TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2011

38

Carlos Vaquero Avilés-Casco

Robust diarization for speaker characterization (Diarización robusta para caracterización de locutores)

Departamento

Ingeniería Electrónica y Comunicaciones

Director/es

Ortega Giménez, Alfonso

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad
Zaragoza
1542

Tesis Doctoral

ROBUST DIARIZATION FOR SPEAKER CHARACTERIZATION (DIARIZACIÓN ROBUSTA PARA CARACTERIZACIÓN DE LOCUTORES)

Autor

Carlos Vaquero Avilés-Casco

Director/es

Ortega Giménez, Alfonso

UNIVERSIDAD DE ZARAGOZA

Ingeniería Electrónica y Comunicaciones

2011

1542 Universidad Zaragoza

Department of Electronic Engineering
and Communications

Ph.D. Thesis

Robust Diarization for Speaker Characterization

Diarización Robusta para Caracterización de Locutores

Carlos Vaquero Avilés-Casco

Thesis Advisor
Dr. Alfonso Ortega Giménez

October 31, 2011

Abstract

The task of speaker characterization, which aims at describing the particular and distinctive peculiarities of a person's speech, is essential for several speech based technologies and applications. The clearest example is voice based biometrics, but also speech recognition can take advantage of speaker characterization using speaker adaptation techniques. Speaker characterization approaches require large datasets with speaker labels to operate, but in several environments, even when there are datasets available, they are not directly useful for speaker characterization. An usual problem is to find that every recording in the dataset contains several speakers, and there are no labels available indicating when every speaker is speaking. The solution to this problem is the use of speaker diarization, which aims at answering the question "Who spoke when?".

This thesis focuses on providing robustness to speaker diarization for real life speaker characterization applications. For this purpose two complementary objectives are pursued: first, the development of very accurate speaker diarization systems is desired in order to ensure that speaker characterization applications will operate correctly when they make use of recordings containing more than a single speaker. Second, quality assessment strategies for speaker diarization are desired in order to detect those recordings that will be reliable for speaker characterization.

To achieve these objectives, we review the traditional diarization solutions and we analyze the impact of diarization errors on a speaker verification task. It is shown that the traditional diarization strategies may not be accurate enough for certain applications. To solve this problem, a new approach for speaker diarization based on the most recent innovations in the field of speaker recognition is proposed, including a novel variability compensation strategy for speaker diarization. These innovations increase the accuracy of speaker diarization and speaker verification when considering two-speaker telephone conversations, an environment quite usual in voice biometrics applications. Then, the analysis is extended to problems involving more that two speakers, and new approaches for speaker clustering are analyzed. The proposed diarization solution also outperforms the traditional ones when the number of speakers is unknown.

Finally, a study on quality assessment strategies for speaker diarization is included. Several confidence measures and a methodology to detect recordings with reliable diarization hypotheses are proposed. This methodology enables us to retrieve a subset of reliable recordings, ensuring that a speaker characterization application will not obtain significant degradation due to speaker diarization errors when the retrieved subset is considered. The methodology is shown to be helpful for speaker characterization applications as speaker verification and speaker clustering. In addition, it is shown to increase the accuracy of speaker diarization when it is combined with a strategy to automatically generate and select several diarization hypotheses for a single recording.

The retrieval of reliable recordings is quite useful to process a dataset in a semi-supervised fashion, since only the subset of recordings that are not detected as reliable should be inspected manually. The subset to inspect manually is expected to be small when accurate speaker diarization systems, as the one proposed in this thesis, are considered.

Resumen

La tarea de caracterización de locutores, cuyo objetivo es describir las peculiaridades particulares y distintivas del habla de una persona, es esencial para muchas tecnologías y aplicaciones basadas en el habla. El ejemplo más claro es la biometría basada en la voz, pero también el reconocimiento del habla puede aprovecharse de la caracterización de locutores utilizando técnicas de adaptación al locutor. Las técnicas de caracterización de locutores requiren grandes bases de datos con etiquetas de locutor para operar, pero en muchos entornos, incluso cuando existen bases de datos apropiadas, éstas no son útiles para la caracterización de locutores. Un problema habitual es que cada grabación de la base datos contiene muchos locutores, y no existen etiquetas indicando cuando habla cada locutor. La solución a este problema es el uso de diarización de locutores, cuyo objetivo es responder a la pregunta "¿Quién ha hablado en cada momento?".

Esta tesis se centra en proporcionar robustez a la diarización de locutores para que sea utilizada en aplicaciones reales de caracterización de locutores. Para ello, se persiguen dos objetivos complementarios: en primer lugar, se requiere el desarrollo de sistemas de diarización precisos, para asegurar que las aplicaciones de caracterización de locutores operarán correctamente cuando utilicen grabaciones con más de un locutor. En segundo lugar, se requieren técnicas para la evaluación de la calidad de las hipótesis de diarización, para detectar aquellas grabaciones que serán fiables para la caracterización de locutores.

Para alcanzar estos objetivos, se revisan las técnicas de diarización tradicionales, así como el impacto que tienen los errores de diarización en una tarea de verificación de locutor. Se demuestra que las estrategias tradicionales de diarización pueden no ser suficientemente precisas para determinadas aplicaciones. Para resolver este problema, se propone una nueva aproximación para diarización de locutores basada en los recientes avances obtenidos en el campo de reconocimiento de locutores. Estos avances incrementan la precisión de la diarización y verificación de locutores cuando se consideran conversaciones telefónicas de dos locutores, un entorno muy habitual en aplicaciones biométricas basadas en voz. Después, el análisis se extiende a problemas con más de dos locutores, y se analizan nuevas técnicas de clustering de locutores. El sistema de diarización propuesto también obtiene mejores prestaciones que los sistemas tradicionales cuando se desconoce el número de locutores.

Finalmente, se incluye un estudio sobre evaluación de calidad para diarización. Se proponen varias medidas de confianza y una metodología para la detección de grabaciones con hipótesis de diarización fiables. Esta metodología permite recuperar una parte de una base de datos dada, compuesta por grabaciones fiables, de forma que una aplicación de caracterización de locutores no obtendrá una degradación significativa debido a errores de diarización. Se demuestra que la metodología es útil para aplicaciones de caracterización de locutores tales como verificación o clustering de locutores. Además se demuestra que, en combinación con una estrategía de generación y selección de hipótesis, incrementa la precisión de la diarización de locutores.

La recuperación de grabaciones fiables es muy útil para procesar una base de datos de forma semisupervisada, ya que sólo es necesario inspeccionar manualmente la porción de la base de datos no detectada como fiable. La porción a inspeccionar será pequeña cuando se consideren sistemas de diarización precisos, como el que se propone en esta tesis.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Speaker Characterization and Diarization

Speaker characterization is the task of describing the particular and distinctive peculiarities of a person's speech. This task is essential for voice based biometrics, which is not the most accurate nor the less intrusive form of Biometrics technology [Jain *et al.*, 2008], but it has a clear advantage: it is the only one that can be performed when only the voice of the subject is available.

This fact, that seems obvious, opens several possibilities for the use of human speech for biometric purposes. The most clear one is the use of telephone channels for biometric applications. The number of mobile phones in use in the world has risen dramatically during the last years, up to a point that in the third quarter of 2010, there were around 5.282 billion mobile phone subscriptions in the world, which implies 76.2 mobile phones per 100 inhabitants. These numbers are more exaggerated in the developed countries, where there were 116.1 mobile phone subscriptions per 100 inhabitants, more than 1 mobile phone per inhabitant [ITU-D, 2010].

In a telephone environment, an environment which is becoming routine for most people day by day, the only natural way to perform biometric identification is the use of speech. The biometric technology that enables this is known as *Speaker Recognition* and it encompass the task of *speaker verification* and *identification*. Some fields of application of speaker recognition include *forensics*, *surveillance* or *identity authentication* and *access control*.

On the other hand, the recent development of the Information and Communication Technologies (ICT) has motivated the generation and availability of a huge and increasing number of multimedia information resources. The most evident example is the mass media. In 1920, there were just 4 licensed radio stations in US, currently there are over 65000 [FCC, 2011]. This effect is much more exaggerated in TV broadcasting. Since the beginning of the TV history, the development of new transmission media such as cable TV or satellite TV increased dramatically the number of licensed (and unlicensed) TV channels. Recently, the development of standards for digital transmission and broadcasting of multimedia signals has enabled cheaper accessibility to a wide variety of TV channels for most TV users, and a much more efficient use of the available spectrum, increasing the number of TV channels that can be available in a region.

But the mass media is not the only source of increasing multimedia content. Nowadays, people have the technological means to cheaply generate and share multimedia documents, so the amount of multimedia information generated across the world day by day is intractable, unless it is classified and indexed in an automatic fashion. Within the indexation process, the identification of well-known speakers (politicians, presenters...) is highly desirable, not only for indexation purposes but also to improve automatic transcription using adapted acoustic models for speech recognition. In the case of audiovisual resources, speaker recognition is one of the few biometric technologies that can be used, and probably the most accurate one. In case of audio resources, such as recordings from the radio, speaker recognition is the only biometric technology that enables us to identify the speakers involved.

Thus, speaker recognition is also necessary for the indexation of multimedia databases. The technology aimed at indexing multimedia resources based on its audio content is known as *audio indexation*. Audio indexation systems extract all available information from an audio signal, including not only the identity of the speakers but also the transcriptions of their speech or any other information that may help to classify and eventually retrieve the audio recording (date, place), and store this information in the form of meta-data. This process of extracting all available information from an audio signal is known by the research community as *Rich Transcription*, since information includes not only the transcription of the speech but additional information as well. Audio indexation systems are usually part of *Spoken Document Retrieval (SDR)* systems [Hansen *et al.*, 2005], which also include an information retrieval engine in order to search and rank those audio documents that best match a user request. Audio Indexation can be useful not only for the mass media but also for many other applications. For example, audio indexation applications include the indexation of meetings for particular use [Kazman *et al.*, 1995], the indexation of lessons in the field of academics [Zizka *et al.*, 2010], or the indexation of conferences or even parliament sessions and speeches [Löffler *et al.*, 2002].

Speaker characterization is a key technology in the environments previously described. In order to characterize a speaker certain amount of data from that speaker is required. However, in these environments is usual to find more than one speaker involved in a single conversation or recording. In this situation, the segregation of the audio segments produced by the desired speakers is essential for a correct speaker characterization. This process aimed at segmenting and classifying the audio into different homogeneous classes according to the speaker that produced every segment is known as *speaker diarization*. Speaker diarization is a subtask of a more general task known as *audio diarization* or *audio segmentation and classification*, which involves not only the annotation of the speakers present in the signal but also different classes such as music or different background noise sources.

Speaker diarization aims at answering the question “who spoke when?” given an audio signal. The word *diarization* refers to the the task of creating a diary of events occurred in the audio signal. Usually, speaker diarization systems work in a unsupervised fashion, in the sense that no prior models of the classes that might be present in the audio signal are considered. Most of the speaker diarization algorithms operate following two steps: *segmentation* and *clustering*. The segmentation of an audio signal is the task of finding the boundaries between the different acoustic sources present in the signal, splitting the single into acoustically homogeneous segments. The clustering step then agglomerates those segments into acoustically homogeneous classes. This procedure is known as *bottom-up* or *agglomerative hierarchical clustering* (AHC) strategy for clustering in speaker diarization systems. There are also *top-down* strategies that starts with a single cluster and fragment it into homogeneous classes.

Speaker diarization is usually a support technology for other tasks such as speaker characterization or recognition, or Automatic Speech Recognition (ASR). Thus there are several applications that include speaker diarization, not as main technology but as an essential support technology. All speaker recognition applications mentioned before as well as audio indexation systems can benefit from a robust speaker diarization system.

Traditionally the domains of application of speaker diarization have been telephone speech, broadcast news and meetings. Telephone speech is a domain quite related to speaker recognition. In fact, speaker diarization systems for telephone conversations started being evaluated by the US National Institute for Standards and Technology (NIST) in 1996, within the Speaker Recognition Evaluations, which are still running. Diarization on broadcast news domain was mainly encouraged by the NIST ARPA Continuous Speech Recognition (CSR) Hub-4 evaluation, and later by the DARPA's EARS program (DARPA Effective, Affordable, Reusable Speech-to-Text) [EARS, 2004]. Hub-4 aimed at automatic speech transcription of broadcast news, and acoustic segmentation and clustering became important for improved transcription using speaker adaptation and normalization techniques. EARS program aimed at rich transcription of broadcasted news content. Finally, diarization on meetings have caught the attention of researchers in part due to the impulse of CHIL (Computers in the Human Interaction Loop) [CHIL, 2006] and AMI (Augmented Multiparty Interaction) projects [AMI, 2006].

This thesis focuses on speaker diarization for applications that need to characterize different speakers. Actually, both speaker diarization and characterization are quiet related: Speaker diarization is needed to perform correct speaker characterization when more than one speaker is involved in the available audio signal. On the other hand, finding those particularities that make distinctive different speakers is undoubtedly helpful for speaker diarization.

1.2 Brief Historic Evolution

Speaker Diarization is a term relatively new in the field of audio processing. It was introduced with the Rich Transcription (RT) Evaluations organized by NIST, and became popular with the beginning of the EARS program and the RT 2004 Fall evaluation. However, the research community has worked on speaker segmentation and clustering before.

In the field of speaker segmentation, the first works date from the early nineties. In [Gish *et al.*, 1991], the problem of segmentation and clustering was formulated as a hypothesis selection problem and solved using the Generalized Likelihood Ratio (GLR). This first work established the framework for acoustic change detection and Agglomerative Hierarchical Clustering (AHC) that most speaker diarization systems follow nowadays. Later, in [Sugiyama *et al.*, 1993], a method for speaker segmentation based on Hidden Markov Models (HMM) and a method for speaker clustering based on Vector Quantization (VQ) were proposed, but assuming that the number of speaker was known.

