottal features.

There is another variable we need to fix. The number of gaussians to train the initial models with the glottal features. MFCC use 5 gaussians initially and then grow when two clusters are merged. For these new features we have decided to carry out experiments varying from 1 to 5 the number of gaussians to train the initial models.

In the charts in Figure 36 the results in DER for every feature studied and for 1 to 5 gaussians for the initial training are included. In every case experiments were carried out with MFCCs also. The database used for these experiments is the C-ORAL-ROM-DEVELSET.

First of all we should point out that the dlif0 and NAQ figures in Figure 36 seem to present results only for the use of 5 gaussians because the numerical results are equal, and the other lines (corresponding to the use of 1 to 4 gaussians) are just the same.

By analysing the rest of the results from Figure 36 we can see some interesting points. The HARMONICS feature is the only one which benefits from the application of a greater number of initial gaussians to train their speaker model, in fact, results for these features are really bad until we reach the 4 and 5 gaussians, when the error rate turn to be one of the best. The same happens for the HARMONICS merged with lif0.

Figure 36: Experiments with the whole range of glottal and F0 features across the weight applied to the new feature and for #initialgaussians to train the model ranging from 1 to 5. Experiments were carried out in combination with MFCC stream. Weights applied to the two streams sum1.

The HNR feature on its own performs reasonably well for 2 initial gaussians, but it is when combined with lf0 when it reaches the best numerical points. There is a clear area for weights between 0.1 and 0.25 in which this feature reduces the error of the system. The other good feature could be the HARMONICS but, unlike HNR, the points with the lowest error (0.2 and 0.4) are separated for several points of higher error. Moreover, even in the best points the DER achieved is not as good as the HNR+lf0.

There is also another particularity. The lf0 feature has a point of minimal DER in 0.15. It could be an outlier because there is only one point with significantly lower DER

and then the error raises but, as its behaviour is much better than lif_0 , we wanted to know what would happen if we merge this lif_0 feature with the others (instead of merging lif_0). Just for checking we computed the experiments combining lif_0 with HNR (our best feature when combined with lif_0). The results are included in the same figure, in the last position.

The best DER is obtained for 3 initial gaussians, and there is, as with HNR+ lif_0 , a flat area around the 0.15 of weight applied. However, these results are not either better than the HNR+ lif_0 or the lif_0 by itself, so we decided not to continue with the experiments merging lif_0 with the rest of the features.

Nearly every feature gives some information, as the error of the system using only MFCC (weight of the glottal feature equal to 0), is improved for many features across different parameters. However, results are not in general very stable and the reduction in the error is modest for most of them. Experiments are carried out using MFCCs as well as the glottal features. Both features carry information about the identity of the speakers. In our previous experiments we combined information of different nature (position vs identity of the speaker) which could account for the more modest improvement of these experiments. Also the database is composed of meetings of various nature, with clean speech in some places, but certainly not in the whole meeting, and even including music. The extraction of the MFCCs and glottal parameters can be greatly affected by this various types of data causing the instability of the results.

The $dlif_0$ features have very similar values for all the frames calculated, and the same occurs for the NAQ features. The low variation of this features may be the reason why the training with more gaussians is not causing any improvement in the results. In fact, when NAQ is combined with the lif_0 feature, the results are similar to those obtained for the lif_0 feature alone (at least more similar than the experiments with other glottal features combined with lif_0).

We can conclude that the glottal features could be of some use to improve a diarization system in particular the f_0 and the features related to the harmonics. In the evaluated case HNR, HARMONICS and lif_0 are the most promising combinations.

6.2 Music detection experiments for speech synthesis

Many of the recordings from the media have music as well as speech. In the C-ORAL-ROM meetings we encounter noises, music and all kind of non-speech events so common in the real life recordings. These noises and music degrades the overall performance of the system. In this part we want to evaluate how much the system could benefit from a music detection preprocessing, especially because with this database we want to use it to extract clean speaker segments for generating automatic personalized speech synthesis voices. Consequently we want to delete any segment that would corrupt our voices. Music and noises are among the events to avoid, as well as speech overlapped with either music or noises.

The VAD module usually labels the segments of music as speech, and then the diarization system assigns them to one speaker, corrupting the model and at the end, adding more error to the system.

In order to use speech collected from the media or from media sharing sites, speech synthesis systems must be robust also to the variation of the acoustic and environmental conditions. The system must be able to robustly cope with noisy automatic-speech-recognition-processed corpora and with challenging data such as interviews, debates, home recordings, political speeches, etc. which will include most likely all kind of noises, overlapped speech and also music. The use of diarization techniques for speaker-turn segmentation will allow the system creating homogeneous voices from heterogeneous recordings, because the number of speakers would be automatically estimated in a fully unsupervised way, and language-independent diarization techniques automatically could provide the temporal labels of the turns of a certain speaker.

In previous work (see (Lorenzo-Trueba, et al., 2012)), a methodology for constructing such tools was sketched, and preliminary experiments with a multi-style database were presented. In that work, only MFCC features were used to perform the

diarization without initial noise reduction of the audio, and the final goal being the extraction of voices suitable for speech synthesis, in particular speech synthesis from different styles. For that task it would become more important to have enough clean data from one speaker than having short segments from different speakers rightly assigned. This means that despite the fact that the DER could be higher, if speakers segmented have mainly clean speech (even if we have had to discard much data) the system would be more desirable. The objective is to develop a system able to take any data and, unsupervisedly, extract enough good speech to be able of using it to synthesize voices, this means, maximize precision of the clusters. This work was presented in (Martínez-González B. , Pardo, Echeverry-Correa, & J.M. Montero, 2014).

Prior to the study of the Music detection module we want to continue with the study in (Lorenzo-Trueba, et al., 2012), by exploring the effects of applying also the noise reduction (wiener filter in Figure 37) and the lif0 features. To combine the MFCCs with the lif0 features the system needs a weight to be applied to each of these vectors. These weights are complementary, adding up to 1.

There is also another consideration, the database used has meetings from different sources (as it would be media data), as interviews, talk shows... Some of them have a lot of speakers, many more than the meetings from RT database. The initial number of clusters in which the system segments the whole meeting is set to 16, and we know that some meetings in this new database have more speakers. In this case some speakers will be lost, as the maximum number of detected speakers would be lower than the actual number in the meeting. Nonetheless, those speakers have few and very short utterances, and we assume that these speakers will not significantly decrease the overall DER. To confirm this supposition we perform experiments doubling the number of initial clusters, which would cover any number of speakers in the current database.

Experiments showing results with and without the application of the noise reduction module and beginning with a segmentation of 16 and 32 clusters are shown in Figure 37. The results show clearly that starting with a much higher number of initial clusters would only degrade the performance. As mentioned, the extra speakers may

participate very short time, and trying to find so many more speakers only confuse the system. About using the Noise reduction (NR in the graph) the system has two possibilities, if using it apply a weight of 0.05 to the lif0 stream, or if not using apply a higher weight (0.15) to the stream.

Figure 37: Results for the C-ORAL-ROM-DEVELSET when applied and not the noise reduction module and with 16 and 32 clusters for the initial segmentation. Systems with noise reduction are labelled with NR, and k is the number of initial clusters. Experiments were carried out with MFCC and lif0 features, which weights added equals 1.

After the noise reduction experiments we explore music detection features. Media data such as programs of interviews or talk shows can very easily contain fragments of music as well as noises that will degrade the performance. There are several previous works on speech and music segmentation. Many of them focus on the use of different features that would help in the discrimination between music and speech. This is the case of (Izumitani, Mukai, & Kashino, 2008), (Gallardo-Antolin & Montero, 2010) or (Panagiotakis & Tziritas, 2005). Other works like (Lavner & Ruinskiy, 2009) focused in system architecture to segment speech and music.

In (Gallardo & San-Segundo, 2010) the UPM-UC3M system for the Albayzin evaluation 2010 on audio segmentation is presented. The best combination of features for the segmentation of music are MFCC, CHROMA coefficients (see (Bartsch &

Wakefield, 2001)), and Entropy features (Misra, Ikbai, Bourlard, & Hermansky, 2004). In this work we have applied this algorithm for the music segmentation.

There are five classes to be recognized: speech, speech+noise, speech+music, music and others. As our database is not labelled with these classes, we cannot train our own models for each of them, so, for the recognition, we used the same models that were trained in the experiments in (Gallardo & San-Segundo, 2010).

Once the segmentation is carried out, we will only remove “music” and “others” segments from the speech segments detected by the VAD module (see diagram in Figure 38). From now on, if not specified the opposite, when the music detection is applied these segments detected will be removed.

Figure 38: Block diagram of the system with Music detection

The music detection module would be very useful to discard the segments of the recording whose features associated are not reliable (noisy parts and music would corrupt the feature extraction procedure). Three kind of experiments have been carried out: apply only the music detection to the baseline system, apply it in combination with lif0 and apply it in combination with lif0 and the noise reduction module. For these experiments the lif0 weight has been set to 0.05 when we apply noise reduction and 0.15 when we do not (these were the two best systems found in initial experiments with lif0 and noise reduction).

As mentioned before what we would really want to achieve at the end with these experiments is to improve the system so it could serve to extract automatically segments of clean speech suitable for synthesis. Discarding many segments then should not be an

issue as long as those recovered are of sufficient precision. We have then calculated also the precision and recall of the clusters and we have included a merit figure too, which weights the precision by two thirds and the recall by one third. This merit figure is intended to help us compare between two similar systems. After all, although precision is the factor we want to maximise, we want to do it discarding as few data as possible. The three new measures that we will analyse will be then:

$$Precision = \frac{T_p}{T_p + F_p} \quad (6-1)$$

$$Recall = \frac{T_p}{T_p + F_n} \quad (6-2)$$

$$merit\ figure = \frac{2}{3}Precision + \frac{1}{3}Recall \quad (6-3)$$

Where T_p refers to the true positives (frames correctly assigned to the speaker), F_p to the False positives (frames incorrectly assigned to the speaker), and F_n is the false negatives (frames from the speaker not assigned to it).

There is also another consideration which is the insertion penalty in the music recognition system. This insertion penalty is a parameter that regulates the percentage of music segments that the system can include in the segmentation. The higher the term the higher the number of segments that could be labelled at the end as “music” or “others”.

In Figure 39, the precision, recall and merit figure for the three studied systems across different insertion penalty values is presented. These three systems reach the best precision values with insertion penalty of 5 or 15. For the system that uses lif0 and applies noise reduction and music detection the best precision is clearly obtained with insertion penalty of 15. For the system with lif0 and no noise reduction the best precision is reached with insertion penalty of 5. The baseline system with just MD applied had their best precision value with insertion penalty 15 though this value is only slightly better than the precision value with insertion penalty of 5 and the recall drops heavily

between this two options. In this situation we had preferred to work with insertion penalty of 5 for this combination. Higher values of this insertion penalty term would allow more changes between classes, which means, at the end, more segments categorized as music and therefore, discarded. In general, even if we lose many speech segments wrongly labelled as music, as long as we discard enough real music segments, the clusters generated with the remaining segments will be purer. Removing more segments, could reduce the amount of speech recovered but, as long as the precision of the clusters increases and the amount of remaining data is enough for the speaker training, the voices generated with these clusters should be more accurate. Of course, if we remove too much speech we are not only reducing the data available for voice building, but the models trained by the diarization system will be less accurate and, therefore, the final segmentation will have more errors.

Figure 39: Precision, recall and merit figure versus insertion penalty of the music recognizer for the C-ORAL-ROM DEVELOPSET. $F0_weight=0.05$ for system with noise reduction (NR) and 0.15 for system without it. MD stands for Music detection applied.

The best numerical results for all the combinations evaluated are included in Table 15. In the table, different options appear, the Baseline, which does not apply noise reduction and uses only MFCC features, and the Baseline plus the modifications made on it. These modifications can include Noise Reduction (NR), use of lif0 (F0), number of initial speakers (K) and Music detection (MD).

The best numerical result (in precision) obtained for the development set is included in the sixth row of Table 15, (system with noise reduction and insertion penalty of 15). However, in the fifth and seventh row, the best result for the two other systems with music detection are presented (no noise reduction, insertion penalty term of 5 and use or not of lif0 features). We can see that even though the precision values are a bit lower, the merit figure of these two systems surpass that of the system with the best precision value (in which we applied noise reduction).

The noise reduction module apparently affects highly to the recall of the system. This can be due to the high insertion penalty defined for the music detection module when using also noise reduction. Though that yielded the best possible precision among the experiments evaluated, the higher insertion penalty would mean higher probability of labelling an audio segment as music, and thus, higher number of speaker segments lost. The number of False negatives (segments of one speaker not associated to him) would grow, making the recall to be reduced.

For comparison purposes we have included also results with the baseline, plus noise reduction, lif0 and Music detection module, when the insertion penalty is 5 (the same insertion penalty used for the two systems without noise reduction). Precision result decreases while recall increases, as well as the merit figure, but not enough to reach the performance (in merit figure) of any of the other two systems where no noise reduction is applied.

Note that the system with noise reduction and lif0 has a merit figure comparable to the best ones. However, this high merit figure is due to the higher recall value. As we are looking to maximize precision, having the same merit figure, the chosen systems would be those with good precision.

System	Insertion penalty of MD module	lifo weight	DER	Precision	Recall	Merit figure	Precision error improvement (%)
Baseline		0.0	19.34	84.66	83.28	84.2	
Baseline+NR+lifo+K32		0.05	20.33	84.14	82.41	83.56	-3.39
Baseline+NR+lifo		0.05	18.87	85.64	83.87	85.05	6.39
Baseline+lifo		0.15	18.58	85.43	84.04	84.97	5.02
Baseline+MD	5	0.0	21.96	88.09	78.86	85.01	22.36
Baseline+NR+lifo+MD	15	0.05	23.52	88.42	77.04	84.63	24.51
Baseline+lifo+MD	5	0.15	21.88	88.17	78.94	85.09	22.88
Baseline+NR+lifo+MD	5	0.05	22.49	88.09	78.24	84.80	22.35

Table 15: Results for the C-ORAL-ROM DEVELOPMENT. Relative precision error improvement is calculated over the baseline. K stands for the initial number of hypothetical speakers, K=16 if nothing indicated. NR stands for noise reduction algorithm and MD stands for music recognition algorithm.

Our task implies maximizing precision but we want to maintain a certain level of recall and considering the variation in the merit figure we cannot yet decide between these options. Experiments with the test set will show if one of them turns clearly better.

6.2.1 Evaluation with the test set

Following experiments will contrast the results of the development set with the C-ORAL-ROM TESTSET, whose meetings were not used until now.

The first modification tried over the development set was to increase the initial number of hypothetical speakers. This modification did not improve diarization for the development set, probably because the extra speakers in some meetings speak for very short time. Therefore, the following experiments were not tested with a higher number of initial clusters but with the usual 16.

The second group of experiments was focused on optimizing the systems using or not lif0 and a noise reduction Wiener filter. At this point it was not clear if we should use or not the noise filtering. Both systems delivered similar performance in precision and merit figure. Thus, we decide to keep both systems and check its performance with the test set

Finally, in the last experiments with the development set, we tried to take advantage of a music detection module. This module is applied alone and in combination with the two previous ones, adjusting for each one the insertion penalty term. In Table 15 we can see that the three of them achieved high relative precision error improvement (24.51%, 22.88% and 22.36%). For comparison purposes, we included also the performance of the system with the best precision result (Baseline+NR+F0+MD) that due to the high insertion penalty had lower recall, but with insertion penalty of 5 achieving a 22.35% of precision error improvement).

The best systems evaluated with the development set still had very similar precision and the best one degrades heavily its recall, and, consequently, its merit figure, so we decided to check all of them with the test set.

The experiments we have carried out with the test set are included in Table 16. When there is no music detection, the use of lif0 decreases the precision error in 17.12% for the system with noise reduction, which is much more than the 3.99% achieved when no noise reduction is applied (second and third row in Table 16). However, the use of noise reduction, as we have seen before, reduces heavily the recall of the system, and the merit figure of these two systems turns very similar (86.14 vs 86.04).

When we include the music detection module, the system with noise reduction has the same problem we have been noticing with the development set. The recall is heavily reduced by the combination of noise reduction and music detection, this time affecting the precision as well, which is increased much less than the two other systems with music detection.

The two systems without noise reduction outperform clearly the rest because not only precision increases, but also the merit figure. In this case, the use of music

detection alone is slightly better than its combination with the lif0 features. The precision, in this case, turns 90.62%, and recall decreases to 84.38%, vs precision 90.22% and recall of 84.01% for the system without noise reduction and with lif0 and music detection. As both the precision and the recall are higher for the system that do not use lif0, the merit figure is much better too.

If the noise reduction is also applied the precisions are 87.76% and 88.95% and the recalls are 80.81% and 78.92% for the system with lif0 and insertion penalty of 15 and 5 respectively.

The best evaluated system turned to be the baseline+MD alone. We obtain with this system a relative decrease of the precision error of 39.64% over the test set.

Though reducing DER was not the primary objective, we can see also, that for the test set, the use of the music detection system decreases the DER value of the baseline in more than one point. This means that we are not discarding much clear speech which is very useful in the training of the speakers because the diarization system can model the speakers better.

System	Insertion penalty of MD module	F0 weight	DER	Precision	Recall	Merit figure	Precision error improvement (%)
Baseline		0.0	18.68	84.46	87.33	85.42	
Baseline+NR+lif0		0.05	19.36	87.12	84.18	86.14	17.12
Baseline+lif0		0.15	18.04	85.08	87.97	86.04	3.99
Baseline+MD	5	0.0	17.28	90.62	84.38	88.54	39.64
Baseline+NR+lif0+MD	15	0.05	20.92	87.76	80.81	85.44	21.23
Baseline+lif0+MD	5	0.15	17.66	90.22	84.01	88.15	37.06
Baseline+NR+lif0+MD	5	0.05	22.57	88.95	78.92	85.60	28.89

Table 16: Results for the C-ORAL-ROM TESTSET. Relative precision error improvement is calculated over the baseline. NR stands for noise reduction algorithm and MD stands for music recognition algorithm

Finally, in Table 17, the results obtained with different styles of the test set are presented. The precision in speaker diarization ranges from 76.18 % for reportages to 99.93% for meteorology recordings.

Note that meteorology meetings are remarkably easier than the rest, having only one speaker to detect. On the other hand the set of reportages is more difficult (it is the only one with precision below 90%) due to noise and the high number of different speakers that can participate (see Table 4). In future work new strategies should be drawn in order to tackle this problem.

Style	Precision	Recall	Merit figure
Interviews	92.48	91.38	92.11
Meteorology	99.93	79.18	93.01
News	96.83	93.12	95.59
Reportages	76.18	72.94	75.10
Scientific press	94.01	79.60	89.21
Sports	91.08	90.07	90.74
Talk shows	92.67	81.59	88.98
ALL	90.62	84.38	88.54

Table 17: Precision, recall and merit figure for the different styles in the C-ORAL-ROM TESTSET.

6.3 Adding glottal features to system with music detection

After experimenting that some of the glottal features (as the HNR or the HARMONICS) might be useful, in this section we present the study of these preselected best features (section 6.1) joined with the music detection system used in the previous section. The music detection system will be used for unsupervised detection of music

intervals as well as other intervals not recognized as speech and their removal for the posterior stages. Once this is done we extract the rest of the features.

As in the previous section, the comparison will be done in terms of precision and recall. The suppression of intervals of music could only benefit the whole system but we can expect that some intervals of speech, probably those of bad quality, are going to be discarded as well as those of actual music. Our main target is to recover speech suitable for synthesis so we do not worry too much about those intervals of speech discarded, as long as we recover enough speech. It is for this reason that we decided to focus on precision results (and recall a bit less) instead of focussing on DER results to make the analysis in this part. The merit figure used to compare two systems is in equation (6-3).

The graphs of the results for the case of using 1 or 2 gaussians and applying music detection are in Figure 40. As values are too close one to another we have included in the right part a zoom of the area between weights 0 and 0.25. In Figure 41 are included results for the cases of using 3, 4 and 5 gaussians. Note that the only features analysed with 4 and 5 gaussians are the HARMONICS and the HARMONICS+lif0 (which were not analysed with only 1, 2 or 3 gaussians). The previous experiments with these features showed that performance with less than 4 gaussians was really poor so we decided to explore only these two number of gaussians for the initial models of the HARMONICS features. Remember that experiments were always carried out with one of the presented new features in combination with MFCC and the music detection.

It still happens that in the case when we use a single Gaussian the best feature we have is lif0, followed by HNR. Results for lif0, our traditional feature, does not reach any of these two.

With two and three Gaussians we can also see that the HNR (not concatenated) has a somewhat larger working area than other features.

In these experiments, even though individually the lif0 feature seems to work well, we have not tried concatenating it with HNR because we had already done it when we did not use music detection and results were not promising.

The features that yielded highest merit figures are lf0 with 1 gaussian, HNR with 2 and HARMONICS+lf0 with 5. The HARMONICS have good performance on their own but they are slightly surpassed by the feature with the lf0 concatenated so we have decided to continue experiments only with it.

The results show that the HNR concatenated with lf0, which achieved the lowest error in absence of music detection, would have the same working point (2 Gaussians for the initial model and weight 0.2) in the system where we do apply music detection. This result is, however, restricted to a single point. It does not happen, as we saw before, that there is an area covering several weights with maximum values of merit figure (equivalent to minimum values of DER). This is not even the feature which achieves the best value of merit figure. Lif0 is worse than lf0, HNR, HARMONICS or HARMONICS+lf0, but we keep it in following experiments for an easier comparison with the previous system. Both lf0 and HNR reaches its best performance for a weight of 0.05 or 0.1 and there is no significant difference between the two points. However, as the HNR show an area of low error variation around 0.1, this will be the weight chosen for it. Remember that the noise reduction module is not utilized, just as we did in the previous section, only the music detection.

For the HARMONICS+lf0, the lowest error is achieved with weight of the attribute equal to 0.2 and 5 gaussians for the initial models (see Figure 41)

The best system using MD and lf0 has worse results than the system that uses the HNR, lf0 or HARMONICS features. The merit figure is higher with these new features and, in the case of the HNR, it remains quite stable with the variation of the weight applied to it. The numerical results for the best points evaluated are included in Table 18. Experiments have been carried out with the test set to evaluate these three new combinations.

Figure 40: Merit figure for the C-ORAL-ROM DEVELSET through various weights applied to the glottalFeature, and using 1 and 2 gaussians for the initial model created with the glottal features. Music detection was applied to discard all intervals of music. All experiments have been done with MFCC features too. The weight applied to MFCC features is 1-weight_glottalFeature.

Figure 41: Merit figure for the C-ORAL-ROM DEVELSET through various weights applied to the glottalFeature, and using 3, 4 and 5 gaussians for the initial model created with the glottal features. Music detection was applied to discard all intervals of music. All experiments have been done with MFCC features too. The weight applied to MFCC features is 1-weight_glottalFeature.

The three extra features improve the results for the development set in more than a 20 percent. The result for the test set are included in Table 19 alongside with the best combinations of previous features for the task at hand. The l_f0 has been checked for the two best weights, 0.05 and 0.1, as they performed very similar in the development set. However, it is noticeable the great variation in the test set for the two weights. We wondered if this great change could be attributable to only some particular meetings whose performance was disastrous but, although there were some that worked significantly worse with a weight of l_f0 equal to 0.05, the differences seem not to be linked to a specific meeting or style (see Table 20 or Table 21). Though results with l_f0 and weight equal to 0.1 are good, the performance of the HNR or HARMONICS+ l_f0 is much better. Having other features more stable and whose performance is as good as l_f0 , we would rather use them.