In 1996, NIST started the Hub-4 evaluation, oriented to speaker independent speech recognition. This evaluation motivated further research on speaker segmentation and clustering as they became key technologies in order to perform speaker adaptation and normalization for ASR. In the context of this evaluation, in [Siegler *et al.*, 1997] the symmetric Kullback-Leibler distance (KL2) was proposed for speaker segmentation and clustering, following the framework proposed in [Gish *et al.*, 1991]. Then, in [Chen and Gopalakrishnan, 1998], the Bayesian Information Criterion (BIC) was proposed as metric for speaker segmentation and clustering, again following the same framework. This last work was a milestone in the field of speaker diarization, and currently most speaker diarization systems rely on BIC for speaker segmentation, clustering or both. A classification for speaker segmentation approaches, which is still in use nowadays, was also proposed in [Chen and Gopalakrishnan, 1998].

During the last decade, the evolution of speaker diarization systems was mainly motivated by the DARPA EARS program and the NIST RT evaluations. Most speaker diarization approaches adopted the state-of-the art technique of iterative clustering and segmentation, presented in [Reynolds and Torres-Carrasquillo, 2005]. In this approach, BIC is considered for acoustic change detection and AHC as in [Chen and Gopalakrishnan, 1998], and once the optimal number of clusters is achieved, several segmentation passes using dynamic programming algorithms as the Viterbi algorithm are performed, considering Gaussian Mixture Models (GMM) for speaker modeling. With the introduction of meetings for the RT evaluations, the research community have focused also on performing speaker diarization when the signals from multiple distant microphones are available.

The relation between speaker diarization and characterization is evident in the literature. In [Gish *et al.*, 1991] the statistical modeling of speakers is cited as a main contribution to make possible speaker segmentation and clustering. GMMs were not used for speaker modeling in speaker diarization until they were validated for speaker characterization [Reynolds, 1995a] [Reynolds and Rose, 1995]. Speaker identification techniques based on Maximum a Posteriori (MAP) adaptation from a Universal Background Model [Reynolds *et al.*, 2000] were also applied for speaker diarization in [Zhu *et al.*, 2005]. Recently, new advances in the field of speaker characterization have been introduced, mainly motivated by the NIST Speaker Recognition Evaluations (SRE). These advances, which include Joint Factor Analysis (JFA) [Kenny *et al.*, 2007] or the use of i-vectors [Dehak *et al.*, 2010] for Speaker Recognition, have encouraged new approaches for speaker diarization [Castaldo *et al.*, 2008], [Kenny *et al.*, 2010], [Vaquero *et al.*, 2010a]. These new approaches for speaker diarization have proved to outperform the traditional ones in telephone environments.

1.3 Motivation of this Work

During the last years, the research community have focused on speaker diarization in meetings, mainly motivated by the recent RT evaluations organized by NIST. Some challenges in these evaluations are dealing with speech recorded using far field microphones, taking advantage of using multiple distant microphones [Pardo *et al.*, 2007]. Another challenges are traditional problems of speaker diarization that have not been solved yet, as determining the actual number of speakers present in a speech signal [Valente and Wellekens, 2004], being able to perform diarization in small segments [Imseng and Friedland, 2009], detecting and dealing with overlapped speech [Boakye *et al.*, 2008], or performing diarization in streaming [Castaldo *et al.*, 2008] or even on-line [Vaquero *et al.*, 2010c], as fast as possible. Recently, the problem of speaker diarization has expanded to multimedia environments and there has been an important effort in using video information to perform speaker diarization, in part motivated by projects as AMI or CHIL.

This thesis focuses on speaker diarization for speaker characterization, and in addition to some of the challenges previously defined, it deals with some issues that have not been deeply analyzed by the research community. Firstly, since speaker diarization is in general a support technology for speaker characterization, the impact of diarization errors on the performance of speaker characterization based applications needs to be studied. The importance and impact of performing diarization for speaker adaptation and normalization for ASR have been broadly studied within the framework of Broadcast News [Gauvain *et al.*, 1999] and meeting [Stolcke *et al.*, 2010] (RT) transcription evaluations. However, the importance of speaker diarization in speaker recognition applications have only been studied recently, in part motivated by the NIST SRE, and only a few works analyze the impact of speaker diarization errors on speaker recognition systems [Reynolds *et al.*, 2009], but the sensitiveness of these systems to speaker diarization errors is not analyzed exhaustively.

In the line of the previous issue, the accuracy of speaker diarization systems will need to be increased to meet the requirements of certain speaker characterization systems. In order to improve speaker diarization, several lines can be studied. Among them, the most promising one seems to be the use of new approaches based on the recent advances obtained in the field of speaker recognition [Castaldo *et al.*, 2008], [Kenny *et al.*, 2010], [Vaquero *et al.*, 2010a]. Also, the study of new initialization techniques is promising: In [Imseng and Friedland, 2010] improvements in the initialization of a speaker diarization system increase the accuracy of the system significantly.

Another interesting line to be studied is the compensation of variability for speaker diarization. Recently there has been a huge effort in compensating for inter-session variability to improve the performance of speaker verification systems. Compensating for this type or other types of variability can be useful for speaker diarization, and the research community have not tried to analyze and compensate variability sources that affect a speaker diarization system.

On the other hand, current approaches for speaker clustering have several problems. The first an most evident is the fact that they do not take into account session variability. Thus, they cannot operate over several different sessions, which may be useful for several speaker characterization applications. In addition, the task of determining the number of speakers over several sessions or within a single session is still not solved. In these lines, speaker characterization techniques that are known to work successfully for speaker recognition can be used for speaker clustering, within speaker diarization systems.

Finally, speaker characterization applications that make use of diarization usually do not have information about the quality or accuracy of the output of a diarization system. Some speaker characterization applications can be severely affected by errors in diarization and may not need all the audio signals available to work correctly. For example, a speaker recognition system that uses diarization to segment conversations and to train speaker models, may not need all available conversations from a given speaker to train the corresponding speaker model. However this speaker model can be affected and the accuracy of speaker recognition degraded if severe diarization errors exist in some of the hypotheses given by the diarization system. In these situations, obtaining confidence measures that give an idea of the accuracy of the diarization system can be highly desirable. Although there is work on quality and confidence measures for speaker recognition [Solewicz and Koppel, 2005] [Garcia-Romero, 2006] [Harriero *et al.*, 2009], and some of them could be applicable to speaker diarization, there is no work prior to this thesis on specific confidence measures for speaker diarization. These confidence measures can be used to improve the performance of a speaker diarization system and to let a speaker characterization application to decide how to deal with every output hypothesis of the diarization system, increasing its reliability.

1.4 Objectives and Methodology

The main objective of this work is to provide the tools to extract the maximum amount of useful information from a given dataset composed of conversations containing several speakers in order to utilize it for speaker characterization.

For this purpose, two complementary objectives are pursued. On one hand, more accurate speaker diarization systems are needed to avoid the use of information from an undesired speaker to characterize another speaker. On the other hand, an automatic method to assess the quality of the diarization hypothesis obtained for each recording of the dataset will help to segregate useful recordings from the dataset, or to process the whole dataset in a semi-supervised fashion.

In order to achieve more accurate speaker diarization systems, the recent innovations in the field of speaker verification (JFA [Kenny *et al.*, 2007] and i-vectors [Dehak *et al.*, 2010]) will be applied to the field of speaker diarization. The sources of desired and undesired variability in speaker diarization will be analyzed and modeled. To assess the quality of the diarization hypotheses obtained, several confidence measures for speaker diarization system will be studied.

As speaker characterization technology, a state-of-the-art speaker verification system is considered. Thus, every improvement obtained in the accuracy of the speaker diarization system will be validated in a speaker verification task, in order to determine whether the improvement obtained by the proposed technique is reflected in a speaker characterization task.

Since most databases for speaker verification that involve several speakers in a single recording are composed of two-speaker telephone conversations, this study focuses on improving diarization accuracy and assessing the quality of telephone conversations containing two speakers. Nevertheless, to expand the proposed innovations in speaker diarization for conversations containing more than a single speaker, new approaches for speaker clustering are also studied. Speaker clustering is analyzed in large datasets as a general speaker characterization application, to particularize then to the case of speaker diarization, where the audio segments to cluster are obtained from a single recording.

1.5 Outline

This thesis is divided into nine Chapters organized in five Parts. The first Part contains two Chapters that analyzes the need of speaker diarization for speaker characterization. In Chapter 2, traditional and recently proposed speaker diarization techniques are deeply reviewed, introducing the state-of-the-art in this research field. Chapter 3 briefly analyzes the state-of-the-art approaches for speaker recognition, introducing the speaker verification task considered to validate the any improvement obtained in speaker diarization. In addition, the importance of speaker diarization for speaker characterization is studied in this Chapter, considering a traditional speaker diarization system.

In the second Part, new techniques for speaker diarization are proposed, with the purpose of improving the diarization accuracy. This Part is composed of two Chapters. Chapter 4 proposes and analyzes an innovative speaker diarization system based on the recent advances developed in the field of speaker recognition. In Chapter 5, the different sources of variability involved in speaker diarization are studied in order to determine which sources help for speaker diarization and which ones should be compensated.

The third Part is composed of a single chapter, Chapter 6. This Chapter studies innovative techniques for speaker clustering in large datasets based on recent advances in the field of speaker recognition. These techniques are then combined to the innovative speaker diarization techniques proposed in Chapters 4 and 5 in order to expand these techniques and face conversations containing an unknown number of speakers.

The fourth Part includes two Chapters that study and validate quality assessment techniques for speaker diarization in two-speaker telephone conversations. In Chapter 7, the concept of quality in speaker diarization for speaker characterization is introduced, and a technique for quality assessment in speaker diarization is proposed. This technique makes use of several confidence measures for speaker diarization, presented in this thesis, in order to determine the quality of a given diarization hypotheses. Chapter 8 validates the proposed technique in two different use cases. The first one makes use of quality assessment to improve the accuracy in speaker diarization, while the second one utilizes quality assessment for diarization in the task of speaker clustering in large datasets composed of two-speaker telephone conversations.

Finally, in the fifth Part, Chapter 9, remarks the main conclusions of this work.

Part I

The need of Diarization for Speaker Characterization

2

State of the Art

This chapter brings together the main approaches for the task of speaker diarization. Firstly, the task of speaker diarization is briefly introduced, and its main subtask described. These subtasks comprehend feature extraction, speaker segmentation and speaker clustering, and the state of the art of each one of them is analyzed. In addition, different techniques to evaluate the performance of speaker diarization systems are introduced. Then, the main environments of application of speaker diarization are commented and some recent studies on these fields are described. The information presented here is not complete, but there are other works that also bring together the main approaches for speaker diarization. A very complete one from 2006 is presented in [Anguera, 2006].

Finally, a brief review of the state of the art of speaker characterization techniques for speaker recognition is also done, analyzing whether or not they may be useful for speaker diarization. Also, the latest works that make use of state of the art in speaker characterization for speaker diarization are introduced.

2.1 Introduction to Speaker Diarization

Speaker diarization refers to the set of techniques that aim at answering the question '*Who spoke when?*' given an audio signal. Usually, speaker diarization systems work in an unsupervised fashion, since in general, no prior information of the speakers involved in the audio signal is available, the number of speakers is unknown, and within the diarization task, every speaker must be labeled with a different label.

The task of speaker diarization date from the early nineties, but the major progress have been achieved from the late nineties to the present. Actually, the term *speaker diarization* dates from the beginning of the two-thousands, when the Rich Transcription (RT) Evaluations organized by NIST started, and the term became popular with the beginning of the EARS program [EARS, 2004] and the RT 2004 Fall evaluation [NIST, 2004].

Conceptually it can be seen as a part of the more general task of audio diarization. The goal of audio diarization is to segment and classify the audio into different homogeneous classes according to the source that produced every segment obtained. These classes include speech, silence, acoustic events or different acoustic environments, such as background noises, music and so on. We refer to speaker diarization when every different class is a unique speaker.

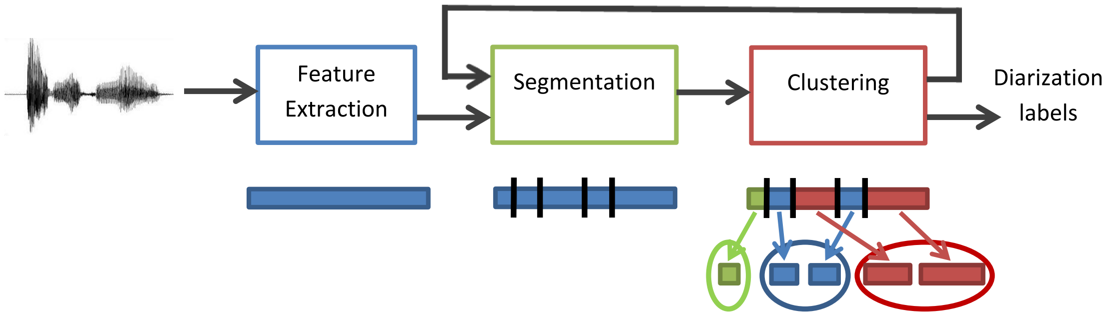


Figure 2.1: *Scheme of a speaker diarization system.*

The detection of silence, background noises, music, acoustic events or from a more general point of view, non-speech classes is a previous task needed to perform speaker diarization, and is usually included in works on speaker diarization as a subtask of speaker diarization. However, in general, the acoustic features and algorithms considered to detect these classes can differ from those designed specifically to work with speakers. In this thesis we focus on techniques to segregate different speakers from an audio signal, and other modules as speech/non-speech detection or acoustic event detection are not studied.

Although there are many different approaches to perform speaker diarization, most of them follow the scheme represented in Figure 2.1. The functionality of every block is explained next:

- Feature Extraction: Firstly, a set of features are obtained from the audio signal. The set of features obtained by the feature extractor should look for a representation of the information in the audio signal where the speakers are easy to separate.

- Segmentation: Once the features are obtained, the signal is segmented into acoustically homogeneous regions. The task of segmentation aims at detecting the boundaries between different and unknown acoustic sources or classes, in the continuous observations of speech. This is done under the assumption that the sources are discrete and every one is acoustically homogeneous, which is reasonable given that the classes are different speakers.

- Clustering: the task of speaker clustering aims at grouping the set of acoustically homogeneous segments obtained as output of the segmentation task into a discrete set of priorly unknown classes, which correspond to different speakers. Every segment is then labeled with a cluster identifier that refers to a unique speaker. The set of time marks provided by the segmentation and the set of speaker labels provided by the clustering compose the output of the complete speaker diarization system.

The information of the different classes present in the audio signal that the clustering stage provides can be used to refine the boundaries between contiguous segments. This is represented by the feedback path shown in Fig. 2.1, Then the refined segmentation can be fed into the clustering stage again, iteratively, until the desired solution is obtained. Note that, in general, segmentation and clustering could be inverted in the figure: a diarization system could perform clustering on the features directly and then segmenting the audio signal according to the set of classes obtained by the clustering. However, in any case, both stages are needed.

2.2 Features for Speaker Diarization

A speaker diarization system needs to segregate the fragments of an audio signal that belong to different speakers. To do so, the features obtained from the audio signal should extract information that enable the system to separate the speakers present in the signal. Therefore, those features suitable for speaker characterization will be useful for speaker diarization. The most popular acoustic features for speaker diarization are those used also for speech and speaker recognition: Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Predictive (PLP), Linear Predictive Coding (LPC) and others.

The mentioned parametrizations are known to obtain good performance in state-of-the-art speaker diarization systems and also in speaker recognition systems, but those features are not designed to capture relevant information for speaker characterization. Actually, these features are designed according to a human audio perception model (MFCC, PLP) or according to a human speech production model (LPC). Probably the most used features for both speaker diarization and recognition are the MFCC and PLP. Since these features have been designed according to human audio perception and humans are capable to perform speaker diarization and recognition they seem reasonable for these tasks. However, the community have always borne in mind the fact that these features are also good for speech recognition (actually they were firstly used for this purpose), which is a task that needs features that represent phonetic information as independently of the speaker as possible.

Thus, it does not seem reasonable to use the same features for both speech recognition and speaker characterization. Following this thought some works have explored other features for speaker diarization and recognition. In [Yamaguchi *et al.*, 2005] a set of features composed by energy, pitch frequency, peak-frequency centroid, peak-frequency bandwidth, temporal feature stability of the power spectra, spectral shape and white noise similarities is used for segmenting the audio into different classes, including speech, silence, noise and crosstalk. This was performed on several signals obtained from different speakers that were using pin tie microphones, where crosstalk was present. Then the same features are considered for speaker identification on the obtained speech segments. In [Huang and Hansen, 2006], three features including Perceptual Minimum Variance Distortionless Response (PMVDR), Smoothed Zero Crossing Rate (SZCR) and Filter-Bank Linear Coefficients (FBLC) are analyzed for a speaker segmentation task, showing that the first and the last outperforms traditional MFCC in noisy conditions. However, the improvement is not significant enough to replace the MFCC features, which, depending con the application, can be reused in posterior stages as in a speaker recognition or an ASR system.

Prosodic and long term features have been suggested to be useful for speaker characterization [Shriberg *et al.*, 2005]. In [Friedland *et al.*, 2009b] a total of seventy features including prosodic and long-term features are studied, and it is shown that certain statistics extracted from pitch and format estimations or from long-term spectrum estimations can be combined with the traditional MFCC improving the performance of a speaker diarization system. Again the improvement is not significant enough to consider these features in most applications.

Feature normalization techniques have been also studied in order to mitigate the influence of background noises and channel variability. Feature warping [Pelecanos and Sridharan, 2001], a technique that applies a non-linear transformation to the features over a sliding window so that the features will follow a Gaussian PDF, was utilized in [Sinha *et al.*, 2005] and [Zhu *et al.*, 2006] successfully. However, it is arguable that this type of normalizations is always useful for speaker diarization since part of the information they remove may help to characterize a speaker even if such information is channel information: for example, when several speakers are recorded in a room using a far field microphone, the channel information removed by feature warping may help to separate the speakers as far as they keep static. In [Kenny *et al.*, 2010], it is shown that the performance of a speaker diarization system for telephone conversations is better using unnormalized features than normalized features.

Finally, other features not directly related to acoustic parameters of the speakers present in a given audio signal have shown to help speaker diarization under certain conditions. In environments where more than one microphone is available to capture the audio signal, the time-delays between microphones, which are related to the position of the speakers, have shown to improve speaker diarization performance, as far as the speakers remain static [Pardo *et al.*, 2007]. In those situations when video information is also available, video features can help to determine the active speaker every moment, as in [Friedland *et al.*, 2009a], where MFCC features are combined with compressed domain video features to improve the performance of a speaker diarization system.

2.3 Speaker Segmentation

Speaker segmentation is the task of finding the boundaries between the different speaker turns present in an audio signal. This task is also known as speaker change detection or speaker turn detection. This task is related to the more general task of audio segmentation, that aims at detecting the boundaries between different acoustic sources in an audio signal, be they speakers or any other source. Audio segmentation includes speaker segmentation, but also acoustic event detection, speech/non-speech segmentation or acoustic environment change detection (background noises, music...). This work is focused on speaker segmentation rather than audio segmentation.

Speaker segmentation systems can be classified according to several criteria. In [Anguera, 2006], speaker segmentation systems are classified into two groups depending on the number of passes they perform on the data: those that perform a single pass to segment the data and those that perform more that one pass. In [Chen and Gopalakrishnan, 1998] the different methods used to perform speaker segmentation are classified into three groups:

- Metric-based: A distance is defined and computed between two neighboring windows whose boundary is placed sample by sample. The local maxima are labeled as hypothetical speaker boundaries, and compared to a threshold to determine whether or not they are speaker changes.

- Model-based: When there is data available to build statistical models for every speaker involved in the audio recording, such models are built and the most likely speaker changes are detected, according to a maximum likelihood criterion.

- Silence-based: Silence-based methods assume that prior speech and non-speech models are available. a speech/non-speech segmentation is performed, and the silences are labeled as possible speaker changes.

A more general classification of speaker segmentation systems can be done according to the assumption of availability of prior information of the classes involved. If it is assumed that prior information of the classes involved in the audio signal is available, model-based methods can be used. If there is no assumption of any prior knowledge, metric-based speaker segmentation needs to be performed. Silence-based speaker segmentation methods make use of a previous speech/non-speech segmentation. Depending on how this speech/non-speech segmentation is performed, these methods may or may not need prior speech and non-speech models. Most state-of-the-art diarization systems [Reynolds and Torres-Carrasquillo, 2005] combines different types of speaker segmentation: usually a first metric-based speaker segmentation method is performed to roughly detect the speaker changes. Then, after a clustering stage, and after obtaining a set of hypothetical speaker models, a model-based segmentation method is considered to refine the speaker boundaries.

2.3.1 Metric-based Speaker Segmentation

Metric-based techniques are among the most popular methods for speaker segmentation. This is probably due to the fact that they do not make use of any prior information of the classes present in the audio signal, so they are suitable for unsupervised tasks such as speaker diarization.

Metric-based methods compute a distance between two contiguous speech segments in order to determine whether or not there is a speaker change between them. Let χ_i, χ_j be two sequences of features of length N_i, N_j extracted for each one of the two contiguous segments i, j under test. Let $\chi_{i,j} = \{\chi_i \cup \chi_j\}$ be the stream composed of the features from both segments. Two different hypotheses can be tested: H_1 or the *null hypothesis* states that both segments belong to the same speaker, and H_2 states that it exists a speaker boundary between segments i and j. The goal of metric-based methods is to obtain a distance or metric $D(i, j)$ between segments i and j, in order to determine the correct hypothesis between H_1, H_2. $D(i, j)$ usually takes into account the differences between the acoustic sequences χ_i and χ_j and the homogeneity of each one of these sequences compared to the homogeneity of the whole stream $\chi_{i,j}$. Usually, the obtained distance is compared to a threshold ϵ in order to select one of the hypothesis.

$$D(i, j) \underset{H_1}{\overset{H_2}{\gtrless}} \epsilon \tag{2.1}$$

Most metric-based approaches for speaker segmentation work on a sliding window of fixed or variable length, following the procedure represented in Figure 2.2. The window w is supposed to contain only two speakers or less (step 1 in Fig. 2.2). Within the window, several hypothetical boundaries b between different speakers are tested (step 2 in Fig. 2.2). Every boundary b splits the window into two sequences of features that depend on b: $\chi_{i_w(b)}, \chi_{j_w(b)}$. For a given window w, the distance $D(i_w(b), j_w(b))$ can be expressed as a function of the hypothetical boundary b within w: $D_w(b)$, and only the boundary b_c obtaining the local maximum value of the distance $D_w(b)$ is considered as candidate for a possible speaker change (see step 3 in Fig. 2.2):

$$b_c = \underset{b}{argmax}\, D_w(b) \tag{2.2}$$

In order to determine whether or not the candidate boundary b_c is a speaker change, Both possible hypotheses H_1, H_2 are evaluated for b_c (step 4 in Fig. 2.2). The evaluation is performed by comparing distance $D_w(b_c)$ to the threshold ϵ. In case b_c is a speaker change,

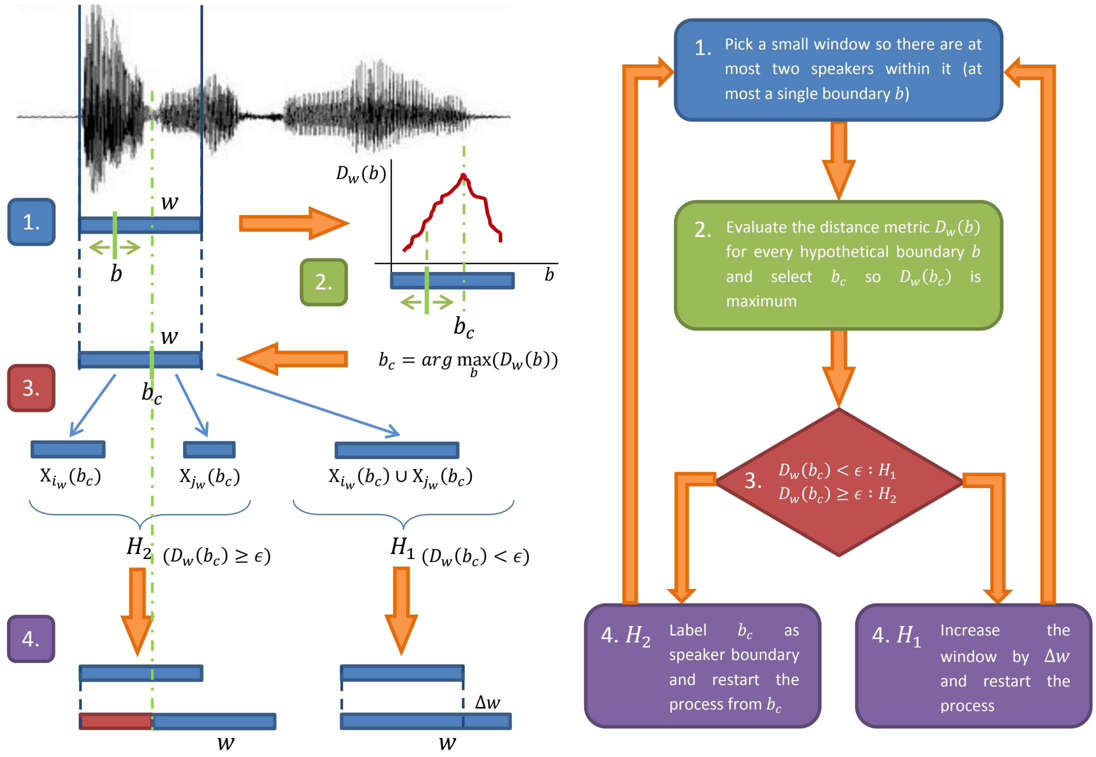

1.
w
b
$D_w(b)$
b
2.
b_c
$b_c = arg\max_b(D_w(b))$
w
3.
b_c
$X_{i_w}(b_c)$
$X_{j_w}(b_c)$
$X_{i_w}(b_c) \cup X_{j_w}(b_c)$
H_2 $(D_w(b_c) \geq \epsilon)$
H_1 $(D_w(b_c) < \epsilon)$
4.
w
Δw
w
1. Pick a small window so there are at most two speakers within it (at most a single boundary b)
2. Evaluate the distance metric $D_w(b)$ for every hypothetical boundary b and select b_c so $D_w(b_c)$ is maximum
3. $D_w(b_c) < \epsilon : H_1$
$D_w(b_c) \geq \epsilon : H_2$
4. H_2 Label b_c as speaker boundary and restart the process from b_c
4. H_1 Increase the window by Δw and restart the process

Figure 2.2: *Scheme of a metric based speaker segmentation system.*

usually the sliding window w is resized to its original size and moved to start processing the audio signal from b_c. In case b_c is not a speaker change, w is expanded by Δw or advanced in the audio signal in order to look for possible boundaries considering new data. This process was first described in [Chen and Gopalakrishnan, 1998] and several variations of this approach have been used in the literature [Delacourt and Wellekens, 2000], [Zhou and Hansen, 2005].

The following subsections describes the most popular distance metrics used in the literature.

- **Bayesian Information Criterion (BIC)**: BIC was firstly proposed to select the model that best explained certain data in [Schwarz, 1978]. Given a set of N samples χ drawn from a random process, and a hypothetical candidate model Θ that describes the available data, BIC is a metric that represent the degree of fitness of Θ to χ, taking into account the complexity of Θ. BIC is computed as follows:

$$BIC(\Theta) = log(\mathcal{L}(\chi|\Theta)) - \lambda\frac{1}{2}\#(\Theta) \times log(N), \tag{2.3}$$

  Where the first term $log(\mathcal{L}(\chi|\Theta))$ is the log-likelihood of the data given the model and explains the goodness of fit of the model to the data, while the second term $\lambda\frac{1}{2}\#(\Theta) \times log(N)$ penalizes the likelihood taking into account the complexity of the model and is usually referred as the complexity penalty. The complexity penalty is introduced to penalize more complex models, since they are expected to obtain higher likelihood even if they do not describe the data distribution properly. The complexity penalty is proportional to $\#(\Theta)$, which is the number of free parameters to estimate the model Θ, to $log(N)$, and to λ, which is a free parameter to adjust the penalty depending on the application.