System	Insertion penalty of MD module	Glottal weight	DER	Precision	Recall	Merit figure	Precision error improvement (%)
Baseline		0.0	19.34	84.66	83.28	84.2	
Baseline+ l_f0		0.15	18.58	85.43	84.04	84.97	5.02
Baseline+MD	5	0.0	21.96	88.09	78.86	85.01	22.36
Baseline+MD+ l_f0	5	0.15	21.88	88.17	78.94	85.09	22.88
Baseline+MD+HNR	5	0.1	21.55	88.53	79.27	85.44	25.23
Baseline+MD+ l_f0	5	0.05	21.47	88.63	79.35	85.54	25.88
Baseline+MD+ l_f0	5	0.1	21.49	88.61	79.32	85.52	25.75
Baseline+MD+HARMONICS+ l_f0	5	0.2	21.21	88.93	79.62	85.83	27.84

Table 18: Results for the C-ORAL-ROM DEVELOPSET for the best combinations of features. MD stands for music detection,

In Table 20 and Table 21 are presented the values per meeting and per style respectively for the new systems evaluated in this section. Although the Baseline system is improved by the use of HNR or HARMONICS+lif0, the MD alone achieved better results for the test set in many of the style of meetings evaluated. In fact, with HNR only the sports recordings obtain higher precision. On the other hand, there are more styles that improve their performance by the use of HARMONICS+lif0.

System	Insertion penalty of MD module	Glottal weight	DER	Precision	Recall	Merit figure	Precision error improvement (%)
Baseline		0.0	18.68	84.46	87.33	85.42	
Baseline+lif0		0.15	18.04	85.08	87.97	86.04	3.99
Baseline+MD	5	0.0	17.28	90.62	84.38	88.54	39.64
Baseline+lif0+MD	5	0.15	17.66	90.22	84.01	88.15	37.06
Baseline+MD+HNR	5	0.1	17.97	89.88	83.69	87.82	34.88
Baseline+MD+lf0	5	0.05	30.58	76.34	71.06	74.58	-52.25
Baseline+MD+lf0	5	0.1	17.97	89.88	83.69	87.82	34.88
Baseline+MD+HARMONICS+lif0	5	0.2	17.43	90.46	84.23	88.38	38.61

Table 19: Results for the C-ORAL-ROM TESTSET for the best combinations of features.

The music detection algorithm and some glottal features (as the HNR and HARMONICS+lif0 evaluated) can be used to improve diarization in absence of TDOA features. Also, as we have seen, if the task requires a specific format of the segments of speech (like being purer rather than longer) the diarization should be adjusted. The use of the music detection have proved to be enough to reach the highest performance, without the need of using any other feature. However, some styles are still better when other features (HNR or HARMONICS) are used. If the task is focusing in a specific kind of

meeting could be interesting to apply these features to the diarization. Moreover, the use of HNR, lf0 or HARMONICS+lf0 is slightly worse for the test set but are the best systems in the development set.

Meeting	Baseline+MD+H NR system precision (weight HNR=0.1)	Baseline+MD+lf 0 system precision (weight lf0=0.05)	Baseline+MD+lf 0 system precision (weight lf0=0.1)	Baseline+MD+H ARMONICS+lf0 system precision (weight HNR=0.1)
emedin01	85.86	94.26	82.48	85.71
emedin05	99.36	98.697	98.43	98.71
emedmt01	99.93	99.927	99.92	99.93
emednw01	96.99	94.099	96.99	96.99
emednw04	96.67	79.5	97.45	97.35
emedrp03	65.69	43.836	66.11	67.90
emedrp05	79.76	57.775	79.76	79.76
emedsc04	93.73	71.534	93.84	94.01
emedsp01	85.81	89.669	86.33	82.93
emedsp03	99.07	98.798	98.92	98.89
emedts01	85.94	76.212	88.83	88.93
emedts02	96.90	38.115	96.90	96.93
emedts05	95.51	77.01	95.51	96.95
emedts10	88.26	74.352	88.56	93.68
ALL	89.88	76.342	89.88	90.46

Table 20: Precision for the C-ORAL-ROM TESTSET meeting per meeting

Style	Baseline+MD system (weight HNR=0.1)	Baseline+MD + HNR system precision (weight HNR=0.1)	Baseline+MD +lf0 system precision (weight lf0=0.05)	Baseline+MD +lf0 system precision (weight lf0=0.1)	Baseline+MD +HARMONICS +lf0 system precision (weight HNR=0.1)
Interviews	92.48	92.15	96.33	89.92	91.77
Meteorology	99.93	99.93	99.93	99.93	99.93
News	96.83	96.83	86.81	97.22	97.17
Reportages	76.18	72.71	50.79	72.92	73.81
Scientific press	94.01	93.73	71.53	93.84	94.01
Sports	91.08	91.66	93.70	91.88	89.97
Talk shows	92.67	92.05	64.72	92.78	94.36
ALL	90.62	89.88	76.34	89.88	90.46

Table 21: precision for the C-ORAL-ROM TESTSET per style of the meetings

6.4 Features for overlap detection

Most diarization systems have not solved yet the problem of the overlap, and noises and other corrupting artefacts could lead to incorrect assignment of speech segments. Several speakers speaking together can only corrupt the models trained for a particular speaker. Noises of any nature, including for instance laughter or slamming of doors, can do the same.

In this thesis we intend to study the possible use of the TDOA and the parameters related (as the xcorr) to the detection of not suitable segments to model single speakers. We will focus on having enough frames of clean speech to model reliable speakers and leave the rest just to be labelled with the most likely speaker. Studies about overlap have shown that impure models can degrade heavily the performance of the system and have mainly focused on correctly detecting and labelling overlap. We search for a method to

detect enough clean speech for diarization. In this initial approach we do not intend to label overlapping segments with more than one speaker, therefore, if the system excludes from the training many more segments than the actual overlapped speech, this is not a problem. These excluded segments will be labelled for the final segmentation with the most likely speaker.

The TDOA and the cross correlation measure studied for the selection of pairs of channels are of particular interest. A high cross correlation would mean a very good match between the two audio segments, which is impossible in the case of overlap. TDOAs also should vary much more in a segment with two or more speakers (if they do not move) than in a long period of only one speaker in which the TDOAs should be very stable. Therefore, the study of the influence of the overlapped speech in the TDOAs and the estimated cross correlation is of particular interest in this thesis.

6.4.1 Analysis of proposed features

There are many possible features that could be used to detect overlap. Acoustic features intend to detect the variations in the speech in regions of overlap while localization features would focus on the placing of the speech source, and, having at least two sources these would be unstable.

In this section we are going to perform only some preliminary analysis on the features used throughout this thesis, and try to assess their utility to detect regions of overlap.

The first features we want to analyse are the TDOA related ones. If two speakers are in a meeting talking each at specific moments, never interrupting each other, the delays are expected to be invariant. During the time when the first speaker is talking the delay between two particular microphones will be X , while when the second one participates the delay of the registered audio signal between these same two microphones will be Y . We expected that, if the two speakers start to speak simultaneously, the estimation of delay during that period of overlapped speech will have a lot of variations. It could happen that the finally estimated delay is neither X nor

Y, but a different one. The other possibility is that the estimated delay changes from X to Y very fast, leaving an inconstant value for a very short period of time. With this hypothesis we have decided to use the delays extracted with the XCorrPairs method (see section 5.1.1) for the detection of overlap regions.

As we have mentioned the value of delays per frame could be different if two speakers are talking, and can also happen that the delays vary very fast between several values of delays (the two delays corresponding to the two speakers talking and maybe more). Then, we have decided to calculate the local entropy of the delay values using windows of 100 frames (1 sec). This entropy will work as a measure of the uncertainty of the delays. The entropy is calculated as:

$$H = - \sum_i p_i \cdot \log(p_i) \quad (6-4)$$

Where H is the entropy, and p_i is the probability of finding the delay equal to i in the analysed window.

Entropy will be 0 if there is only one value of delay among the 100 frames evaluated, and near 0 if there are few frames with a different value assigned. On the other hand, if as many different delay values as the number of frames evaluated are present, the entropy will turn maximum.

There is a situation in which the entropy will not grow heavily, even if there is great variation of the delay values. If the number of possible delays is only two (probably from the only two talking speakers) and it changes in every frame, the entropy will be the same as if half of the time one speaker talks (first delay) and the other half the other speakers (second value). To identify this situation we have included another parameter, the frequency of these changes.

Finally, other two parameters will be analysed, the mean and the variance of the TDOAs distribution in the analysed window.

We have been working with another parameter alongside TDOAs, the cross correlation. The cross correlation have proved to be very useful in the determination of

the best pairs to use in diarization. Used as a quality requirement, it allows the system to work with only the channels whose TDOAs are reliable. Still, in the set of delay values extracted for one meeting some of them are bound to have been estimated with a very low cross correlation. It is possible that the signal coming from two microphones cannot be time aligned properly to extract the correlation because there are two or more speakers talking, and the voice signals arrive at different moments to the microphones. In this situation, even if the best alignment correspond to one of the real speaker delays, the correlation associated could be much lower than in the case of a single speaker talking. Of course this situation could happen not only when two speakers talk but when there are other noises, laughs, etc... Still, we would ideally want to exclude those kind of segments from the segmentation prior to the training of the speaker models.

Apart from the xcorr value associated to every frame we have also decided to analyse the entropy, frequency of changes, mean and variance of the xcorr (the same processing applied to the TDOAs). However, cross correlation values tend to be very similar with only slight variations in the decimal places. The entropy needs to calculate the probability of every value in the window. The variations were so high that entropy was always maximum. The same occurs with the frequency of the changes, nearly every frame has a different value of xcorr. There are still two other parameters which we expect to show some information, the mean and the variance. The smoothing performed by the mean in the distribution of xcors (in time) would allow to notice if some part of the recording has significantly lower cross correlation than other.

Summarising we propose, the following possible features: the TDOA, the xcorr, and their entropy, frequency of changes, mean and variance.

We want to add also some other delay and xcorr related features. The calculation of delays extracts not only the most probable delay value but also some other possibilities. These possibilities are less probable because their cross correlation is smaller than the one in the first place. Many times the estimated delay of a region of overlap corresponds to one of the speakers talking, the dominant speaker. The signal recorded by two microphones is then aligned thanks to this speaker. Even if the first best option does correspond to a specific single speaker and does not vary during the

overlap region, we wonder if the second would. We have then computed all the modifications applied to the first TDOA and first best xcorr to the second best TDOA and the associated xcorr.

Now we have the first and second TDOA and XCorr, so we define some relations between them. In a region of overlapped speakers talking, the intrinsic variations in the first and second TDOA and xcorr could not be as related to the presence of overlap as the relation between, for instance, the first and second TDOA. The first TDOA of a region with a single speaker that is correctly calculated would be stable and would have a high cross correlation associated. For the second TDOA option it is expected that the delay is much less probable and the cross correlation much lower. On the other hand, in a region of overlap, the system which estimates the delays would have to choose between two equally bad options, delivering a first TDOA and a second TDOA with lower xcorr associated than in the case of non-overlapped speech. This is expected to happen because it would have to align the combined signal of two or more speakers recorded with one microphone (voices arriving to this microphone with a delay between them dependent of the position of the microphone and the speakers) with the combined signal recorded with another microphone, located in a different place of the room, and whose distance to the speakers could be very different. Even if the first delay does correspond to one of the speakers, the second may not, or may not be as stable as the first option. The expected variation could be more apparent in the second best delay. We have then included in the analysis the relationship between the first and the second delay (just computing $TDOA1/TDOA2$). We have also computed the relation between the first and second entropy, frequency of changes, mean and variance.

In the same manner the relation between the two cross correlations could be quite informative. If a single speaker is talking and the delay estimated is correct, the cross correlation is expected to be high, and the second best choice would have a very low cross correlation. In the situation of overlapped speech, the first option, even if estimated correctly, would have a lower xcorr much similar to the second one. After all, if there are many voices arriving to microphones at slightly different moments (depending on the relative position of the speakers to the microphones), the system would find many suitable alignments, but with less confidence. The relation between

the two best xcorr, as well as the relation between their entropies, frequency of changes, mean and variance have been studied as well.