When more that one model Θ is proposed to describe the available data χ, BIC can be used to select the simplest model that successfully describes the data. This is the main application of BIC: a model selection criterion. The model obtaining higher BIC will be selected to explain the data.

BIC can be used to evaluate whether a change point occurs between two segments i and j. For this purpose two BIC values are computed. One BIC value for H_1 ($BIC(H_1)$), assuming that the data from the two segments $\chi_{i,j} = \chi_i \cup \chi_j$ can be described by a single model $\Theta_{i,j}$. Another BIC value is computed for H_2 ($BIC(H_2)$), assuming that the data from every segment is explained by a different model. Given that the complexity of a model for a given segment is fixed, the model $\Theta_{i,j}$ for H_1 is simpler, but using different models Θ_i and Θ_j to describe the segments i and j will yield a better fitness, i.e. will increase the likelihood of the data given the models. To obtain a distance metric from both values, the difference between both BIC values is computed:

$$\Delta BIC = BIC(H_2) - BIC(H_1) = R(i,j) - \lambda P, \tag{2.4}$$

where $R(i,j)$ is the difference between the log-likelihoods obtained for every hypothesis, and P is the excess complexity penalty of H_2 with respect to H_1

Usually, every segment is assumed to be modeled by a Gaussian distribution with full covariance, so $\chi_i \sim \mathcal{N}(\mu_i, \Sigma_i)$. With this assumption, the $R(i,j)$ term can be expressed as:

$$R(i,j) = \frac{N}{2} log(|\Sigma_{i,j}|) - \frac{N_i}{2} log(|\Sigma_i|) - \frac{N_j}{2} log(|\Sigma_j|), \tag{2.5}$$

while the excess complexity penalty can be expressed as:

$$P = \frac{1}{2}(p + \frac{1}{2}p(p+1))log(N), \tag{2.6}$$

GMMs are also usually considered to model every segment. When such models are used, there is no simplification in the Likelihood computation, so $R(i,j)$ must be computed as a log-likelihood ratio between H_1 and H_2 while P will be a function of the number of free parameters that the considered GMM have.

The ΔBIC value can be used as distance metric in the algorithm presented in Fig. 2.2. Actually, the use of this ΔBIC metric for speaker segmentation was first proposed along with the algorithm described in Fig. 2.2 by Chen and Gopalakrishnan in [Chen and Gopalakrishnan, 1998]. Since then, ΔBIC has become the most popular metric for speaker segmentation and also for speaker clustering.

One of the problems of the ΔBIC is the presence of the λ parameter in its formulation. The λ parameter was added to adjust the complexity penalty, but its presence introduces a hidden threshold in the ΔBIC. In [Ajmera and Wooters, 2003], the excess complexity penalty term is removed by considering a model $\Theta_{i,j}$ for H_1 with a number of free parameters equal to the sum of the number of free parameters of the models Θ_i, Θ_j estimated for H_2.

- **Generalized Likelihood Ratio (GLR)**: Given the sequences of features χ_i, χ_j from two segments i and j, GLR is obtained as a likelihood ratio between the likelihood for the given feature sequences computed under the assumption that both segments belong to the same speaker (H_1), and the likelihood computed under the assumption that both segments belong to different speakers (H_2). GLR differs from the standard Likelihood Ratio (LR) in that the probability density function PDF for every hypothetical speaker must be estimated from the data available in i and j when computing the GLR, while for LR, the PDF for every hypothetical speaker is priorly known.

  Therefore, the available data is modeled in a different way for each one of the hypothesis. Assuming that the model of every hypothetical speaker k within every hypothesis is a PDF whose parameters are given by Θ_k, for H_1, both segments are assumed to belong to a single speaker and $\Theta_{i,j}$ is estimated with all the data $\chi_{i,j}$, while for H_2, every segment is assumed to belong to a different speaker, and two models are estimated: Θ_i from χ_i and Θ_j from χ_j. The GLR is then computed as follows:

$$GLR(\frac{H_1}{H_2}) = \frac{\mathcal{L}(\chi_{i,j}|\Theta_{i,j})}{\mathcal{L}(\chi_i|\Theta_i)\mathcal{L}(\chi_j|\Theta_j)}, \tag{2.7}$$

  where $\mathcal{L}$ denotes likelihood. Usually, the acoustic observations χ_k that are assumed to belong to a single speaker k are modeled with a Gaussian PDF or a GMM. The distance is obtained as the log of the GLR, $D(i,j) = -log(GLR(\frac{H_1}{H_2}))$.

  GLR was the first proposed metric for change detection in [Willsky and Jones, 1976], but it was not used for acoustic change detection until much later. In [Delacourt and Wellekens, 2000], speaker segmentation is performed in two steps, first using GLR smoothed by a low-pass filter and then using BIC. In [Bonastre *et al.*, 2000], the GLR is used as a single step to perform speaker segmentation in a speaker tracking task. In [Liu and Kubala, 1999] the GLR is penalized depending on the available data in segments i and j.

- **Kullback-Leibler Divergence (KL)**: Given two distributions P and Q, their KL divergence is defined as:

$$KL(P||Q) = \mathbb{E}_x[log\frac{p(x)}{q(x)}], \tag{2.8}$$

  where p and q denote the densities of P and Q. The KL Divergence measures the expected number of extra bits required to code samples from P using a code based on Q.

  Although the KL divergence is related to the difference between two distributions, is not a distance metric strictly speaking, since KL divergence is not symmetric. To overcome this problem the symmetrized KL or KL2 distance is considered. KL2 is defined as follows:

$$KL2(P||Q) = KL(P||Q) + KL(Q||P), \tag{2.9}$$

The KL2 divergence can be easily applied to the task of speaker segmentation, considering the framework presented in Fig. 2.2. For a given window w, considering the hypothesis H_1 of having a speaker boundary b within w, a model is built by estimating a PDF on both segments χ_i, χ_j. The KL2 divergence computed for the obtained PDFs can be used as a distance between both segments. Usually, Gaussian PDFs are considered, since there is no close form solution of the KL2 for GMMs, but the KL2 divergence can be easily approximated when both GMMs are adapted from the same background GMM, for example by means of MAP [Do, 2003].

The first study that considered the KL2 distance for speaker segmentation was presented in [Siegler *et al.*, 1997]. In this work, the KL2 distance is used for acoustic segmentation, and also for speaker segmentation in Broadcast News environments. Later works have considered this measure for its fast computation as part of a multiple step speaker segmentation system, usually to obtain an initial rough segmentation that is later refined using other metrics [Delacourt and Wellekens, 2000].

- **Other distance metrics**: Many other distance metrics have been presented in the literature for the task of speaker segmentation. In [Hung *et al.*, 2000], the Mahalanobis or Bhattacharyya distances are proposed for this task, and compared to the KL2 divergence. In [Kemp *et al.*, 2000], the entropy loss of coding the data in two segments instead of only one is proposed in comparison to the KL2 divergence. In [Zhou and Hansen, 2000], the Hotelling's T^2 distance is proposed to perform speaker segmentation when the analysis window w is small, and BIC is considered as the window w becomes larger.

  Other proposed metrics are modifications of the well known distance metrics previously described. Some examples in the literature are the Gish distance [Gish *et al.*, 1991] or the Cross Likelihood Ratio (CLR) [Barras *et al.*, 2006], based on the GLR, the Cross-BIC [Anguera, 2005], based on the BIC, or the Divergence Shape Distance (DSD) [Lu and Zhang, 2002] based on the KL2 divergence.

2.3.2 Model-based and Silence-based Speaker Segmentation

Metric-based methods for speaker segmentation assume no prior knowledge of the classes present in the audio signal to process. However, if prior information is available, every class can be modeled and audio or speaker segmentation can be performed by means of a decoding process. The techniques that make use of prior models to perform speaker segmentation are known as model-based techniques.

Model-based techniques make use of a close set of models to classify the acoustic observations extracted from the audio signal into the desired classes. These classes may be the speakers present in the audio signal, but also more general classes that also can give information of speaker changes as male-female or telephone-wideband. Usually, GMMs are considered to model every class and the acoustic observations are classified according a Maximum Likelihood (ML) criterion or using Viterbi decoding [Gauvain *et al.*, 1999], [Kemp *et al.*, 2000], [Kubala *et al.*, 1996].

Other approaches make use of the given audio signal to train the speaker models, considering an initial segmentation and refining iteratively the boundaries between the different speakers. When the number of speakers is known (telephone conversations) or after it has been estimated (clustering), this process is known as re-segmentation, since its goal is simply to refine the speaker boundaries [Reynolds and Torres-Carrasquillo, 2005]. In those environments where the number of speakers is unknown, model-based segmentation is usually combined with clustering techniques to iteratively reestimate the number of speakers and refine the boundaries among them [Ajmera *et al.*, 2002].

On the other hand, if a robust speech/non-speech segmentation is available, the silence segments can be considered as candidate speaker boundaries. The speaker segmentation techniques based in this principle are known as silence-based techniques. Some works in the literature make use of an energy detector to find the silence segments [Kemp *et al.*, 2000]. In [Huang and Hansen, 2006], several features for speech/non-speech segmentation are presented. Other works make use of a complete speech recognition system in order to find the silence segments [Kubala *et al.*, 1996]. When a speech recognition system is considered for this purpose, the segmentation system is known as decoder-guided.

Silence-based speaker segmentation systems are not the most popular for this task since most silence segments do not correspond to a boundary between speakers, so additional techniques are needed in order to determine whether or not a silence segment is also a speaker change. In addition, there may be speaker boundaries not associated with a silence segment, specially when overlapped speech is present.

2.4 Speaker Clustering

Speaker clustering is the task of grouping a set of audio segments into a discrete set of priorly unknown classes, which correspond to different speakers. Traditionally, this task has been associated to speaker diarization, and in this framework, speaker clustering techniques operate over the audio fragments obtained after speaker segmentation. However, speaker clustering involves a more general task, since the audio segments to be clustered do not need to belong to the same recording. This is the case of the recent works presented in [Brummer and De Villiers, 2010], [van Leeuwen, 2010], where different recordings containing a single speaker are considered as input for speaker clustering. In fact, in [Brummer and De Villiers, 2010], the speaker clustering task is referred as the speaker partitioning problem, and it is presented as a generalization of the speaker detection/verification problem.

In this section the speaker clustering or speaker partitioning problem is presented, and the most popular approaches to solve this problem as part or apart from a speaker diarization system are presented.

2.4.1 The Speaker Partitioning Problem

The speaker detection/verification problem has as input $N = 2$ speech segments $X = \{\chi_1, \chi_2\}$, and there are $K = 2$ possible hypotheses $\{H_1, H_2\}$. H_1 is the null hypothesis or target hypothesis, and it states that both segments belong to the same speaker, while H_2 is the non-target hypothesis, and it states that the segments belong to different speakers.

The speaker clustering problem can be seen as a generalization of the speaker detection/verification problem where a set Ω of $N \geq 2$ speech segments $X = \{\chi_1, \chi_2..., \chi_N\}$

is available as input, and the partition of the set that clusters those speech segments belonging to the same speaker together is desired. The desired partition is unique, and must be selected among all possible hypothetical partitions $H_1, H_2, ..., H_{B_N}$, from the coarsest partition H_1 that assumes that all segments belong to the same speaker, to the finest partition H_{B_N} that assumes that every segment contains a unique speaker. B_N denotes the N^{th} Bell number, that is in fact defined as the number of partitions for a set of N members.

To solve this problem, we assume that it exist a framework, for example a generative model Θ, that enables us to obtain a score, or following the proposed example, a likelihood, for every hypothetical partition. Every hypothetical partition H_k is composed of S non-overlapping clusters $\mathcal{C}_1(k), \mathcal{C}_2(k), ..., \mathcal{C}_S(k)$, which together contains all elements of Ω. This way, the likelihood for H_k is defined as:

$$\mathcal{L}(H_k) = \prod_{j=1}^{S} \mathcal{L}(\mathcal{C}_j(k)) \propto \prod_{j=1}^{S} P(\mathcal{C}_j(k)|\Omega, \Theta), \tag{2.10}$$

where it is assumed that the model Θ has fixed, known parameters and also has suitable independence assumptions so the likelihood for a hypothetical partition H_k is the product of the likelihoods for the non-overlapping subsets of H_k.

If it is also assumed that the likelihoods obtained in this framework enable us to compare hypothetical partitions in order to select the most likely one as the solution of the speaker partitioning problem, then the problem can be solved computing the likelihoods for all B_N partitions and selecting the partition obtaining the highest likelihood.

This solution is optimal in the sense that all partitions are evaluated and the most likely one is selected. However, it is not feasible in most real cases, since the number of hypothetical partitions B_N for a dataset increases dramatically as the number of segments in the dataset N increases. For example, to solve a speaker clustering problem that has three segments as input, five hypothetical partitions need to be evaluated, but for a set containing ten segments, which is a small number of segments for most diarization or speaker recognition problems that make use of speaker clustering, a total of 115975 hypothetical partitions need to be evaluated.

In order to avoid the computation of all possible hypothetical partitions for a given set of segments, suboptimal methods need to be introduced. The most popular suboptimal technique for speaker clustering is known as Hierarchical Clustering. Hierarchical Clustering is a greedy approach that reduces the number of hypothetical partitions by making locally optimal choices, so that the solution to the speaker clustering problem is feasible.

2.4.2 Hierarchical Clustering

Hierarchical clustering techniques starts from a given partition of the available set of speech segments (usually the coarsest one or the finest one), and the clusters are iteratively split or merged until the optimum number of speakers is reached. Hierarchical clustering approaches can be classified into two main groups:

- **Bottom-up**: Bottom-up Hierarchical Clustering or Agglomerative Hierarchical Clustering (AHC) methods start from a large number of clusters or speech segments (the finest partition) and merge the closest segments iteratively until a stopping criterion is met. These techniques are the most used for speaker clustering in the field of speaker diarization, since it is straightforward to apply them on the output

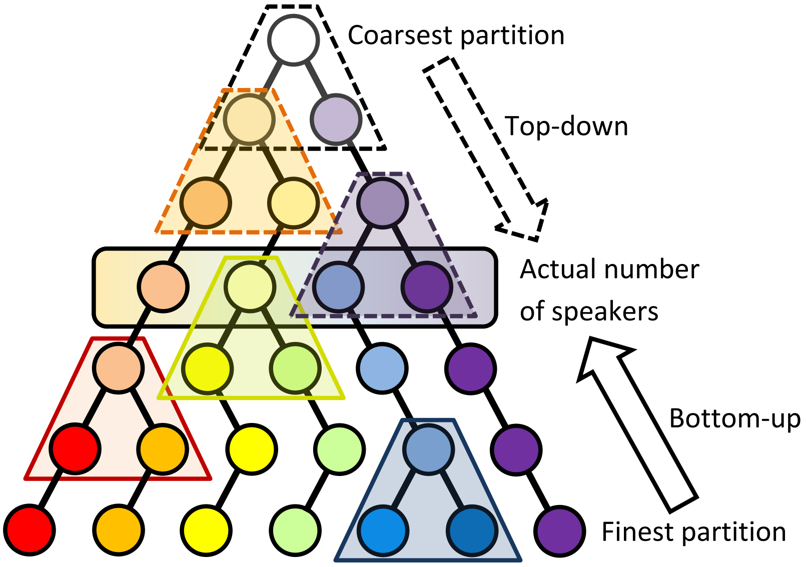


Figure 2.3: *Bottom-up and Top-down hierarchical clustering strategies.*

set of segments obtained from a speaker segmentation system. Usually, a matrix of distances between every possible pair of clusters is computed. Then, the closest pair is merged, the merged clusters are removed from the distance matrix, and the matrix is updated with the distances between the new merged cluster and all remaining clusters. This process is done iteratively until the stopping criterion is met. Alternatively, the process can be done until a single cluster is obtained, building a binary tree, and then the stopping criterion will determine the optimal level of the binary tree. This approach for clustering has been used for many years in pattern classification [Duda and Hart, 1973], but is was first considered for speaker clustering in [Jin *et al.*, 1997] and [Siegler *et al.*, 1997].

- **Top-down**: Top-down Hierarchical Clustering methods start from a small number of clusters (usually a single cluster, the coarsest partition) containing several speech segments, and the initial clusters are split iteratively until a stopping criterion is met. There are fewer systems that make use of top-down clustering methods than systems that make use of bottom-up methods. Some examples of top-down speaker clustering strategies can be found in the literature in [Johnson and Woodland, 1998], [Reynolds and Torres-Carrasquillo, 2005]

Figure 2.3 represents the two most common ways of performing hierarchical clustering. Usually, bottom-up clustering techniques analyzes every possible pair of clusters, merging only the closest pair on every iteration, while top-down clustering techniques analyzes every cluster and splits a single cluster on every iteration. Both techniques make the decision of merging or splitting the clusters locally, expecting that the final solution reached will be the global optimum, but this is not guaranteed.

In order to design hierarchical clustering methods two issues must be taken into account:

- **Distance metric of acoustic similarity**: A metric of acoustic similarity is needed to decide whether or not two clusters must be merged (bottom-up clustering) or split (top-down clustering). The same metrics previously presented for speaker segmentation can be considered.

- **Stopping criterion**: A stopping criterion is needed to determine when the optimal number of clusters (speakers) has been reached.

In the following sections, the most common distance metrics and stopping criteria for hierarchical speaker clustering presented in the literature are described.

2.4.2.1 Distance Metrics for Hierarchical Speaker Clustering

Distance metrics of acoustic similarity for speaker clustering aims at determining whether two clusters belong to the same speaker or to different speakers. Therefore, the distance metrics for this task have the same objective as in the task of speaker segmentation when metric-based techniques are considered. In fact, most of the metrics proposed for speaker segmentation have been also considered for hierarchical speaker clustering. As an example, we can find works where the Gish distance [Gish *et al.*, 1991] is considered for this task [Jin *et al.*, 1997]. In [Siegler *et al.*, 1997] the KL2 distance is compared to the Mahalanobis distance showing that KL2 is more accurate for speaker clustering. The KL2 distance is also considered in [Zhou and Hansen, 2000]. In other works ([Rougui *et al.*, 2006] and [Ben *et al.*, 2004]) every cluster is modeled with a GMM in order to take advantage of the amount of data present in clusters containing several speech segments, and approximations of the KL distance are used to compute distances between GMMs.

The GLR has been also very popular for the task of speaker clustering. It was firstly used for the clustering of speech segments into two known classes in [Siu *et al.*, 1992], and has been also considered as distance metric for hierarchical clustering in [Gauvain *et al.*, 1999], [Solomonoff *et al.*, 1998] and [Barras *et al.*, 2004]. But the most popular metric for this task is the ΔBIC [Chen and Gopalakrishnan, 1998], as for the the task of speaker segmentation. The works presented in [Chen *et al.*, 2002] and [Ajmera and Wooters, 2003] make use of BIC as distance metric for hierarchical speaker clustering.

Other works have proposed distance metrics based on speaker identification techniques. In [Barras *et al.*, 2004] and [Zhu *et al.*, 2005], MAP adaptation is considered to train cluster models from a Universal Background Model (UBM). Then, as distance metric, it is considered the cross likelihood distance (CLR) [Reynolds *et al.*, 1998], defined as:

$$D(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{N_i} log \frac{P(\chi_i|\Theta_j)}{P(\chi_i|\Theta_{UBM})} + \frac{1}{N_j} log \frac{P(\chi_j|\Theta_i)}{P(\chi_j|\Theta_{UBM})}, \tag{2.11}$$

Where $\mathcal{C}_i$ denotes the cluster i, χ_i is the sequence of acoustic observations obtained for all the speech segments belonging to cluster i, N_i is the number of acoustic observations in cluster i, Θ_i is the MAP adapted model for $\mathcal{C}_i$ and Θ_{UBM} is the UBM model.

In [van Leeuwen, 2010], the score of a speaker verification system is directly considered as distance metric for the task of speaker clustering. In this case, clustering is performed over a dataset containing different recording sessions, each of them containing a single speaker. The speaker verification technique utilized is known as GMM-*Support Vector Machine* (GMM-SVM) with Nuissance Attribute Projection (NAP) for channel compensation [Campbell *et al.*, 2006], but the proposed approach can make use of any other speaker verification system to obtain a distance metric for hierarchical speaker clustering.

2.4.2.2 Stopping Criteria for Hierarchical Speaker Clustering

One of the most critical aspects of a hierarchical speaker clustering system is to determine the actual number of speakers present over all the input speech segments. Since hierarchical clustering approaches keeps reducing (bottom-up) or increasing (top-down) the number of clusters iteratively, the task of determining the number of speakers is reduced to knowing when to stop the iterative process. Note that in some tasks the actual number of speakers is priorly known, so the stopping criterion reduces to reaching the known number of speakers. In this section we study stopping criteria for hierarchical clustering approaches assuming that the number of speakers is unknown.

The most widespread stopping criterion for speaker clustering is the use of a threshold in the distance metric considered for the task of speaker clustering. The threshold is usually obtained experimentally, so this approach may not be very robust against training and testing mismatch. For example, in [Gauvain *et al.*, 1999], the GLR is used as distance metric and compared to a threshold to determine the actual number of speakers. In [Barras *et al.*, 2004] and [Zhou and Hansen, 2005], the cross likelihood distance is considered as stopping criterion as well as a distance metric for the hierarchical clustering process. Other methods considered as stopping criterion for hierarchical speaker clustering include the minimization of the estimated cluster and speaker purity, as in [Solomonoff *et al.*, 1998].

Again, the most popular stopping criterion is the ΔBIC distance, specially for bottom-up clustering systems. These systems usually finish the clustering process when all pairs obtain $\Delta BIC < 0$. This stopping criterion is considered in [Chen and Gopalakrishnan, 1998], or [Chen *et al.*, 2002]. The main problem of this stopping criterion is that, although the threshold is fixed, the λ parameter, that adjust the penalty term, must be tuned on a training dataset experimentally, and thus a hidden threshold is introduced. To avoid this effect, the ΔBIC can be estimated as proposed in [Ajmera and Wooters, 2003].

Other measure that have been proposed recently as stopping criterion for speaker clustering is the student´s t-test [Nguyen *et al.*, 1998]. In this work, the populations for intra-cluster and inter-cluster distances are obtained for a given partition and the student´s t-test is obtained for these two populations. The partition that maximizes the t-test is selected. The advantage of this approach is that it does not need any development data to set a threshold.

When the score obtained from a speaker verification system is considered for speaker clustering, the most straightforward stopping criterion is again to set a threshold for the obtained scores. This threshold can be set experimentally as in [van Leeuwen, 2010], or through a calibration process, as it is usual in speaker verification tasks [Brummer and Dupreez, 2006].

2.4.3 Other Approaches for Speaker Clustering

In this section we present another approaches to solve the partitioning problem that do not make use of hierarchical clustering. Among these techniques, one of growing interest in the last years is the Variational Bayesian (VB) learning [Attias, 2000], [Bishop, 2006]. The VB framework enables us to learn the model parameters and adjust the complexity of the model depending on the given amount of training data within a single algorithm. The first to apply VB learning to the task of speaker clustering was F. Valente [Valente and Wellekens, 2004]. Recently, the VB framework has been used for speaker clustering in telephone conversations

[Kenny *et al.*, 2010], a task where the number of speakers is previously known and limited to two. The proposed approach can be easily expanded to a higher number of speakers or to a unknown number of speakers.

When the number of speakers is known, well known clustering techniques as Vector Quantization or K-means can be used. In [Lapidot, 2003] Self-Organizing Maps (SOM) [Lapidot *et al.*, 2002] are proposed for speaker clustering. This technique uses a VQ algorithm for training the code-books representing each one of the speakers. In [Castaldo *et al.*, 2008] and [Vaquero *et al.*, 2010a] the Joint Factor Analysis (JFA) paradigm for speaker recognition [Kenny *et al.*, 2008] is considered to extract compact representations of the speaker present over small segments, which are known as speaker factors. In [Vaquero *et al.*, 2010a], these speaker factors are clustered using PCA and K-means.

Finally, an optimal solution can be obtained for the the speaker partitioning problem. In [Brummer and De Villiers, 2010], the i-vector paradigm [Dehak *et al.*, 2010] is considered to find an optimal solution for the partitioning problem, using linear Gaussian models [Bishop, 2006] to represent between-speaker and within-speaker variability. This approach for modeling the i-vectors is known as the *two covariance model.* Results are shown considering two input segments (speaker verification) and three input segments. The problem of this solution is that the number of input segments cannot be high since the number of partitions to evaluate will increase dramatically.

2.5 Speaker Diarization Systems

In the previous sections, the most common techniques for speaker segmentation and speaker clustering have been introduced. Most systems combine the techniques previously described to perform speaker diarization. In this section, the state-of-the-art speaker diarization systems are presented, detailing how the problems of speaker segmentation and clustering are solved.

Recent research effort in speaker diarization has been mainly focused on the meeting environment, partially motivated by the European Projects AMI and CHIL, and the last NIST RT evaluations. The meeting environment has encouraged new research lines such as the use of multiple microphones to extract features related to the speaker position [Pardo *et al.*, 2007], or the use of multimedia information, concretely video information [Friedland *et al.*, 2009a], in order to improve speaker diarization. However, this work is focused on single microphone environments, where only acoustic information is available, and speaker diarization systems in these conditions have not evolved remarkably in the framework of meetings during the last years. Most research effort has been focused on the adjustment of the design parameters in order to provide robustness to the well-known speaker diarization systems.

Traditional speaker diarization systems for meeting environments inherit most techniques from those developed for the Broadcast News environment. A good compilation of traditional diarization approaches used firstly for Broadcast News and later for meeting environments can be found in [Tranter and Reynolds, 2006]. Traditional diarization systems perform a first speaker segmentation pass with a metric-based approach, usually considering BIC. Then, the obtained clusters are merged iteratively using a bottom-up hierarchical clustering approach, also considering BIC in most cases both as distance metric and as stopping criterion. Usually, the clusters are modeled using a full covariance Gaussian

[Anguera, 2005], or GMMs trained from the available data in the cluster [Wooters *et al.*, 2004]. Some works build cluster models adapting the GMMs from a UBM using MAP adaptation [Meignier *et al.*, 2006]. A final re-segmentation pass is performed using model-based speaker segmentation approaches, usually building a Hidden Markov model (HMM) where every state represents a speaker whose observation distribution is given by the speaker GMM. The re-segmentation step can be performed every time two clusters are merged. The purpose of this re-segmentation is to refine the speaker boundaries.

Recent work in speaker diarization in meetings has been focused on providing robustness to these traditional systems. Some studies have shown the variability in performance that these approaches present depending on the initial segmentation [Imseng and Friedland, 2010], proposing novel initialization methods [Imseng and Friedland, 2010].

The Influence of overlapped speech in speaker diarization systems has also been studied. Some methods to model and identify the overlapped speech have been proposed in [Boakye *et al.*, 2008]. In this work, the detection of overlapped speech improves the performance of the speaker diarization system when the overlapped speech is removed in order to build speaker models, and also when the speakers present in overlapped speech segments are identified using GMM classifiers.

A recent approach for speaker diarization that is completely different to the traditional approaches is the one presented in [Fox *et al.*, 2009]. In this work, Hierarchical Dirichlet Processes (HDP) are considered in order to develop a Bayesian nonparametric approach to HMMs, in which the number of states is unknown a priori. This framework enables to perform automatic segmentation and clustering with unknown number of speakers.

On the other hand, the recent advances in the field of speaker recognition have encouraged new approaches for speaker diarization. In [Castaldo *et al.*, 2008], [Kenny *et al.*, 2010], [Vaquero *et al.*, 2010a] a set of speaker factors, which are compact speaker representations based on eigenvoice modeling and the JFA framework [Kenny *et al.*, 2008] for speaker verification, are extracted from short segments. Then the speaker factors are agglomerated using traditional clustering approaches [Vaquero *et al.*, 2010a], or VB [Kenny *et al.*, 2010]. Usually, a final re-segmentation considering MFCC features is performed to refine speaker boundaries, since the speaker factors have low temporal resolution. These techniques have shown to outperform the traditional approaches for speaker segmentation in telephone environments, when the number of speakers is limited to two.