Every parameter mentioned up to now uses information from only one pair of microphones. We have been working with the best 4 pairs, being chosen with the XCorr-Pairs strategy, and we have computed the average for all pairs of all the parameters explained until now. The microphones are supposed to change from one meeting to another so the average of the TDOA, xcorr, entropies, frequency of changes, means and variance, as well as their relation with their corresponding second best, will be more useful to extract general conclusions (as, for instance, we can never assure that the first pair will be the same best pair for other meeting).

One consideration to have in mind is that the delay estimation program as presented in (Anguera X. , 2006) performs a final smoothing of the delays extracted to avoid outliers and any non-probable delay. As the objective is precisely to find the regions with these abnormal situations we have removed this smoothing.

Apart from the delay and xcorr related parameters we have studied in this thesis the use of some glottal features (see section 6.1) for diarization. Some of them, even if they were not good to separate speakers, could give some information about whether the speech region belongs to only one speaker or more. As we have been doing, the values of the glottal features have been pre-processed to add some modifications. We will have then the lif0, HNR, H1H2, HARMONICS, LSF, LSFSOURCE and NAQ and their entropy, frequency of changes, mean and variance. Note that the only F0 related value included in this analysis is the lif0.

The whole set of parameters and their relations are too many variables to experiment directly with the diarization system. We have decided to initially analyse the relation between the set of parameters and the presence of overlapped segments.

6.4.2 Experiments and results

With these experiments we want to be able to select, among those presented in the previous section, the most useful features to detect overlap. We have decided to

focus only in the detection of overlap in this preliminary analysis. The detection would allow the system to discard these segments in the training of the models of the speaker. It is therefore interesting to discard as many segments as possible, while keeping enough clean speech to train reliable models. In these situation, although our objective is to detect overlap, the inclusion in the overlapped detected class of segments of noisy speech or silence should not be something very worrying for the ultimate goal of improve diarization.

The database used for the experiments is the AMI database. Most works on overlap use this database because it comprises more segments of overlap segments than other databases. Also, it would make easier any comparison of the features presented with others in literature. We have decided to use the set of meetings used by (Zelenak, Segura, Luque, & Javier, 2012). As the objective is only to perform a preliminary study of the features presented in this thesis for the task of overlap, the experiments will not be as exhaustive, and we will work initially only with the singlesite database. Taking into account that we are doing only initial experiments to select possible features for overlap detection and we are not going to focus on developing a good overlap classifier, the results are not expected to be competitive, although a rough comparison could be done.

We have decided to use the program Weka (Frank, Hall, & Witten, 2016) to test the capabilities of every presented feature. As the number of examples of non-overlapped speech are much higher than those of overlapped speech, we have initially truncated the data to have the same number of examples of overlap than non-overlapped speech. Also, the references of the database had to be modified to match the expected outcome: singleSpeaker vs multiSpeaker.

We have extracted the Pearson correlation of every feature with the class distribution. Most of the features explained had a very low correlation with the class, the best ones are listed in Table 22. There are other tools in Weka that would allow the selection of a group of features instead of just ranking them. Some initial experiments were carried out with some of this approaches but results were not very different so,

for this exploratory analysis, we decided to continue with the ranked features in Table 22. A more in depth analysis could

Though correlation is not very high, if some information can be utilized to detect overlapped segments (or even unreliable segments) the diarization system could improve. The rest of the parameters had lower correlations and we have decided to keep only those in Table 22 for the following stages. Also, although we evaluated not only the average values but their individual microphone pair values, we have discarded them if only two pairs had correlation with the class distribution higher than any of the selected ones. This has actually happened only for one feature from one pair so all the selected values are the average of the same values per pair of microphones.

#	Feature	explanation	Pearson coeff
1	XCORR2_ave_vars	Average variance of the second best xcorr	0.4533714
2	XCORR2_ave_means	Average mean of the second best xcorr	0.4413671
3	DEL_ave_entropies	Average entropies of first best delay	0.3474287
4	DEL_ave	Average first best delay	0.3277311
5	RATIODEL1DEL2_ave_entropies	Ratio of the average entropies of delays: $\frac{\text{average first_best_delay entropy}}{\text{average second_best_delay entropy}}$	0.3205553

Table 22: Pearson correlation coefficient of the features used to detect overlap segments

If we visualize the data used for the training using the best feature (XCORR2_ave_vars) always in the x axis we can see that the multispeaker examples cannot be clearly separated. Still, as shown in Figure 42 some overlapped examples has, as expected, much higher variance of the xcorr (XCORR2_ave_vars).

It is important to notice that the separation is clearer due to the preprocessing of the data made to be able to analyse results. In real data the examples of non-overlap are many more than those of overlap and therefore, the final amount of examples of no overlap whose values of xcorr and delay go to the regions of overlap is also higher. Though still a high percentage of frames would remain in the area of single speakers, the area of overlapped speech would be highly corrupted with speech from single speakers.

Figure 42: Distribution of the training multispeaker (red dots) vs single speaker (blue dots) for the training AMI set processed to have equal number of overlapped segments than non-overlap segments. X axis in all figures is $XCORR2_ave_vars$. Y axis is the following: Figure 42.a: $XCORR2_ave_means$; Figure 42.b: $DEL_ave_entropies$; Figure 42.c: DEL_ave ; Figure 42.d : $RATIODEL1DEL2_ave_entropies$.

We have performed with Weka some classification experiments, using the AMI training set with a 10-fold validation. The objective is not the classification itself but the testing of the preliminary possibilities of the selected parameters in the detection of overlapped segments. Therefore, the classifiers used were just some of the default classifiers in the Weka program. We want to check if we could reduce even more the list of parameters, so we performed the classification using the best 5 features and then the best 4 and 3.

We do not seek to improve the classification of overlap vs non-overlap, but to be able of discarding as much overlap as possible while keeping enough clean data to model reliable speakers. This would mean having enough data from a single speaker and not too corrupted with other kind of data (i.e. maintain high recall and precision) and, at the same time, having as much overlapped speaker data as possible recognized like that (i.e. having a high recall of the overlapped speech cluster). As the objective is to exclude this data from the training of the speaker models, it is not really important if the overlapped segments include many other segments of non-overlapped speech. To compare the classifiers we have defined a merit figure that weights this parameters as in equation (6-5).

$$\text{Merit_figure} = 0.5 * \text{Fscore}_{\text{single speaker}} + 0.5 * \text{Recall}_{\text{overlapped speech}} \quad (6-5)$$

$$\text{Fscore} = \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6-6)$$

As can be seen in Table 23, the merit figure calculated for every classifier is very similar for many of them. Note that values are now expressed as a fraction of unity (instead of the usual percentage). The best segmentation is obtained with the SMO classifier when using the 5 parameters selected. The merit figure for this classifier is 0.79, the next best is the multilayer perceptron and the bayes net. In general the classification is slightly better with 5 features than with 3. Only the Bayes net seem to perform slightly better with only 4 features. Still, these results are obtained with the pre-processed database, which contain equal number of samples of singlespeaker than overlap.

Classifier	# Features used	Precision	Recall	Merit figure	class
ZeroR	-	0.5	1.0	0.333	singleSpeakers
		0.0	0.0		OverlappedSpeech
OneR	5	0.692	0.740	0.693	singleSpeakers
		0.720	0.671		OverlappedSpeech
	4	0.692	0.740	0.693	singleSpeakers
		0.720	0.671		OverlappedSpeech
	3	0.692	0.740	0.693	singleSpeakers
		0.720	0.671		OverlappedSpeech
Naïve bayes	5	0.701	0.804	0.703	singleSpeakers
		0.770	0.657		OverlappedSpeech
	4	0.698	0.838	0.699	singleSpeakers
		0.797	0.637		OverlappedSpeech
	3	0.689	0.864	0.687	singleSpeakers
		0.817	0.609		OverlappedSpeech
BayesNet	5	0.725	0.727	0.725	singleSpeakers
		0.726	0.724		OverlappedSpeech
	4	0.733	0.760	0.734	singleSpeakers
		0.750	0.722		OverlappedSpeech
	3	0.729	0.782	0.732	singleSpeakers
		0.765	0.709		OverlappedSpeech
Multilayer perceptron	5	0.729	0.773	0.731	singleSpeakers
		0.758	0.713		OverlappedSpeech
	4	0.726	0.776	0.728	singleSpeakers
		0.759	0.707		OverlappedSpeech
	3	0.725	0.777	0.727	singleSpeakers
		0.759	0.705		OverlappedSpeech
SMO	5	0.722	0.818	0.792	singleSpeakers
		0.790	0.685		OverlappedSpeech
	4	0.722	0.799	0.726	singleSpeakers
		0.775	0.693		OverlappedSpeech
	3	0.717	0.801	0.720	singleSpeakers
		0.775	0.684		OverlappedSpeech

Table 23: Precision and recall using a 10-fold validation for the AMI training set processed to have similar number of single speaker frames than multiple speaker. The features used are the best # Features ranked in Table 22

When we try to classify overlap vs no overlap in real meetings the results are far worse for the precision error. On the other hand, even if many single speaker segments are labelled as overlap in the detection stage, as long as there are still enough frames classified as single speaker the performance of the system could improve (because, as it

was mentioned before, the idea is to discard the segments of overlap in the training phase but latter assign them to one speaker).

When trying to classify frames from a whole real meeting the results for the classifiers presented previously turn to be much less precise as the majority of the frames classified as multiple speaker are actually single speaker frames (see Table 24).

Classifier	# Features used	Precision	Recall	Merit figure	class
ZeroR	-	0.951	1.000	0.487	singleSpeakers
		0.075	0.631		OverlappedSpeech
OneR	5	0.954	0.996	0.515	singleSpeakers
		0.431	0.056		OverlappedSpeech
	4	0.954	0.996	0.515	singleSpeakers
		0.431	0.056		OverlappedSpeech
Naïve Bayes	5	0.966	0.945	0.655	singleSpeakers
		0.246	0.356		OverlappedSpeech
	4	0.975	0.954	0.642	singleSpeakers
		0.266	0.324		OverlappedSpeech
BayesNet	5	0.967	0.947	0.657	singleSpeakers
		0.257	0.358		OverlappedSpeech
	4	0.965	0.959	0.636	singleSpeakers
		0.277	0.311		OverlappedSpeech
Multilayer perceptron	5	0.954	0.997	0.516	singleSpeakers
		0.523	0.058		OverlappedSpeech
	4	0.954	0.997	0.512	singleSpeakers
		0.492	0.05		OverlappedSpeech
SMO	5	0.951	1.000	0.487	singleSpeakers
		0.075	0.631		OverlappedSpeech
	4	0.951	1.000	0.487	singleSpeakers
		0.075	0.631		OverlappedSpeech

Table 24: Precision and Recall using a 10-fold validation for the AMI training set when classifying non-overlapped speech and overlapped speech regions. Meetings have many more samples of non-overlapped speech than overlapped. The features used are the best # Features used best ranked in Table 22.

The classifiers used for these preliminary experiments could be greatly improved to suit the task. Looking at classification with real data it is clear that the great difference in the number of samples of single speakers vs overlapped speakers is affecting the classification. Still, now that the analysis has been carried out with many more data the only system which worked better with 4 features than 5 (bayes net) requires now the 5 features for the best segmentation.

We wanted to check the performance of this best classifier if we use more features (at the end the best system with real data uses the 5 best features preselected so could make use of some more).

We carried out experiments with a bayes net classifier using from the best 5 features ranked in Table 22 to the best 20 features, adding 1 new feature each iteration, the result in merit figure is included in Figure 43, where it is clear that the merit figure nearly stops growing from 11 features. This experiments used as training the AMI training set and as test set the AMI development set. We can conclude that, the rest of the features have almost no new information. Note that an increase in merit figure is ruled by an increase in recall of overlapped segments but, of course the precision in overlap detection is reduced. For the system with 11 features the precision has been reduced to 0.135 and the recall has grown to 0.51. The 6 attributes added to the 5 in Table 22 are listed in Table 25.

Figure 43: Merit figure of the development AMI dataset across the number of attributes used to train the classifier.

#	feature	explanation
6	RATIOXCORR1XCORR2_ave_means	Ratio of the average means of the best xcorr and the second best
7	RATIODEL1DEL2_ave_freqChanges	Ratio of the average frequency of changes of the best delay and the second best
8	RATIODEL1DEL2_ave_means	Ratio of the average means of the best delay and the second best
9	DEL_ave_freqChanges	Average frequency of changes of the best delay
10	RATIOXCORR1XCORR2_ave_vars	Ratio of the average variance of the best xcorr and the second best
11	RATIODEL1DEL2_ave_vars	Ratio of the average variances of the best delay and the second best

Table 25: features ranked with Pearson coefficient from 6th to 11th position when analysing its correlation with overlap segments.

Remember that we are not trying, at this point, to improve our DER performance using these classifiers, but to analyse if some of this features could have some information about the presence of overlap. Most systems in the literature use much more complex classifiers, adjusted for the task, and also combine different kinds of features (like using MFCCs). A real classifier adapted to the detection of overlap and combining other kind of features could make use of these 11 features or maybe less, but hardly more as the rest of the evaluated features do not add new information.

The studied features can represent some characteristic of the overlap data, but to detect overlap and only overlap are not useful (at least if used alone). Recall is too low and precision even more so. The question arises, however, of if the frames incorrectly classified as overlap would be actually segments with no so clean speech. There are some works using TDOA related features to detect and label overlap (Zelenak, Segura, Luque, & Javier, 2012) with successful results. The modification of the classifier to adapt it to the task (taking into account the sequence of states for instance) as well as the use of these features with acoustic features could improve the performance.

7 *Segmentation algorithm*

The final part of this work was inspired during the study of the baseline diarization system used. The “segmentation and agglomerative clustering of speech regions” in Figure 5, initially performs a cluster initialization followed by an iterative segmentation and training and cluster comparison (see Figure 6). We are now focusing in this part of the process.

Remembering the process in Figure 6, the segmentation begins with an initialization process which segments the speech into K blocks (equivalent to an initial hypothesis of K speakers or clusters) uniformly distributed. As mentioned before, K has been set to 16 empirically. Every cluster is modelled using a Gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for [mfcc] and 1 for [tdoa] streams). After the initial segmentation a set of training and re-segmenting steps is carried out using Viterbi decoding. Then the merging step takes place.

Then, when a merging takes place, the GMM for the new cluster is retrained with the data now assigned to it and the number of parameters (mixtures) of the merged model is the sum of the number of mixtures of the component models. The segmentation and clustering steps are repeated until a stopping criterion is reached. To decide which clusters to merge, and when to stop the merging, the ΔBIC criterion has been used. When all possible merge pairs give a negative ΔBIC , the merging is stopped. A frame purification algorithm is also applied before computing the BIC distance to eliminate the short silence segments wrongly labelled as speech that are corrupting the models. In (Anguera X. , 2006) variants of a likelihood-based metric are proposed to determine if a frame corresponds to a silence segment and then, filter it out..

The segmentation stage decides, using the speaker models, which segments of the meeting belong to which speaker. For some systems as our baseline, the segmentation is influenced by a weighting parameter dependent on the number of speakers, which initially, as we said before, can only be hypothesised. The goal of this part is to study the effect on the segmentation stage of this parameter and improve the diarization by making segmentation independent of the number of active speakers.

7.1 Statement of the problem

The diarization system starts the segmentation and clustering stage with a division of the recording into several parts (cluster initialization). Each part is assigned to a different cluster (speaker) and used to train the corresponding GMM. The next stage, named “Segmentation and Training” in Figure 6, uses these GMMs to decide which frames belong to which cluster. The system uses an ergodic Hidden Markov Model (HMM) where each state corresponds to each cluster and then performs a Viterbi search to find out the new best sequence of turns according to the likelihood of each frame to belong to each speaker.

Each state is composed of a number of sub-states which imposes the minimum duration mentioned already. As can be seen in Figure 44, the system has to go through all the sub-states before being allowed to change to the first sub-state of other state (cluster). Each one of the sub-states has a probability density function modelled via a Gaussian mixture model (GMM). The same GMM model is tied to all sub-states in any given state. Probabilities of changing or remaining in the same state at the last sub-state have been set to 1 following the recommendation in (Anguera X. , 2006) and (Anguera, Wooters, & Hernando, 2006) who proposed to set $\alpha=\beta=1$ to avoid unrealistic fast changes of speaker, leaving the process as represented in Figure 45. Now the probability of remaining in the same cluster or changing to a different one is mainly ruled by acoustic data, with only one additional transition factor (T_r) dependent on the number of hypothetical active speakers at every iteration ((in Figure 45, $T_r=1/M$, with M the

number of active clusters). This T_r could be easily related with the interword penalty of a speech recognition system, but T_r is variable, as M is always changing, and it was never adjusted in previous works. This T_r parameter results in an effective penalization factor (though, admittedly, not as harmful as having $\alpha=0.9$) applied to the probability of speaker changes (see Figure 45).

Figure 44: Cluster models with minimum duration (from (Anguera, Wooters, & Hernando, 2006))

Figure 45: Diagram of the dependency of turn speaker changes with the number of active speakers.

The number of active speakers begins at 16 and decreases, one by one, every time that the system decides to merge two clusters. The associated transition weight will increase accordingly, as if the probability of having speaker turns would be higher if less speakers are involved in the meeting which is not related to the social activity of the meeting or the number or role of the actors intervening in it. This factor is not usually tuned in diarization systems and this section will focus on the study of this characteristic and the proposal of a better alternative. The factor $1/M$, being M the number of hypothesised speakers at the moment, constitutes therefore a prior variable penalization to changes of speaker.

During the segmentation, in the Viterbi search implemented to decide the best sequence of speaker turns, the value stored in each final sub-state is the sum of the previous “Minimum Duration” log-likelihoods for each cluster and frame. The Minimum Duration (Md), as it was previously mentioned, is the minimum number of frames that has to be assigned to one cluster after a change of speaker. The system calculates this value for each frame and keeps the cluster with the highest log-likelihood at its final sub-state.

A change of speaker will occur when the sum of the last Md log-likelihoods from the current cluster is lower than the sum of the last Md log-likelihoods from any other cluster plus the logarithm of the transition weight. Therefore, a transition between speakers will take place if the following condition is met:

$$\sum_i^{i+MIN_DUR} \log\mathcal{L}(cl_j; fr_i) < \log(T_r) + \sum_i^{i+MIN_DUR} \log\mathcal{L}(cl_u; fr_i) \quad (7-1)$$

Being $\log\mathcal{L}()$ the log-likelihood, cl_j the current cluster, cl_u the candidate cluster, T_r the transition weight (in the baseline system this is $T_r=1/M$) and fr_i the frame being evaluated.

In an extreme situation a very high or very low value of the transition weight could surpass, for every possible data, the difference of the log-likelihoods of the current cluster and the candidate cluster. In practice this would force, or forbid, the transition

to other speakers, independently of the information given by the acoustic data and the speaker models. An extremely low transition weight (whose logarithm would be negative) would make impossible the change of speaker, resulting in a final solution with only one speaker. The opposite situation would result in a solution where any speaker turn is no longer than the minimum duration established at the beginning because changes are maximized.

In the current baseline system, the lowest possible value, and thus the highest opposition to changes, takes place at the beginning of the algorithm, when the number of speakers is 16. From that moment, the transition weight will be increased in every iteration, which would mean that there is higher probability of changes when fewer speakers are present in a meeting. Note that it is impossible in the baseline system to have an iteration in which changes are governed only by the acoustic information, because that situation would happen only when the number of hypothesized speakers is equal to 1, meaning that there will be no speaker turns at all.

Although speaker turns could have some dependency on the number of participants, there is no prior information that would make us to think so. This dependency, if it exists, could be related to the number of participants but more likely to the role of the speakers or the context or content of the meeting (a lecture would have probably much less speaker changes than an informal gathering).

In this section we will carry out experiments to eliminate this variability of the transition weight, making it therefore independent of the number of active speakers at any moment. We want also to study the possibility of using this transition term to actually favour the transition between speakers. As mentioned before a very high value of the transition weight would increase the number of speaker turns drastically, and a too low value would make them nearly impossible. A study of this term is necessary in this case to assure that neither of these situations are encountered.

A value of the transition weight equal to 1 is a special situation where there is no influence of this term, thus it would not either penalise or favour transitions. Any transition would be determined only by the likelihood of the cluster models given the data.

The transition probability is expected to modify the speaker turns length, which is also greatly affected by the minimum duration. This term was established as 250 frames with the baseline system, which makes use of the variable transition weight ($T_r=1/M$). Due to this fact, we have decided also to carry out the experiments considering both the transition weight and the minimum duration, because both terms have influence in the duration of the speaker turns (Martínez-González B. , Pardo, San-Segundo, & Montero, 2016).

7.2 Experiments, results and discussion

The transition weight in the baseline system is $1/M$, being M the number of active speakers. M is reduced in every iteration, going from 16 at the beginning of the process to, at most, 1, which would mean that the system has found only one speaker. As a result this factor will be different in every iteration, while the clusters move from the first 16 to the final hypothesized number.

As we said before, with any number of speakers higher than 1 the transition weight will turn lower than 1, resulting in a penalization of the speaker changes. As there is also a forced minimum duration of the speaker turns, penalising further the changes between speakers makes no sense.

The special case when the transition weight is equal to 1, and therefore, neither penalise nor favour the change of speaker, happens when $M=1$. This situation is theoretically possible for the baseline system, but with only one active speaker, changes are impossible and the diarization process would end.

The experiments focus in two concepts: make the transition weight constant throughout the diarization process (independently of the number of clusters at any moment), and also check the use of values higher than 1 for the transition weight. The first would make the changes independent of the number of clusters at any stage of the

segmentation/clustering process. The second would favour the speaker changes instead of penalising them, as it happens in the baseline system.

Experiments were carried out, as with the TDOA experiments, using the RT database and MFCC and TDOA features. The initial baseline system will now use the TDOA features calculated with the XCorr-Pairs. To measure the error we have used the Speaker Error Rate (SER from now on). This value removes from the typical Diarization Error Rate (DER) the error due to the VAD module which is constant in our experiments. For future reference Missed Speaker plus False Alarm Speaker error is 7.43 for the development set (RT-DEVELSET) and 8.70 for the test set (RT09).

In Figure 46 the results are presented for the development set for values of transition weight (T_r in equation (7-1)) ranging from 0.01 to 5, different values of minimum duration (150, 200, 250 and 300 frames) and the baseline SER in a dotted line. We can see that the speaker error rate is very dependent on the transition weight and the minimum duration. The baseline system works only in the region where the transition weight is lower than one, which penalises speaker changes. The results show that favouring these changes instead of penalising them has better results. In the case of minimum duration equal to 200 frames the majority of the values of the transition weight reduces the error of the system, and furthermore, the variability across transition weights is also reduced.

Four points have been chosen to evaluate its performance with the test set, transition weight equal to 1, 2, 3 and 4, all with minimum duration of 200 frames. The performance of the baseline system for all the sets used (development and test sets) is shown in Table 26 second row. The confidence bands included were estimated with equation (5-8). For the baseline the minimum duration parameter was set empirically to 250 sub-states, which means a minimum duration of 2.5 secs. Experiments have been done using MFCC and TDOA values, whose probabilities have been combined using a weight of 0.85 for the MFCCs and 0.15 for TDOAs.

Figure 46: SER of the RT-DEVELSET for minimum duration of speech of 150, 200, 250 and 300 frames against transition weight

Transition weight	Minimum duration	RT-DEVELSET	Relative improvement over RT-DEVELSET	RT09	Relative improvement over RT09
1/M (Baseline)	250	4.11±0.03		7.82±0.07	
1/M	200	4.07±0.03	1.94%	7.73±0.07	1.15%
1.0	200	3.57±0.03	13.14%	8.45±0.07	-8.05%
2.0	200	3.29±0.03	19.95%	7.72±0.07	1.28%
3.0	200	3.22±0.03	21.65%	7.46±0.07	4.6%
4.0	200	3.06±0.03	25.55%	7.57±0.07	3.20%

Table 26: SER for every point evaluated varying the transition weight of the segmentation stage for the RT-DEVELSET and the RT09 set. M is the number of active clusters at each iteration. Weight applied to MFCC and TDOA features is 0.85 and 0.15 respectively.