2.6 Evaluation of Speaker Diarization

In this section the most common metrics utilized to evaluate the accuracy of speaker diarization systems are presented, as well as other metrics useful to evaluate speaker segmentation or clustering approaches.

Several metrics have been used to evaluate the task of speaker segmentation. In the earlier works, the most popular metrics have been the rate of missed and the rate of false alarm speaker changes [Chen and Gopalakrishnan, 1998], [Zhou and Hansen, 2005]. Also, the Equal Error Rate, defined as the miss rate or the false alarm rate when both error rates are equal has been reported as measure of segmentation performance.

The presented measures encourage speaker segmentation systems to detect as many speaker boundaries as possible, introducing low false alarms, which is the actual goal of the system, but they have some problems when considered as accuracy metrics for speaker segmentation. The first problem is the fact that the bias of the detected boundary with respect to the actual boundary is not taken into account. Some works [Chen and Gopalakrishnan, 1998] report the percentage of speaker changes biased more that a certain duration. In [Hansen *et al.*, 2005], this problem is overcome introducing a new metric called Fused Error Score (FES) where the miss and false alarm rate is weighted by the average mismatch of the detected boundaries measured in milliseconds. The second problem is that these measures consider that all speaker boundaries are equally important to detect, which is in general not true, specially if the output of the speaker segmentation system will be the input of an AHC algorithm. In fact, missing a speaker boundary between two long speaker turns will merge both speakers and the clustering process will not be able to separate them. On the other hand, missing the two boundaries that delimit a short speaker turn during the speech of another speaker will not be as critical as the first example.

To overcome the previous problem, some works [Gauvain *et al.*, 1999] present the percentage of time incorrectly classified or time error rate. The time error rate is simply the percentage of time that has been assigned to an incorrect class. This metric is only helpful to evaluate the accuracy of speaker segmentation when the number of classes is priorly known, or after the clustering process have been performed, since an assignation of every segment to a single class needs to be performed.

The task of speaker clustering have been traditionally evaluated using purity metrics, usually clustering and speaker purity. Their counterparts, cluster and speaker impurities, are very well described in [van Leeuwen, 2010]. Given a set of N segments Ω, that contains R different speakers with $R < N$, we define the (relative) frequency of the speaker r in the segment n as:

$$f_r(n) = \frac{L_r(n)}{L(n)}, \tag{2.12}$$

where $L_r(n)$ is the number of acoustic observations or frames of segment n that belong to the speaker r, and $L(n) = \sum_{r=1}^{R} L_r(n)$ is the total number of speech acoustic observations of segment n. Note that when a segment contains a single speaker i (there are no segmentation errors, or every segment is a different recording containing a single speaker), $f_r(n) = 1$ if $r = i$, and $f_r(n) = 0$ otherwise.

For a hypothetical speaker partition H that gives S clusters $\mathcal{C}_s, s = 1, ..., S$, we define the frequency of a speaker r in a cluster C_s as:

$$f_r(\mathcal{C}_s) = \frac{L_r(\mathcal{C}_s)}{L(\mathcal{C}_s)} = \frac{\sum_{n \in \mathcal{C}_s} f_r(n) L(n)}{\sum_{n \in C_s} L(n)}, \tag{2.13}$$

where $L_r(\mathcal{C}_s)$ is the number of frames of all the segments in cluster $\mathcal{C}_s$ that belong to the speaker r, and $L(\mathcal{C}_s)$ is the total number of speech frames in cluster $\mathcal{C}_s$. Note that the frequency of r in $\mathcal{C}_s$ can be obtained as the weighted average frequency of r for all segments in $\mathcal{C}_s$, where the weights are given by the number of frames. Usually, in speaker clustering problems that deals with a set of different recordings as segments, as the one expressed in [van Leeuwen, 2010], all recordings are equally weighted and $f_r(\mathcal{C}_s)$ reduces to the average of $f_r(n)$ for all $n \in \mathcal{C}_s$.

From this definition of $f_r(\mathcal{C}_s)$, the cluster purity of a single cluster $\mathcal{C}_s$ can be expressed as the frequency of the speaker r that obtains highest frequency in $\mathcal{C}_s$:

$$P_{cluster}(\mathcal{C}_s) = \underset{r}{max}(f_r(\mathcal{C}_s)). \tag{2.14}$$

And thus, the cluster purity for the whole set of segments Ω, given the hypothetical speaker partition H, is defined as the weighted average of the cluster purities for all clusters:

$$P_{cluster}(\Omega|H) = \frac{\sum_{s=1}^{S} L(\mathcal{C}_s) P_{cluster}(\mathcal{C}_s)}{L(\Omega)}, \tag{2.15}$$

where $L(\Omega) = \sum_{s=1}^{S} L(\mathcal{C}_s) = \sum_{n=1}^{N} L(n)$ is the total number of speech frames in the whole set of segments. In some cases all segments are assumed to have the same weight independently of the number of speech frames they contain. In these situations, the weight considered for every cluster is the number of segments that the cluster contains $N_{\mathcal{C}_s}$, and the normalization term is the total number of segments N.

The cluster purity can not measure the performance of a clustering task on its own. A high cluster purity indicates that for each cluster, most segments belong to a single speaker, i.e. there is a speaker that clearly dominates every cluster. However, it does not take into account the fact that a unique speaker may be fragmented and present in several clusters. For example, assuming that the segments contain a single speaker, the finest partition will always get $P_{cluster} = 100\%$, but it is not the solution of our speaker clustering problem.

To take this effect into account, the concept of speaker purity is introduced. In this case, we define the frequency of the segment n in a speaker r as:

$$g_n(r) = \frac{L_r(n)}{L_r(\Omega)}, \tag{2.16}$$

where $L_r(\Omega)$ is the number of frames that belong to speaker r in the whole set of segments. This frequency can be very low if the speaker r is present in several segments. From this definition, we can obtain the frequency of the cluster $\mathcal{C}_s$ in a speaker r, that can be expressed as the sum of all segment frequencies in the speaker r.

$$g_{\mathcal{C}_s}(r) = \frac{L_r(\mathcal{C}_s)}{L_r(\Omega)} = \sum_{n \in \mathcal{C}_s} g_r(n). \tag{2.17}$$

This frequency will increase if segments containing the same speaker r are merged in a single cluster $\mathcal{C}_s$, up to a point where $g_{\mathcal{C}_s}(r) = 1$ if all segments that contain the speaker r are contained in $\mathcal{C}_s$.

Similarly to the cluster purity, the speaker purity for a single speaker r can be obtained as the frequency of the cluster $\mathcal{C}_s$ that obtains higher frequency in the speaker r:

$$P_{speaker}(r) = \underset{\mathcal{C}_s}{max}(g_{\mathcal{C}_s}(r)). \tag{2.18}$$

And the speaker purity for the whole set of segments, given the hypothetical partition H, is the weighted average of the speaker purities for every speaker:

$$P_{speaker}(\Omega|H) = \frac{\sum_{r=1}^{R} L_r(\Omega) P_{speaker}(r)}{L(\Omega)}, \tag{2.19}$$

As it happened with the cluster purity, the speaker purity cannot measure the performance of the clustering task on its own. A high speaker purity indicates that for each speaker most segments containing the speaker are assigned to the same cluster. But it does not take into account the fact that different speaker must be separated. For example, the coarsest partition will always obtain a speaker purity of $P_{speaker} = 100\%$, but this is not the desired partition in this task.

Both clustering and speaker purity measures complement each other. Assuming a clustering problem where all segments contain a single speaker, the only partition that will obtain both $P_{cluster} = 100\%$ and $P_{speaker} = 100\%$ is the actual solution of the clustering, problem, the one where every cluster contains a single speaker, and all clusters contain different speakers. Starting from this partition, the act of merging clusters will degrade the clustering performance, and it will be reflected in a decrease of the $P_{cluster}$, while the $P_{speaker}$ will not change. On the other hand, splitting a cluster will also degrade the clustering performance, but this time the degradation will be reflected in a decrease of the $P_{speaker}$, and the $P_{cluster}$ will not change.

It is usual to find in the literature the counterparts of the cluster and speaker purities, the cluster and speaker impurities, defined as:

$$I_{cluster}(\Omega|H) = 1 - P_{cluster}(\Omega|H) \tag{2.20}$$

$$I_{speaker}(\Omega|H) = 1 - P_{speaker}(\Omega|H) \tag{2.21}$$

When solving the partitioning problem we want to minimize both impurities, $I_{cluster}$ and $I_{speaker}$. Since there is a trade-off between both impurities, a cost and an operating point can be set for a clustering task. Depending on the application, keeping a low $I_{cluster}$ may be more important than keeping a low $I_{speaker}$ or vice-versa. Assuming that both impurities are equally important (which is usual in a speaker diarization application), we can define the Equal Impurity (EI) as the value of any of the impurities for the partition H_{EI} that makes both impurities equal:

$$EI(\Omega) = I_{cluster}(\Omega|H_{EI}) = I_{speaker}(\Omega|H_{EI}) \tag{2.22}$$

The concept of EI is closely related to the concept of Equal Error Rate (EER) in detection tasks. It sets an operating point for the clustering task but any other operating point could be defined. The advantage of the EI is that it summarizes the performance of the clustering task in a single measure, and that this measure is usually obtained for a partition that gives a number of clusters which is close to the actual number of speakers. Other operating points will tend to obtain fewer clusters than speakers (low $I_{speaker}$) or more clusters than speakers (low $I_{cluster}$).

Finally, the most utilized measure to evaluate the accuracy of a system that performs the complete task of speaker diarization, including segmentation and clustering, is the Diarization Error Rate (DER). Given a diarization hypothesis obtained by a system for a recording, that is a set of time marks and labels indicating who is speaking when, the DER is roughly defined as the total time incorrectly assigned divided by the total time to be assigned in the recording.

There are four possible diarization errors accounted by the DER.

- **Speaker Error**: The speaker error is the fraction of the total time to be assigned to different speakers that has been assigned to an incorrect speaker. This measure does not account errors in overlapped speech or in the speech/non-speech detection. It only considers errors where the actual speaker present in a fragment and the hypothetical speaker obtained by the diarization system in that fragment are different. It can be defined as:

$$E_{spk} = \frac{T_{speech}(s_h \neq s_a)}{T_{speech}}, \tag{2.23}$$

where $T_{speech}(s_h \neq s_a)$ denotes the total time where the hypothetical speaker s_h is different from the actual speaker s_a, and T_{speech} denotes the total speech time to be assigned.

- **False alarm speech**: The false alarm speech is the time incorrectly detected as speech divided by the total time to be assigned. This error is mostly due to speech/non-speech segmentation errors (the speech/non-speech segmentation system labels a non-speech segment as a speech segment), and it is usually not related to speaker segmentation or speaker clustering errors. It can be defined as:

$$E_{fa} = \frac{T_{non-speech}(speech_h)}{T_{speech}}, \tag{2.24}$$

where $T_{non-speech}(speech_h)$ denotes the total non-speech time that has been labeled as speech.

- **Missed speech**: The missed speech is the speech time incorrectly detected as non-speech. This is error is again mostly due to speech/non-speech segmentation errors, and it is usually not related to speaker segmentation or speaker clustering errors. It can be defined as:

$$E_m = \frac{T_{speech}(non - speech_h)}{T_{speech}}, \tag{2.25}$$

where $T_{speech}(non-speech_h)$ is the total time speech incorrectly labeled as non-speech.

- **Overlapped speech error**: When multiple speakers are present in a speech segment, the diarization system should detect all the speakers present and assign the segment to all of them. Errors in the detection of the speakers present in an overlapped speech segment are accounted as overlapped speech errors. However, these errors always fall in one of the previous categories: speaker error if a speaker is detected to be present in a overlapped speech segment and the speaker is actually not present, false alarm speech if more speakers than the actual number of speakers are detected in the overlapped speech segment, or missed speech if fewer speakers than the actual number of speakers are detected in the segment.

Thus, the total DER is defined as:

$$DER = E_{spk} + E_{fa} + E_m + E_{ov}, \tag{2.26}$$

where E_{ov} denotes the overlapped speech errors. Most state-of-the-art systems cannot deal with overlapped speech, and the approaches to do so [Boakye *et al.*, 2008] are usually independent systems not included in the diarization engine that process the audio signal previously in order to remove the overlapped speech segments from the input of the diarization system. The presence of overlapped speech is a problem for speaker diarization systems that is still far to be solved.

Therefore, it is usual to analyze the accuracy of a speaker diarization system ignoring the overlapped speech segments. Usually, overlapped speech segments are not removed from the input of the diarization system, so their presence affects the operation of the system, but they are not considered as time to be scored. This means that the decisions on overlapped speech are not evaluated. In this situation, the accuracy of the speech/non-speech segmentation system and the accuracy of the speaker diarization system can be evaluated separately, assuming that speaker diarization (speaker segmentation and speaker clustering) only works on speech fragments. E_{fa} and E_m only take into account the errors obtained by the speech/non-speech segmentation system, while E_{spk} only takes into account the speaker segmentation and clustering errors.

2.7 Speaker Characterization

Speaker characterization refers to the set of techniques that allow extracting and modeling the features that enables an automatic system to distinguish between different speakers. These techniques have evolved significantly during the last decade, due to the effort of the research community in the field of speaker recognition.

Speaker characterization is closely related to speaker diarization, and the techniques developed for speaker characterization can be easily applied on the speaker diarization task. On the other hand, in some situations, speaker diarization is needed as a previous step to perform speaker characterization, for example, in environments where multiple speakers are present in a signal recording and it is desired to train a speaker model for some of the speakers. In this section we summarize the state-of-the-art techniques in the field of speaker characterization that have yield recent advances in speaker diarization, and we also analyze some studies that show the importance of speaker diarization for speaker characterization.