To prove that the improvements obtained are not only due to the reduction of the minimum duration parameter (250 frames in the baseline and 200 frames for the proposed systems), we have checked the performance of the system when the minimum duration is the only parameter modified and the transition weight is $1/M$. Therefore, this experiment and the baseline differs only in the minimum duration established by the user. Its speaker error, calculated only for comparison purposes, is included in third row of Table 26. The next rows show performance of the diarization algorithm when transition weight is kept constant and higher than one, and minimum duration is equal to 200.

As it can be seen in Table 26, cancelling the transition weight (transition weight=1) reduces the SER for the development set but increases it for the test set. On the other hand favouring transitions between speakers (transition terms 2.0, 3.0, and 4.0) improves the performance of the system for both the development and the test set. The difference is also statistically significant for every value of the development set and for the test set when transition weight term is 3.0 and 4.0.

Though two parameters have been changed from the baseline to the system with transition weight equal to 3.0 or 4.0, the results show that the modification of the minimum duration is not the main responsible of the improvement. When a fixed transition weight is used the improvements are higher and significant.

Our experiments demonstrate also that the transition weight should be modified together with the minimum duration to obtain the best results.

One characteristic that is found in speaker diarization of meetings is that there is a high variability of results across different sessions in different rooms and disperse microphone locations and unknown number of speakers so it is very difficult to demonstrate advancements of new methods (Mirghafori & Wooters, Nuts and flakes: A study of data characteristics in speaker diarization, 2006). Although the improvement of results on the test set are smaller than the ones of the development set, if we consider both sets as an ensemble there is a definite improvement using this new approach. If we take into account that the data that we use is a community standard and that we experimented with an extensive amount of meetings (35 meetings), we can conclude

that the new method that we propose has better performance than the previous one, which is extensively used by different laboratories at this moment.

8 *Conclusions and future work*

8.1 Conclusions

In the first part of this thesis we analysed the effect in diarization of the TDOA features and propose several ways of selecting, among the available pairs of microphones, the best possible pairs. Delay features have proven to be very useful in general for the diarization task. Sometimes, however, the use of the delay computed between two particular microphones can significantly degrade the performance of our system. If the audio recorded by one microphone is not easily aligned with the audio recorded by any other microphone, the delay calculated will not be the true delay, and due to its variation throughout the whole meeting, it can degrade the general performance.

We have shown that much better performance can be obtained by using only a reduced set of TDOA values when they have been selected properly. We proposed several new ways of selection, which follow different strategies, most of them achieving better results than the original system which performed no selection of pairs.

The methods based on the quality of the cross-correlation of the signals coming from different microphones have proven to be the most useful. The results for the test set improve in a range from 11.26% to 55.68% of the relative SER and show competitive performance compared to previously published methods. Furthermore, the selection methods significantly reduce the computational time, as it limits very much the number

of parameters to be calculated. The computational cost is reduced by around 20% with the method that achieves the best performance for the test set.

Actually two methods achieve the same values in error reduction and computational time savings. The XCorr-Pairs and XCorr-Channels methods achieved error rates with no statistically significant differences. However, we already introduced that the selection of channels was not as interesting because the system could never increase the number of channels selected if we want to maintain the time savings (choosing only one extra channel the number of pairs would double). Also, even if the meeting has a very good pair it would never be selected if their channels are not the selected ones. These constraints make the XCorr-Pairs more attractive for future experiments.

The XCorrThreshold strategy is a clear example of the utility of the cross correlation as a quality measure for the estimated delays, as any combination of selection algorithms with this minimum quality measure delivers better results than their counterparts without the XCorrThreshold.

Among the other methods developed we would single out PCA and Kmeans+Silhouette technique, which achieve also good results for the development and the test set utilized.

The Kmeans+Silhouette chooses pairs according to the accuracy of the performed clustering (silhouette coefficient would be nearer to 1 if the distance within a cluster is minimized while maximizing the distance inter clusters), therefore, the final chosen pairs are less prone to lead to diarization errors in the final speaker clustering. Its performance is not as good as the PCA, but the improvement over the baseline is still noticeable and very good for the development dataset.

The PCA method achieves better performance than all the rest for the development set but the test set, though low in value, is not as good as the already mentioned XCorr related methods. The reason could be that the PCA works with every information available (or the information available once removed the really bad pairs with the XCorr threshold), and bad pairs could degrade heavily the system.

However, there is a main disadvantage of these systems (PCA and Kmeans), which is that not only they have worse DER for the test set, but that these results are obtained with a much slower system. This effect will be present for any other method that needs to calculate delays for every possible combination of pairs of microphones.

The histogram based selection methods (DM, ER and MC) were implemented with a bin width of the histograms of 5ms, which could be a very high value that would join in the same bin delays that belong to more than one speaker. The final TDOA values calculated will have a much higher resolution but the selection may be affected for this so high bin width. The histograms are only used to check which pairs have the highest dynamic margin (in DM), the highest number of high delay values (in the ER), or the highest number of the common values (in the MC method), and later the original delay values are used for diarization. As all the histograms from all the pairs use the same bin width the comparison between pairs could still be carried out, but could happen that the difference between delay values was lower than those 5ms and then, even if the delays are easily separated, these histogram based selection techniques would consider it a same value and would not select that pair. This could be one reason why none of these techniques achieved very good results.

It was clear when analysing results that every meeting needed a different number of final pairs selected. In this work, we have developed some methods to reduce the error and the computational time, but it requires some preliminary analysis to decide the number of pairs to select which is not optimum for each meeting. On average, the best number of pairs can be 4, but the best number can vary according to the meeting. Still, most of the algorithms rely, in average, in a lower number of pairs which is indicative that the selected pairs are of higher quality than those used in previous system.

In this work we have also analysed the task of unsupervised diarization focused on obtaining pure speaker recordings in order to synthesize voices from media data with only one audio channel available. With this purpose we have modified slightly the traditional task of diarization. Now we have focused on recovering pure speaker clusters, even if we have to discard many segments, or speakers, overlapped with other speakers

or noises. For such objective we have defined a merit figure that weights the precision and the recall. We have studied the application of the UPM music detection module. We have proved that by using the music recognition module we can decrease the precision error 22.36% for the development set and 39.64 % for the test set, improving also the merit figure.

The noise reduction module in combination with the music detection module makes the system to lose too many segments of speech, reducing the recall, and thus the merit figure, and making this combination undesirable.

Results using lif0 in combination with music detection were slightly better for the development set and slightly worse for the test set, therefore, we cannot prove its usefulness for this task. However, when combined with the glottal features the best system does not need anymore the lif0. The HNR or the HARMONICS+lif0 can be used in absence of TDOA to enhance the result of the diarization. The lif0 tested have also very good results when the weight applied to the stream is 0.1 but varying the weight only to 0.05 the error increases drastically which makes us think that it may not be the best option. HARMONICS were not tested alone because the best point for the development set was found with the features concatenated with lif0, however, results with the development set were as good as those with the HNR, which makes us think that they could also be used.

Most of the glottal features evaluated were not clearly useful in diarization, but they were only a reduced set of features to begin the exploration. Even though there were not a very high amount of features, all of them improved the baseline for some working points and even some of them proved its utility with the whole system.

The glottal features along with other TDOA and XCorr related features were studied for its use in the detection of overlap segments. The whole set of features was ranked in accordance to correlation with the two classes to be evaluated and then some simple classifiers were used to test the usefulness of the group of features selected. We have worked with the first 5 features, the average variance of the second best xcorr of the best 4 pairs, the average mean of the second best xcorr of the best 4 pairs, the average entropies of the delays from the best 4 pairs, the average delays from the best

4 pairs and the ratio between the average entropies of the best delays and the second best. Although our experiments use mostly these 5 attributes, we have found that we could use up to 11 features. From that point, the system has no gain, so using more features among those designed for these experiments is not necessary,

Experiments with equal samples of overlap and single speaker showed that the selected features, though not enough to carry out the task perfectly, had some relation with the presence of overlapped speech. However, these features seem not enough to detect correctly the overlap regions in a whole meeting with much more single speech data than overlap data. Although the detection of overlap was quite imprecise, the question remains as if the recall of the single speaker would be enough to generate reliable clusters. If so, the discarded regions of single speaker (wrongly labelled as overlap) could be simply be assigned afterwards to the most suitable speaker,

It is important in this case to differentiate between detecting overlap and recognizing the speakers involved. One feature could be of much use for the first task and not so much for the second. Also, as the diarization is greatly affected for the wrongly trained speakers models, the fact that the overlap detection module is not actually detecting only overlap but regions of noisy speech could still be of use. Focusing on maximising recall the detected segments would include as much overlap as possible, and, if the number of frames of speech not discarded are still big enough to train reliable models, the final diarization will be greatly benefited. So the system could simply discard the suspect segments in the training phase of the speaker models (always ensuring that the quantity of clean speech remains big enough for the training of reliable models). These discarded segments would be assigned later to the closest speaker, or to the closest two if they are finally found to be overlapped speech.

Finally, in this thesis we have found and solved one problem of the diarization system that we were not aware of. The segmentation stage make use of one parameter related to the number of active speakers, which at the end means that changes of speakers are not ruled solely by acoustics and that the penalization applied to changes varies as if it were more probable to have a change of speaker if there are few speakers in the room. We have proved that the segmentation stage of a speaker diarization

algorithm can be improved by not penalizing transitions between different speakers, or by making them more probable. The variability of results is reduced when these transitions are not penalised and remain constant throughout the segmentation/clustering iterations.

The work presented in this thesis has produced the following publications:

- **B. Martínez-González, J. M. Pardo, J. D. Echeverry-Correa, R. San-Segundo,** "[*Spatial Features Selection for Unsupervised Speaker Segmentation and Clustering*](#)". Expert Systems With Applications No. 73, 1 May 2017, pp 27-42
- **B. Martínez-González, J.M. Pardo, J.D. Echeverry-Correa, J.M. Montero,** "[*New experiments on speaker diarization for unsupervised speaking style voice building for speech synthesis*](#)". Procesamiento del Lenguaje Natural, [S.l.], v. 52, p. 77-84, mar. 2014
- **B. Martínez-González, J.M. Pardo, R. San-Segundo, J.M. Montero,** "[*Influence of Transition Cost in the Segmentation Stage of Speaker Diarization*](#)". Odyssey 2016, June 21-24, pp 385-392, 2016, Bilbao, Spain.
- **B. Martínez-González, J.M. Pardo, J.D.Echeverry-Correa, J.A. Vallejo-Pinto, R. Barra-Chicote,** "[*Selection of TDOA Parameters for MDM Speaker Diarization*](#)", InterSpeech 2012, 13th Annual Conference of the International Speech Communication Association. Portland, Oregon. September 9-13, 2012

8.2 Future work

The TDOA selection methods have been proven to be of much utility in reducing the diarization error. However, in all the experiments we have decided to select a fixed number of final pairs, which will not be optimum for every meeting. A future work would be to modify the algorithm introducing a condition to decide, for every meeting, which

could be the best number of pairs to use. This parameter could be related, for instance, to the average xcorr or the silhouette coefficient of the different pairs. Instead of choosing a fix number of pairs, choosing only those that fulfil some characteristic, as it is roughly done in the XCorrThreshold, could greatly improve the adaptation of the system to new data. The difference with XCorrThreshold would be that XCorrThreshold is always applied in our experiments in combination with other selection methods which set the final maximum number of pairs. An adaptive system is always more desirable as the meetings with high number of usable pairs would not discard them while those with only bad pairs could even work with no TDOA and improve the overall result.

For instance the SER of the XCorr-Pairs system would decrease from 4.11 to 2.72 (a 33% relative improvement) if the system were able to select its best number of pairs for each meeting. It would be then interesting to find a relationship between the best number of pairs and any parameter of the data, so the final number of pairs selected would not be fixed.

Also, all the selection methods developed select pairs of channels without considering the relation between their delay values. Another future line of research would be to modify the selection strategy to select the pairs one by one, always trying to choose new pairs whose delay values have a different distribution of the delays when compared to the already chosen pairs (for instance minimising correlation between them). With this approach two pairs with identical information would not be selected, even if both have very high quality.

About the DM, ER and MC selection systems, we mentioned in the previous section that the bin width used in the histograms (5ms) could be too high for the actual delay values and may be preventing the selection of the really useful pairs. In future work some experiments could be carried out reducing this bin width to check if the performance improves.

Continuing with the selection of pairs, a possible future research alternative to PCA, would be the use of a Deep Neural Network to extract bottleneck features from delay features of preselected pairs of microphones and combine them with MFCC features or with bottleneck features extracted from MFCC features. We consider it an

alternative to PCA as the objective would be to use every pair available and then reduce dimensionality.

We have also studied several glottal features and applied them to other tasks, with different goals in mind but which could make use of a diarization preprocessing. In future work we could study more features traditionally used in diarization or in other areas of speech technologies to tackle the problem of the automatic retrieving of clean speech voices. Preferably features that could be used if there is no more than one channel which would allow the use of many real life audio recordings.

It could happen that some features would be more useful to diarize some specific kind of meetings. The classification of the features according to their utility in the diarization of meeting styles would also be a possible future direction.

The music detection algorithm in combination with the glottal features have been tested for the C-ORAL-ROM database, which contain music and was recorded with only one microphone. A future task will be to test the utility of this music detection system and the glottal features with the RT database.

We have proved also that the overlap detection is a very demanding task in speaker diarization. Exploring the features susceptible of being used to detect overlap is still a very important area of research, as well as the development of new algorithms to classify the speech segments into overlap or non-overlap. We can talk about xcorr related features, f_0 and TDOA, but also the silhouette coefficient or any feature that could be affected for the presence of more than one speaker. The use of the standard deviation, instead of the variance, could also be beneficial.

Another future line of research will be the use of the features selected, joined with MFCCs and maybe other features, to detect overlap, followed by its application in our diarization system (discarding the recognized frames of overlap in the training process).

Finally, the diarization system makes use of many different modules as the music detection module applied in this thesis. A future task must be the inclusion of a better

voice activity detection module as most of the error in the RT database for the development set is now due to the speech/non-speech (which includes the overlap errors).

References

- Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., . . . Sivasdas, S. (2002). Qualcomm-ICSI-OGI features for ASR. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*. Denver, Colorado.
- Ajmera, J., & Wooters, C. (2003). A robust speaker clustering algorithm. *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, (pp. 411-416).
- Ajmera, J., McCowan, I. A., & Boulard, H. (2004). Robust Speaker Change Detection. *IEEE Signal Processing Letters*, 11, 649-651.
- Alku, P., Bäckström, T., & Vilkmán, E. (2002). Normalized amplitude quotient for parameterization of the glottal flow”,. *Journal of the Acoustical Society of America* , 112(2), 701-710.
- Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. 50(1-2), 5-43.
- Anguera, X. (2006). *Robust Speaker Diarization for Meetings*. Barcelona: Universidad Politecnica de Cataluña, UPC.
- Anguera, X., Aguiló, M., Wooters, C., Nadeu, C., & Hernando, J. (2006). Hybrid speech/non-speech detector applied to speaker diarization of meetings,. *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006*.
- Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2), 356-370.
- Anguera, X., Wooters, C., & Hernando, J. (2006). Automatic cluster complexity and quantity selection: Towards robust speaker diarization. *MLMI 2006, Lecture Notes on Computer Science 4299* (p. 248=256). Springer Verlag.

- Anguera, X., Wooters, C., & Hernando, J. (2007, Sept.). Acoustic beamforming for speaker diarization of meetings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7), 2011–2022.
- Anguera, X., Wooters, C., Pardo, J. M., & Hernando, J. (2007, April). Automatic Weighting for the Combination of TDOA and Acoustic Features in Speaker Diarization for Meetings}. 2007 *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4, IV-241-IV-244.
- Anguera, X., Wooters, C., Pesking, B., & Aguiló, M. (2006). Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. *Proceedings of the Second International conference on Machine Learning for Multimodal Interaction*, 402-414.
- Barra-Chicote, R., Pardo, J. M., Ferreiros, J., & Montero, J. M. (2011). Speaker Diarization Based On Intensity Channel Contribution. *IEEE Transactions on Audio, Speech and Language*, 19(4), 754-761.
- Barras, C., Zhu, X., Meignier, S., & Gauvain, J. L. (2004). Improving speaker diarization. *Proc. DARPA RT04*. Palisades, NY.
- Bartsch, M. A., & Wakefield, G. H. (2001). To catch a chorus: using chroma-based representations for audio thumbnailing. *Bartsch, M. A. and G. H. Wakefield. 2001. To catch a chorus: using chroma-based represenIEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 15-18.
- Ben-Harush, O., Ben-Harush, O., Lapidot, I., & Guterman, H. (2012). Initialization of Iterative-Based Speaker Diarization Systems for Telephone Conversations. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 414-425.
- Ben-Harush, O., Lapidot, I., & Guterman, H. (2010). Online diarization of telephone conversations. *Odyssey*.
- Boakye, K., & Vinyals, O. G. (2011). Improved Overlapped Speech Handling for Speaker Diarization. *Interspeech 2011*, (pp. 941-944). Florence, Italy.
- Boakye, K., Trueba-Hornero, B., Vinyals, O., & Fiedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, (pp. 4353-4356). Las Vegas, NV.

- Boehm, C., & Pernkopf, F. (2009). Effective metric-based speaker segmentation in the frequency domain. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, (p. 4081).
- Bozonnet, S., Evans, N., & Fredouille, C. (2010). The LIA-EURECOM RT'09 Speaker Diarization System: enhancements in speaker modelling and cluster purification. *proceedings ICASSP 2010*. Dallas, Texas, USA.
- Carletta, J. (2006, January-March). Announcing the AMI Meeting Corpus. *The ELRA Newsletter* 11, 3-5.
- Carter, G. C., Nuttal, A. H., & Cable, P. (1973). The smoothed coherence transform. *Proceedings of the IEEE*, 61(10), 1497-1498.
- Çetin, Ö., & Shriberg, E. (2006). Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. *Proceedings ICASSP*, (pp. 357-360). Toulouse, France.
- Cettolo, M., & Vescovi, M. (2003). Efficient audio segmentation algorithms based on the BIC. *Acoustics Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 6, 537-540.
- Chang, H.-H., Kim, N. S., & Mitra, S. K. (2006). Voice activity detection based on multiple statistical models. 54(6), 1965-1976.
- Charlet, D., Barras, C., & Liénard, J.-S. (2013). Impact of overlapping speech detection on speaker diarization for broadcast news and debates. *Acoustics, Speech and Signal Processing (ICASSP), 2013. IEEE International Conference on*, (pp. 7707-7711).
- Chen, S. S., & Gopalakrishnan, P. S. (1998). Speaker, Environment And Cannel Change Detection And Clustering Via The Bayesian Information Criterion. *DARPA Broadcast News Transcription and Understandin Workshop*, (pp. 127-132). Landsowne, Virginia, USA.
- Chen, S. S., Eide, E., Gales, M., R.A., G., Danvesdy, D., & Olsen, P. (2002). Automatic transcription of Broadcast News. *Speech Communication*, 37(1-2), 69-87.
- Cheng, S. s., & min, W. H. (2004). METRIC-SEQDAC:a hybrid approach for audio segmentation. *Interspeech*.
- Chu, S. M., Tang, H., & Huang, T. S. (2009). Fishervoice and semi-supervised speaker clustering. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*.

- de Campos Niero, M., de Lima Veiga Filho, A., & Adami, A. G. (2014). A comparison of distance measures for clustering in speaker diarization. *2014 International Telecommunications Symposium (ITS)*, 1-5.
- Delacourt, P., & Wellekens, C. (2000). DISTBIC: a speaker-based segmentation for audio data indexing. *32*(1-2), 111-126.
- Dellwo, V., Huckvale, M., & Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production & description for speaker classification. In C. Muller, *Speaker classification I* (pp. 1-20). Berlin, Heidelberg: Springer-Verlag .
- Dighe, P., Ferràs, M., & Bourlard, H. (2014). Modeling Overlapping sSpeech using Vector Taylor Series. *ODYSSEY: The Speaker and Language Recognition Workshop*.
- Eckart, C. (1952). *Optimar rectifier systems for the detection of steady signals*. UC San Diego: Scripps Institution of oceanography: Scripps Institution of oceanography.
- Ellis, D. P., & Liu, J. C. (2004). Speaker turn segmentation based on between-channel differences. *ICASSP*.
- Ellis, D. P., & Liu, J. C. (2004). Speaker turn segmentation based on between-channel differences. *NIST Meeting Recognition Workshop at ICASSP 2004*.
- Evans, N. W., Fredouille, C., & Jean-François, B. (2009). Speaker diarization using unsupervised discriminant analysis of inter-channel delay features. *Acoustics, Speech and Signal Processing 2009, ICASSP 2009, IEEE International Conference on*, 4061-4064.
- Evans, N. W., Fredouille, C., & Jean-François, B. (2009). Speaker diarization using unsupervised discriminant analysis of inter-channel delay features. *Acoustics, Speech and Signal Processing 2009, ICASSP 2009, IEEE International Conference on*, 4061-4064.
- Evans, N., Bozonnet, S., Wang, D., Fredouille, C., & Troncy, R. (2012). A Comparative Study of Bottom-Up and Top-Down Approaches to Speaker Diarization. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(2), 382-392.
- Fernandez, D., Otero, P., & Mateo, C. (2009). An adaptive threshold computation for unsupervised speaker segmentation. (pp. 840-843). Brighton, UK: Proceedings Interspeech.
- Ferrand, C. T. (2002). Harmonics-to-noise-ratio: An index of vocal aging. *16*(4), 480-487.

- Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Fourth Edition.
- Fredouille, C., Bozonnet, S., & Evans, N. (2009). The lia-eurecom rt 09 speaker diarization system. *The Rich Transcription 2009 Meeting Recognition Evaluation Workshop*.
- Friedland, G. (2012). *Using a GPU, Online Diarization = Offline Diarization*. ICSI Technical Report.
- Friedland, G., Janin, A., Imseng, D., Anguera Miro, X., Gottlieb, L., Huijbregts, M., . . . Vinyals, O. (2012). The ICSI RT-09 Speaker Diarization System. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 371-381.
- Friedland, G., Vinyals, O., Huang, Y., & Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5), 985-993.
- Gallardo, A., & San-Segundo, R. (2010). Upm-uc3m system for music and speech segmentation. *Jornadas de Tecnología del Habla FALA 2010*.
- Gallardo-Antolin, A., & Montero, J. (2010). Histogram Equalization-Based Features for Speech, Music, and Song Discrimination. *IEEE Signal Processing Letters*, 17(7), 659-662.
- Geiger, J. T., Eyben, F., Evans, N., Schuller, B., & Rigoll, G. (2013). Using Linguistic Information to Detect Overlapping Speech. *Interspeech 2013*. Lyon, France.
- Geiger, J. T., Vipplerla, R., Bozonnet, S., Evans, N., Schuller, B., & Rigoll, G. (2012). Convolutional Non-Negative Sparse Coding and New Features for Speech Overlap Handling in Speaker Diarization. *Proceedings Interspeech, 2012*.
- Hannan, E. J., & Thomson, P. J. (1973). Estimating Group Delay. *Biometrika*, 60(2), 241-253.
- Huijbregts, M., van Leeuwen, D. A., & Wooters, C. (2012). Speaker Diarization Error Analysis Using Oracle Components. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 393-403.
- Huijbregts, M., van Leeuwen, D., & de Jong, F. (2009). Speech Overlap Detection in a Two-Pass Speaker Diarization System. *Interspeech 2009*, (pp. 1063-1066).
- Imseng, D., & Friedland, G. (2010). Tuning-Robust Initialization Methods for Speaker Diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8), 2028-2037.