2.7.1 Speaker characterization for speaker diarization

The progress in speaker diarization has always been linked to the advances in the field of speaker characterization. Regarding acoustic features, the traditional MFCC of PLP features considered in both domains were previously used for ASR, but some features initially developed for speaker recognition have been later used for speaker diarization. for example, prosodic and long term features that have been suggested to be useful for speaker characterization [Shriberg *et al.*, 2005], have been applied later for speaker diarization [Friedland *et al.*, 2009b]. Another example is the use of feature normalization techniques to mitigate the influence of background noises channel variability. Feature warping, originally developed for speaker recognition [Pelecanos and Sridharan, 2001], has been used for speaker diarization [Sinha *et al.*, 2005] and [Zhu *et al.*, 2006].

But the main ideas that speaker diarization techniques have borrow from speaker characterization for speaker recognition are related to the statistical modeling of speakers. Current speaker diarization systems model speakers using GMM [Wooters *et al.*, 2004] as proposed previously for speaker recognition in [Reynolds, 1995b]. Speaker adaptation techniques starting from a UBM has also been inherit from speaker recognition (that also took them from ASR techniques). For example the UBM-MAP framework for speaker verification [Reynolds *et al.*, 2000] is considered for speaker diarization in [Barras *et al.*, 2004] and [Zhu *et al.*, 2005].

Recently, the development of techniques that study the different sources of variability in order to capture the desired variability (for example, the one that best separates different speakers) and compensate the undesired variability (that due to environment) for speaker recognition has motivated new approaches for speaker diarization. In the field of speaker recognition, these techniques try to capture the variability present among different speakers, namely inter-speaker variability, and to compensate the variability present among different recordings that contain the same speaker, namely inter-session variability.

Among these techniques, one of the most popular approaches is the Joint Factor Analysis (JFA) for speaker recognition [Kenny *et al.*, 2007]. The JFA approach models a speaker s using a GMM-supervector (GMM-sv) adapted from the UBM. The GMM-sv in the JFA paradigm speaker model can be expressed as:

$$\mathbf{m}_s = \mathbf{m}_{UBM} + \mathbf{V}\mathbf{y}_s + \mathbf{D}\mathbf{z}_s, \tag{2.27}$$

where $\mathbf{m}_s$ is the speaker dependent GMM-sv, obtained concatenating all the Gaussian means of the speaker GMM, $\mathbf{m}_{UBM}$ is the UBM mean supervector, $\mathbf{V}$ is a low rank matrix, $\mathbf{D}$ is a diagonal matrix, and $\mathbf{y}_s$ and $\mathbf{z}_s$ are normally distributed random vectors. $Vmat$ is known as the eigenvoice matrix and $\mathbf{y}$ as speaker factor vector. The columns of $\mathbf{V}$ span the subspace where most inter-speaker variability is confined. $\mathbf{D}$ models the remaining inter-speaker variability and $\mathbf{z}$ are the remaining variability factors.

However, the audio signal containing the desired speaker is always recorded under certain conditions, that may vary from one recording session to another. To model this aspect, the JFA defines the speaker GMM-sv in a given session n as:

$$\mathbf{m}_s(n) = \mathbf{m}_s + \mathbf{U}\mathbf{x}(n) = \mathbf{m}_{UBM} + \mathbf{V}\mathbf{y}_s + \mathbf{D}\mathbf{z}_s + \mathbf{U}\mathbf{x}(n), \tag{2.28}$$

where $\mathbf{m}_s(n)$ is the speaker and session dependent GMM-sv, $\mathbf{U}$ is a low rank matrix and $\mathbf{x}(n)$ is a normally distributed random vector. $\mathbf{U}$ is known as the eigenchannel matrix and $\mathbf{x}(n)$ is the channel factor vectors. The columns of $\mathbf{U}$ span the subspace where most inter-session variability is confined.

According to this model, once the hyperparameters $\mathbf{V}$, $\mathbf{U}$ and $\mathbf{D}$, and the UBM are estimated, we can obtain a speaker model given a session recording session that contains the desired speaker, simply removing the inter-session variability component from the model.

The idea of modeling inter-speaker variability to improve speaker recognition has motivated new approaches that make use of this variability model to separate two speakers present in the same recording, that is to perform speaker diarization. Some systems based in this approach are presented in [Castaldo *et al.*, 2008], [Kenny *et al.*, 2010] or [Vaquero *et al.*, 2010a].

2.7.2 Speaker diarization for speaker characterization

Speaker diarization is usually a support technology for other applications. Usually, the applications that make use of speaker diariation are those that face audio signals containing several speakers and need to use speaker dependent information in order to operate correctly. These applications need speaker diarization as a previous step to perform speaker characterization. For example, in surveillance applications, in telephone environments, target speakers may be involved in a conversation where another speaker is present. For large-scale deployments, an automatic system should segregate the two speakers to avoid manual segmentation, which might be not feasible. Other examples include automatic speaker adaptation or model selection for ASR: when large amounts of data containing different speakers need to be transcribed, prior models of some of the speakers involved or building adapted models for the most repeated speakers will enhance the Large Vocabulary Continuous Speech Recognition (LVCSR) task. In this case, speaker diarization is mandatory to take advantage of speaker adaptation techniques.

There are works that study the importance of correct speaker diarization for the correct operation of a speaker recognition system or an ASR system that uses adapted speaker models. In [Castaldo *et al.*, 2008], [Reynolds *et al.*, 2009], or [Vaquero *et al.*, 2010a], it is shown that the accuracy of a speaker recognition system that faces two speaker conversations as testing segments is severely degraded if no diarization is performed. In [Hansen *et al.*, 2005], speaker diariation is considered to perform unsupervised speaker adaptation to transcribe large broadcast news corpora, for Spoken Document Retrieval (SDR). The combination of speaker diarization and speaker adaptation improves significantly the accuracy of LVCSR.

3
Diarization for Speaker Characterization

Speaker Diarization is usually a support technology for applications that need to perform speaker characterization to operate. Traditionally, speaker diarization has been studied in broadcast news and meeting environments, with the purpose of adapting the acoustic models of a LVCSR system. It has been shown that unsupervised speaker adaptation improves the accuracy of a LVCSR in both environments [Gauvain *et al.*, 1999], [Stolcke *et al.*, 2010].

Speaker diarization has also been studied in telephone environments. Nevertheless, this environment has not been as popular as broadcast news or meetings until the recent years, with the significant progress in the field of speaker recognition. The telephone environment is quite usual in speaker recognition, since it is the only biometric technology that is straightforward to apply in these environments. However, speaker diarization has not been deeply considered for this applications. Traditionally, it was assumed that the speakers present in a telephone conversation were recorded in different channels, or that manual diarization could be performed.

In real life applications, it is usual to find telephone conversations where the two speakers are recorded over a single channel, and datasets containing two-speaker conversations that are so large to perform manual diarization. This fact has motivated the analysis presented in this chapter. Below, the importance of speaker diarization for speaker recognition when working with two-speaker conversations is studied. Also, the impact of diarization errors on the speaker recognition performance is analyzed.

3.1 Speaker Recognition

Speaker recognition is a biometric modality based on recognizing a person from his or her voice. Speaker recognition systems can perform two kinds of tasks: *speaker verification* and *speaker identification*. Speaker verification is the task of determining whether a person is the one that he or she claims to be. In this task, the system needs to be able to cope with unknown impostors (Open-set task). On the contrary, speaker identification is the task of determining who is talking among a known group of speakers (Closed-set task). In this thesis we consider the task of speaker verification.

A speaker recognition system has three stages of operation: *development* stage, *enrollment* stage and *testing* stage. The *development* stage is carried out in first place,

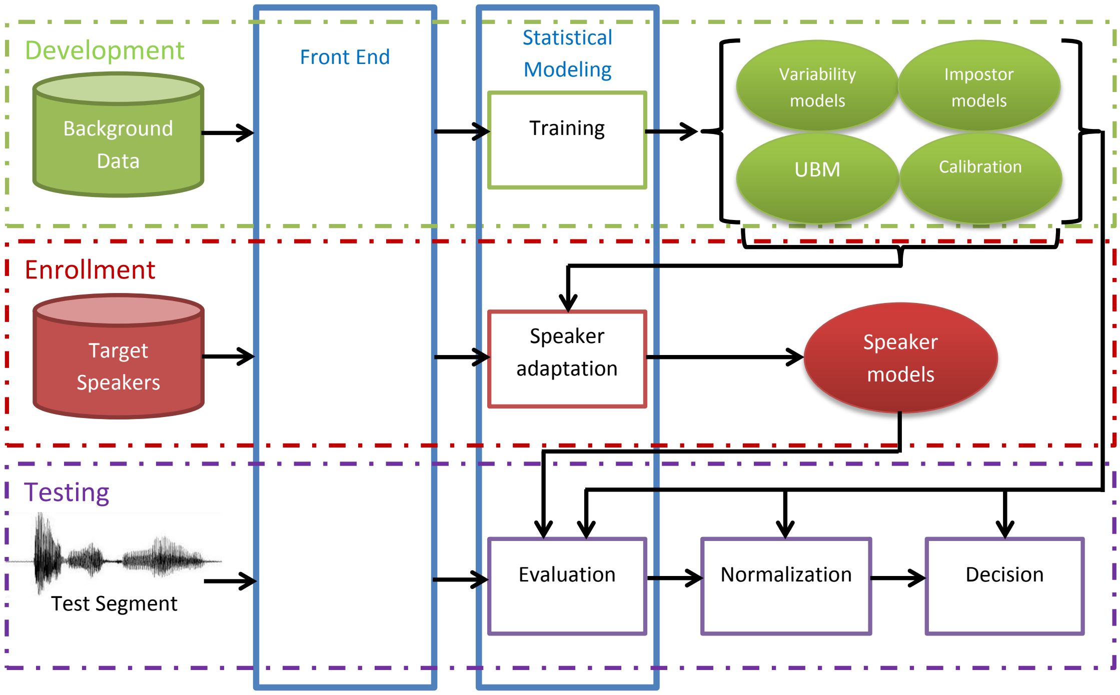


Figure 3.1: *Scheme of a speaker recognition system.*

and it involves the creation of background models, impostor models, variability models or the calibration of the system for the desired application. During the *enrollment* stage, statistical models are created for each one of the desired or target speakers. Finally, during the *testing* stage, a new speech segment is compared to the available models to make a decision about the identity of the speaker present in the speech segment.

Figure 3.1 depicts the main blocks of a typical speaker recognition system, showing the stage they are involved in, and the relation among the different blocks. These blocks are described in the following sections:

3.1.1 Front End

The Front End block extracts a set of features from the audio signal that must represent the speaker identity and enable the posterior stages to operate correctly. Features traditionally considered for speaker recognition include [Kinnunen and Li, 2010] short-term spectral features, voice source features, spectro-temporal features, prosodic features and high-level features. The spectral features try to characterize the resonances of the vocal track. The voice source features try to characterize the glottal flow. Spectro-temporal and prosodic features try to capture intonation and rhythm. Finally, high level features try to capture particular word usage (idiolect), related to learned habits, dialect and style.

Currently, short-term spectral features are the most used for the task of speaker recognition. Among them, MFCC [Davis and Mermelstein, 1980] and PLP [Hermansky, 1990] are the most popular ones. Other types of features have shown to carry additional information useful for this task, but they yield lower speaker recognition accuracy on its own compared to short-term spectral features. However, the fusion of systems using different types of features have shown to increase the final accuracy [Kajarekar *et al.*, 2009].

Another important aspect of the Front End is the frame selection for speaker recognition.

The task of frame selection involves the selection of the frames that are suitable to perform speaker recognition. Traditionally, frame selection is equivalent to Voice Activity Detection (VAD) or speech/non-speech segmentation as considered in the field of speaker diarization, and techniques such as *long-term spectral divergence* (LTSD) VAD [Ramirez *et al.*, 2004] have been used for this purpose [Villalba *et al.*, 2009].

However, the concept of frame selection is wider than this. It involves not only discarding silence frames or frames containing stationary background noise, but also frames containing acoustic activity different from the speech of the desired speaker. This acoustic activity can be due to the presence of music, non-stationary acoustic events, or another speaker in the audio signal. In case of audio signals containing several speakers, a speaker diarization system may be used as part of the Front End.

Finally, the front end usually includes feature normalization techniques. The purpose of these techniques is to mitigate the effect of the recording channel or background noise on the features. Techniques commonly used for this purpose in speaker recognition include cepstral mean subtraction (CMS) [Bimbot *et al.*, 2004], relative spectral (RASTA) filtering [Hermansky and Morgan, 1994] or feature warping [Pelecanos and Sridharan, 2001].

3.1.2 Statistical Modeling

The Statistical Modeling block establishes the framework to perform the development of the speaker recognition system given the features extracted from a background database, to obtain the target speaker models given the features extracted from a dataset of desired speakers and to provide a score or set of scores related to the likely that a test audio segment is to be uttered by a target speaker. Next, the main modules of this task are described.

- **Training**: The training module performs all the background modeling needed for the correct operation of the system. Usually, this includes creating the Universal Background Model (UBM), capturing desired and undesired variability, training impostor models, and calibrating the system for the desired application, among other tasks.

  The UBM tries to represent the complete feature space and currently most speaker recognition systems use it as starting point to obtain the target speaker models through adaptation techniques as MAP [Reynolds *et al.*, 2000].

  The variability models try to capture the desired variability present in the audio signals, and to remove the undesired sources of variability. The desired variability is the variability existing among different speakers, namely inter-speaker variability. Undesired variability includes that introduced by the acoustic channel or the background noise and is usually referred as inter-session variability. A further study on the variability present in several audio signals is presented in Chapter 5.

  Impostor models are needed to perform normalization techniques during the *testing* stage. These techniques are described later in Section 3.1.3.

  Finally the calibration of the system is crucial in open-set speaker recognition applications. One of the main objectives of the calibration process is to determine the optimal operating point of the system, that will depend on the application.

- **Speaker Adaptation**: The speaker adaptation module trains the target speaker model given an audio or set of audio signals containing the desired speaker. Traditional adaptation techniques considered to obtain speaker models are MAP [Reynolds *et al.*, 2000] or Eigenvoices [Kuhn *et al.*, 2000]. Inter-session variability compensation is usually considered in this stage in order to remove undesired sources of variation, such as channel or background noise.