- Izumitani, T., Mukai, R., & Kashino, K. (2008). A background music detection method based on robust feature extraction. *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 13-16.
- Kim, H.-G., Ertelt, D., & Sikora, T. (2005). Hybrid Speaker-Based Segmentation System Using Model-Level Clustering. *ICASSP*.
- Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1), 12-40.
- Knapp, C. H., & Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *Acoustic, Speech and Signal Processing, IEEE Transactions on*, 24(4), 320-327.
- Koh, E. C., Sun, H., Nwe, T. L., Nguyen, T. H., Ma, B., Chng, E.-S., . . . Rahardja, S. (2008). Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I2R-NTU Submission for the NIST RT 2007 Evaluation. *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA*, 484--496.
- Kumatani, McDonough, J., Lehman, J. F., & Raj, B. (2011). Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. *Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*, 1-6.
- Kwon, S., & Narayanan, S. (2005). Unsupervised Speaker indexing using generic models. *Speech and Audio Processing, IEEE Transactions on*, 13(5), 1004-1013.
- Lavner, Y., & Ruinskiy, D. (2009). A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. *Lavner, Y. and D. Ruinskiy. 2009. A Decision-Tree-Based Algor* *EURASIP Journal on Audio, Speech, and Music Processing*.
- Lin, S.-N., & Chern, S.-J. (1998). A new adaptive constrained LMS time delay estimation algorithm. *Signal Processing*, 71(1), 29-44.
- Liu, D., & Kubala, F. (1999). Fast speaker change detection for broadcast news transcription and indexing. *Proceedings of the 6th European Conference on Speech Communication and Thechnology*. 1031-1034.
- Lorenzo-Trueba, J., Barra-Chicote, R., Raitio, T., Obin, N., Alku, P., Yamagishi, J., & Montero, J. M. (2012). Towards Glottal Source Controlsability in Expressive Speech Synthesis. *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*. Portland, Oregon.

- Lorenzo-Trueba, J., Martínez, B., Barra-Chicote, R., López-Ludeña, V., Ferreiros, J., Yamagishi, J., & Montero, J. M. (2012). Towards an Unsupervised Speaking Style Voice Building Framework: Multi-Style Speaker Diarization. *InterSpeech 2012*. Portland (Oregon).
- Lu, L., & Zhang, H.-J. (2002). Real-Time Unsupervised Speaker Change Detection. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*.
- Lu, L., & Zhang, H.-J. (2005). Unsupervised speaker segmentation and tracking in real-time audio content analysis. *Multimedia Systems*, 10(4), 332-343.
- Luque, J., Segura, C., & Hernando, J. (2008). Clustering initialization based on spatial information for speaker diarization of meetings. *Interspeech*, (pp. 383-386).
- Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 561-580.
- Martínez-González, B., Pardo, J. M., Echeverry-Correa, J. D., & J.M. Montero, J. (2014, mar). New experiments on speaker diarization for unsupervised speaking style voice building for speech synthesis. *Procesamiento del Lenguaje Natural*, 52, 77-84.
- Martínez-González, B., Pardo, J. M., Echeverry-Correa, J. D., & San-Segundo, R. (2016). Spatial Features Selection for Unsupervised Speaker Segmentation and Clustering. *Expert Systems With Applications*.
- Martínez-González, B., Pardo, J. M., Echeverry-Correa, J. D., Vallejo-Pinto, J. A., & Barra-Chicote, R. (2012). Selection of TDOA parameters for MDM speaker diarization. *Interspeech*. Portland (OG).
- Martínez-González, B., Pardo, J. M., San-Segundo, R., & Montero, J. M. (2016). Influence of transition cost in the segmentation stage of speaker diarization. *Odyssey* (pp. 385-392). Bilbao: odyssey.
- Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, 20(2-3), 303-330.
- Mirghafori, N., & Wooters, C. (2006). Nuts and flakes: A study of data characteristics in speaker diarization. *International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Toulouse.
- Mirghafori, N., Stolcke, A., Wooters, C., Pirinen, T., Bulyko, I., Gelbart, D., . . . Ostendorf, M. (2004). "From switch-board to meetings: Development of the 2004 ICSI-SRI-UW meeting

- recognition system. *Int'l Conf. Spoken Language Processing (ICSLP '04)*. Jeju island, Korea.
- Misra, H., Ikbai, S., Bourlard, H., & Hermansky, H. (2004, May). Spectral entropy based feature for robust ASR. *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '04)*, 1, 1, 193-6 vol.1.
- Moattar, M. H., & Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, 54, 1065-1103.
- Moreno-Sandoval, A., G., D. I., Alcántara, M., González, A., Guirao, J., & De la Torre, R. (2005). The spanish corpus" C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages. (pp. 135-161). Amsterdam: John Benjamins Publishing Company.
- Murphy, P. J. (1999). Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *The Journal of the Acoustical Society of America*, 105, 2866.
- Nguyen, T., & Sun et al., H. (2009). The IIR-NTU Speaker Diarization Systems for RT 2009. *NIST RT09 Evaluation Workshop, Florida*.
- Nwe, T. L., Sun, H., Li, H., & Rahardja, S. (2009). Speaker diarization in meeting audio. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 4073-4076.
- Nwe, T. L., Sun, H., Ma, B., & Li, H. (2012). Speaker Clustering and Cluster Purification Methods for RT07 and RT09 Evaluation Meeting Data. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 461-473.
- Otterson, S. (2007). Improved location features for meeting speaker diarization. *Proc. Interspeech*, 1849–1852.
- Otterson, S., & Ostendorf, M. (2007). Efficient use of overlap information in speaker diarization. *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, (pp. 683-686).
- Panagiotakis, C., & Tziritas, G. (2005). A Speech/Music Discriminator Based on RMS and Zero-Crossings. *IEEE Transactions on Multimedia*, 7(1).

- Pardo, J. M., Anguera, X., & Wooters, C. (2006). Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. (pp. 2194–2197). *Proceedings of International Conference on Spoken Language Processing*.
- Pardo, J. M., Anguera, X., & Wooters, C. (2007, September). Speaker diarization for multiple-distant-microphone meetings using several sources of information. *Computers, IEEE Transactions on*, 56(9), 1212–1224.
- Pardo, J. M., Barra-Chicote, R., San-Segundo, R., Córdoba, R., & Martínez-González, B. (2012). Speaker diarization features: The UPM contribution to the RT09 evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2), 426-435.
- Pardo, J. M., Barra-Chicote, R., San-Segundo, R., Córdoba, R., & Martínez-González, B. (2012). Speaker diarization features: The UPM contribution to the RT09 evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2), 426-435.
- Raitio, T. A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., & Alku, P. (2011). HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), 153-165.
- Reed, F. A., Feintuch, P. L., & Bershad, N. J. (1981). Time Delay Estimation Using the LMS Adaptive Filter-Static Behavior. *Acoustics, speech, and signal processing, IEEE*, 29(3), 561-571.
- Rich Transcription Evaluation Project, National Institute of Technology (NIST), 2002-2009*, . (n.d.). Retrieved 2012, from <http://www.itl.nist.gov/iad/mig/tests/rt>
- Roth, P. R. (1971). Effective measurements using digital signal analysis. *Spectrum, IEEE*, 8(4), 62-70.
- Rougui, J.-e.-., R., Anoutajdine, D., Gelgon, M., & Martinez, J. (2006). Fast Incremental Clustering of Gaussian Mixture Speaker Models for Scaling up Retrieval In On-Line Broadcast., 5.
- Shriberg, E. (2007). Higher-Level Features in Speaker Recognition. In C. Müller, *Speaker Classification I. Series Lecture Notes in Computer Science* (Vol. 4343, pp. 241-259). Springer Berlin Heidelberg.
- Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proceedings EUROSPEECH*, (pp. 1359-1362). Aalborg, Denmark.

- Siegles, M., Jain, U., Raj, B., & Stern, R. (1997). Automatic segmentation, classification and clustering of broadcast news audio. (pp. 97-99). Chantilly: Proceedings of DARPA Speech Recognition Workshop.
- Sinclair, M., & King, S. (2013). Where are the challenges in speaker diarization? *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*.
- Sinclair, M., & King, S. (2013). Where are the challenges in speaker diarization? *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*.
- Soong, F., & Juang, B. (1984). Line spectrum pair (LSP) and speech data compression. *CASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing, 9*, pp. 37-40.
- Stafylakis, T., Datsouros, V., & Carayannis, G. (2009). Redefining the bayesian information criterion for speaker diarization. *Proccedings Interspeech*.
- Sun, H., Nwe, T. L., Ma, B., & Li, H. (2009). Speaker diarization for meeting room audio. *Proccedings of Interspeech*, (pp. 900-903).
- Talkin, D. (1995). *A Robust Algorithm for Pitch Tracking (RAPT)*. (W. B. Kleijn, & K. K. Paliwal, Eds.) new york: elsevier.
- Thomas, S., Ganapathy, S., & Hermansky, H. (2008). Recognition of reverberant speech using frequency domanin linear prediction. *15*, 681-684.
- Tranter, S. E., Yu, K., Evermann, G., & Woodland, P. C. (2004). Generating and avaluating segmentations for automatic speech recognition of conversational telephone speech. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*.
- Tranter, S., & Reynolds, D. (2006). An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(5), 1557–1565.
- van Leeuwen, D. A. (2008). Progress in the AMIDA speaker diarization system for meeting data. In *Multimodal Technoogies for Perception of Humans. Lecture Notes in Computer Science* (pp. 475-483). Springer Berlin Heidelberg.
- Vaquero, C., Vinyals, O., & Friedland, G. (2010). A hybrid approach to online speaker diarization. *Interspeech, 2010*, (pp. 2638-2641).
- Vijayasenan, D., & Valente, F. (2012a). Dimensionality Reduction of Large TDOA Vectors for Speaker Diarization. *SAPA-SCALE conference 2012*, 64-67.

- Vijayasenan, D., & Valente, F. (2012b). Speaker diarization of meetings based on large TDOA feature vectors. *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 4173-4176.
- Vijayasenan, D., Valente, F., & Boulard, H. (2009). Mutual information based channel selection for speaker diarization of meetings data. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 4065-4068.
- Vijayasenan, D., Valente, F., & Boulard, H. (2010). Multistream speaker diarization beyond two acoustic feature streams. (pp. 4950-4953). *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*.
- Vijayasenan, D., Valente, F., & Boulard, H. (2011). An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2), 431-438.
- Vinyals, O., & Fiedland, G. (2008). Modulation spectrogram features for speaker diarization. *Proceedings of Interspeech*.
- Wiener, N. (1949). *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*.
- Wooters, C., & Huijbregts, M. (2008). The icsi rt07s speaker diarization system. In *Lecture Notes in Computer Sciences* (Vol. 4625, pp. 509–519).
- Wu, C.-H., & Hsieh, C.-H. (2006). Multiple change-point audio segmentation and classification using an MDL-based Gaussian model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2), 647-657.
- Yang, J., Zhang, D., Jin, Z., & Yang, J.-y. (2006). Unsupervised Discriminant Projection Analysis for Feature Extr. *18th International Conference on Pattern Recognition (ICPR'06)*, 1, 904-907.
- Yang, Y., & Jiu, L. (2014). Exploring the Large-Scale TDOA Feature Space for Speaker Diarization. *HCI International 2014- Posters' Extended Abstracts Communications in Computer and Information Science*, 434, 551-556.
- Yella, S. H. (2012). Speaker diarization of overlapping speech based on silence distribution in meeting recordings. *Interspeech 2012*. Portland, Oregon.

- Yella, S. H., & Boulard, H. (2014). Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12), 1688-1700.
- Yokoyama, R., Nasu, Y., Iwano, K., & Shinoda, K. (2013). Detection of overlapped speech using lapel microphones in meeting. *Speech Communication*, 55(10), 941-949.
- Zelenak, M., Segura, C., Luque, J., & Javier, H. (2012). Simultaneous Speech Detection With Spatial Features for Speaker Diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 436-446.
- Zhou, B., & Hansen, J. (2000). Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. (pp. 714-717). Beijing, China: Proceedings ISCLP 2000.
- Zhu, X., Barras, C., Meignier, S., & Gauvain, J.-L. (2005). Combining speaker identification and bic for speaker diarization. *European Conference on Speech Communications Technology*. Lisbon.