- **Evaluation**: The evaluation module receives as input a target speaker model and a test segment and outputs a score indicating how likely is the test segment to be uttered by the target speaker or not. This score will be processed by posterior stages (Normalization and Decision) to make the final decision on the identity of the speaker present in the test-segment.

The statistical modeling technique considered in a speaker recognition system is one of the most distinctive aspects of the system, and most of the recent advances obtained in the field of speaker recognition have been developed in this area. In fact, the system architecture shown in Figure 3.1 has been consolidated during the last decade, with the introduction of the GMM-UBM framework for speaker verification [Reynolds *et al.*, 2000]. In this framework, the UBM is a GMM obtained during the development stage, and target speaker models are obtained by means of MAP adaptation of the UBM. The evaluation is performed simply obtaining the likelihood ratio between the likelihood of the test segment features for the speaker model and the likelihood of the test segment features for the UBM.

Another statistical modeling framework is the GMM-sv (GMM-supervector), where the GMM are represented using supervectors obtained as function of the GMM parameters (usually concatenating all the Gaussian means). This framework has enabled several modeling techniques including GMM-SVM-NAP (GMM-Support Vector Machines-Nuisance Attribute Projection) [Campbell *et al.*, 2006], JFA [Kenny *et al.*, 2007], or the i-vector paradigm [Dehak *et al.*, 2010].

The GMM-SVM-NAP technique is similar to the UBM-GMM in the sense that the speaker models are adapted using MAP from the UBM, but the evaluation is performed using SVM classifiers trained on the GMM-sv space. To train the SVM classifiers, a cohort of impostor speakers is needed as negative samples. SVM-GMM-NAP also includes inter-session variability compensation through NAP [Campbell *et al.*, 2006].

The JFA [Kenny *et al.*, 2007] paradigm, previously commented in Section 2.7.1, is based on the GMM-sv framework, and it combines MAP and Eigenvoices [Kuhn *et al.*, 2000] for speaker adaptation. The later provides fast adaptation when a small number of samples of the speaker is available, while the former provides correct asymptotic behavior when enough number of samples is available. It also performs inter-session variability compensation by means of eigenchannels. The JFA paradigm models inter-speaker and inter-session variability as low-rank non-orthogonal subspaces in the GMM-sv space, so every speaker and session could be represented using a small set of parameters (see Section 2.7.1). Although originally developed for speaker recognition, this technique has been successfully applied in the field of speaker diarization [Castaldo *et al.*, 2008], [Kenny *et al.*, 2010], [Vaquero *et al.*, 2010a].

Finally, state-of-the-art speaker recognition systems are based on the i-vector paradigm [Dehak *et al.*, 2010]. This paradigm is derived from the JFA technique where the inter-speaker and inter-session variabilities are modeled using different low-rank subspaces in the GMM-sv space. Instead of this, the i-vector approach defines a single space that is referred as *total variability space* that contains both inter-speaker and inter-session variability simultaneously. Then, the GMM-sv for a given recording session n of certain speaker s is given by:

$$\mathbf{m}_s(n) = \mathbf{m}_{UBM} + \mathbf{T}\phi_s(n), \tag{3.1}$$

where $\mathbf{m}_s(n)$ is the speaker and session dependent GMM-sv, $\mathbf{m}_{UBM}$ is the UBM mean supervector, $\mathbf{T}$ is a low rank matrix defining the total variability space and $\phi_s(n)$ is a normal distributed vector. $\phi_s(n)$ are the total variability factors or the i-vector that represents the relevant information extracted from the session n and speaker s. This approach enables us to deal with the speaker recognition task as a usual pattern recognition task, since the dimension of the i-vectors is much smaller that the dimension of the GMM-sv. Actually, in [Dehak *et al.*, 2010], the i-vector extraction is proposed as a Front End for speaker recognition, in order to facilitate the statistical modeling techniques to apply later.

Some of this techniques make use of Linear Discriminant Analysis (LDA) or Within-Class Covariance Normalization (WCCN) to compensate inter-session variability captured by the i-vectors, and obtain the cosine distance between i-vectors or train an SVM to obtain the evaluation score [Dehak *et al.*, 2009]. More recently a two covariance model has been proposed to represent the two main sources of variability among i-vectors (again inter-speaker and inter-session variability) [Brummer, 2010].

The two covariance model assumes that both inter-speaker and inter-session variability can be observed in all directions of the total variability (i-vector) space. In [Kenny, 2010] a Probabilistic LDA (PLDA) model [Prince and Elder, 2007] is considered to represent the i-vectors. The PLDA model is similar to the JFA model described in Section 2.7.1, and it assumes that the inter-speaker, the inter-session or both sources of variability are confined in low-rank subspaces. In this paradigm, an i-vector can be represented as follows:

$$\phi_s(n) = \mathbf{m} + \mathbf{V}\mathbf{y}_s + \mathbf{U}\mathbf{x}(n) + \varepsilon(n), \tag{3.2}$$

where $\phi_s(n)$ is the i-vector obtained for the session n of the speaker s, $\mathbf{m}$ is the estimated mean of i-vectors, $\mathbf{V}$ is a low rank matrix whose columns span a subspace where most inter-speaker variability is confined, $\mathbf{y}_s$ is a vector of speaker factors, since they determine the speaker component of the i-vector, $\mathbf{U}$ is a low rank matrix whose columns span a subspace where most inter-session variability is confined, $\mathbf{x}(n)$ is a vector of channel factors, since they determine the channel (session) component of the i-vector, and $\varepsilon(n)$ is a random vector that follows a normal distribution with mean 0 and diagonal precision matrix Λ_d. The speaker $\mathcal{S}$ and session $\mathcal{C}_n$ components are determined as:

$$\mathcal{S} = \mathbf{m} + \mathbf{V}\mathbf{y}_s \tag{3.3}$$
$$\mathcal{C}_n = \mathbf{U}\mathbf{x}(n) + \varepsilon(n) \tag{3.4}$$

If we consider $\mathbf{V}$ and $\mathbf{U}$ to be full-rank, this model reduces to the two covariance model [Brummer, 2010]. It is usual to consider that inter-speaker variability is confined in a low rank subspace defined by $\mathbf{V}$ but that inter-session variability is present in all directions of the total variability space. In this case, the $\mathbf{U}\mathbf{x}(n)$ term is removed and $\varepsilon(n)$ is supposed to follow a normal distribution with mean 0 and full precision matrix Λ, and to model the session component $\mathcal{C}_n$ completely:

$$\phi_s(n) = \mathbf{m} + \mathbf{V}\mathbf{y}_s + \varepsilon(n), \tag{3.5}$$

Given two i-vectors ϕ_i and ϕ_j, the PLDA paradigm for speaker verification provides an score in the form of a likelihood ratio, where two hypothesis are confronted: H_1 assumes that both i-vectors belong to the same speaker and H_2 assumes that every i-vector belongs to a different speaker. The score is then computed as follows:

$$score_{PLDA} = \frac{P(\phi_i, \phi_j|H_1)}{P(\phi_i|H_0)P(\phi_j|H_0)}, \tag{3.6}$$

3.1.3 Score Normalization

The motivation of score normalization is the fact that the distributions of the scores obtained for impostor and target speakers present large variance. This effect was firstly analyzed in [Li and Porter, 1988], and to compensate this effect the normalization of the impostor score distribution was proposed according to:

$$s' = \frac{s - \mu_{impostor}}{\sigma_{impostor}} \tag{3.7}$$

where s is the score obtained by the evaluation module, and $\mu_{impostor}$ and $\sigma_{impostor}$ are the mean and standard deviation of the impostor score distribution.

Among all normalization techniques, Z-Norm and T-Norm [Auckenthaler *et al.*, 2000] are the most popular. In Z-Norm (zero normalization), the μ and σ statistics are computed scoring a cohort of non-target test segments against the training model. On the contrary, in T-Norm (test normalization), the μ and σ are calculated scoring the test segment against a cohort non-target training models. Recently, with the introduction of the PLDA paradigm, that produces symmetric scores, the S-norm (symmetric normalization) [Kenny, 2010] has also been introduced.

Nevertheless, recent works have suggested that within the PLDA framework, score normalization is not necessary anymore [Brummer *et al.*, 2010], since the accuracy of the speaker recognition system does not seem to be altered by this technique. However, normalization techniques are helpful not only for increasing the accuracy of speaker recognition, but for obtaining a known score distribution for impostor trials as well. This helps a lot during the deployment of a speaker recognition system, making the calibration process simpler and more robust (see Section 3.1.4).

3.1.4 Decision Making

Once a score is obtained for a target speaker model and a test segment, and this score is normalized, a single step remains to complete the speaker verification process: Decision making. Decision is usually made comparing the normalized score with a threshold ϵ previously obtained during the development stage. The choice of this threshold is a trade-off between both possible errors that can appear during the operation of a speaker verification system. These two errors are the false rejections or misses and the false acceptances or false alarms. A false rejection occurs when a valid target speaker is rejected. A false acceptance happens when an impostor is accepted. Depending on the application, the threshold can be tuned to produce low miss rates (P_{miss}) and high false alarm rates (P_{fa}) of vice-versa. When a value for the threshold ϵ is set, a operating point of the speaker verification system is defined. A usual operating point to compare different speaker verification approaches is the point where $P_{miss} = P_{fa}$, known as *Equal Error Rate* (EER).

To analyze the performance of a speaker verification system, and also to compare different approaches for speaker verification, it is usual to represent the trade-off between P_{miss} and P_{fa} by means of *Detection Error Trade-off* curves [Martin *et al.*, 1997]. This curve represent the P_{miss} against the P_{fa}. Assuming that the score distributions for *target* (the test segment contains the target speaker) and *non-target* (the test segment does not contain the target speaker) trials are Normal and have the same deviations, DET curves are straight lines that enable us to compare different systems easily.

Every point of the DET curve is an operating point and has a threshold associated. The operating point may vary from one application to another and is usually defined by a cost function which weights both possible errors with different coefficients (c_{miss} and c_{fa}) and also takes into account the prior probability for finding a *target trial* (P_{tar}):

$$C_{det} = c_{miss} P_{tar} P_{miss} + c_{fa}(1 - P_{tar}) P_{fa} \tag{3.8}$$

The parameters c_{miss}, c_{fa} and P_{tar} are set according to the application. The optimum operating point is defined by the pair (P_{miss},P_{fa}) that minimizes the C_{det}.

Usually, C_{det} is normalized by dividing it by the best cost that could be obtained without processing the input data (i.e. by either always accepting for rejecting all test segments):

$$C_{norm} = \frac{C_{det}}{min(C_{det}(P_{miss} = 1, P_{fa} = 0), C_{det}(P_{miss} = 0, P_{fa} = 1))} \tag{3.9}$$

C_{norm} is easier to interpret than C_{det}, since it is a value in the range $[0, 1]$, where $C_{norm} = 0$ means that the system is not producing errors and $C_{norm} = 1$ means that the system is useless. Note that it is possible to obtain $C_{norm} > 1$, which means that the system is even worst than the best we can do making always the same decision.

The Decision making module in speaker verification can be seen as a random process that generates events X governed by a binomial distribution. These events can take two values: $X = +1$ or *target* and $X = -1$ or *non-target*. A prior for the binomial distribution $B(P_{tar})$ is provided. For every trial, we assume that it is possible to define a likelihood ratio as presented in [Gonzalez-Rodriguez *et al.*, 2003]:

$$LR(tar, nontar) = \frac{P(s|X = +1, \Theta_{tar})}{P(s|X = -1, \Theta_{nontar})}, \tag{3.10}$$

where s is the score provided by the speaker verification system, and Θ_{tar} and Θ_{nontar} are the score distributions for target and non-target trials. Note that these distributions may depend on the enrollment and testing sides of a given trial. Therefore, it is always interesting to apply score normalization techniques, ensuring that Θ_{nontar} is a standard normal distribution and it does not depend on the enrollment or the testing side of the trial.

Therefore, for every trial, given the prior for the binomial distribution and the defined likelihood ratio, which is related to the score obtained by the speaker recognition system, the posterior $B(P'_{tar})$ can be easily computed from the *log-odds*, since:

$$logit(P'_{tar}) = logit(P_{tar}) + LLR(tar, nontar), \tag{3.11}$$

where the *logit function* is defined as $logit(P) = log(\frac{P}{1-P})$ and LLR denotes log(LR).

Usually, the parameter P_{tar} of the prior distribution is defined depending on the application or computed considering a development set of trials. The main problem is to obtain $LR(tar, nontar)$ from the output scores of the speaker verification system. For this purpose, *Linear Logistic Regression* is usually considered, since it provides a methodology to obtain $LR(tar, nontar)$ from the speaker verification scores, given a development set. The $LR(tar, nontar)$ values are usually referred as *well-calibrated likelihood ratios* to differentiate them from the "uncalibrated" likelihood ratio values that some speaker verification systems provide as score.

The advantage of this posterior P'_{tar} as final measure to perform a decision is that it provides an easy interpretation of the results and the selected threshold for the desired application. In fact, the remaining parameters usually considered to determine the operating point and thus the threshold of the system (c_{miss}, c_{fa}) can be included in this methodology. There are tools available on-line that enable us to obtain *well-calibrated likelihood ratios* and the posterior P'_{tar} for the desired application depending on the parameters c_{miss}, c_{fa} and P_{tar}, given development set of trials, as the FoCal toolkit for logistic regression [Brummer, 2005].

This process of obtaining *well-calibrated likelihood ratios* and setting the threshold for the desired application is known as *Calibration* and is part of the development stage. Some speaker verification applications do not need to make a hard decision but to provide a posterior for the probability of target (P'_{tar}) for a given trial, so there is no need to set a threshold during the calibration process. Some systems set the threshold considering directly the normalized scores and do not obtain *well-calibrated likelihood ratios.* Since the *logistic function* or *sigmoid function* (inverse function of the *logit function*) is monotonically increasing, the relative ordering of the scores for a set of trials does not change when converting them to *well-calibrated likelihood ratios*, so the final accuracy does not change. However, it is a good practice to obtain *well-calibrated likelihood ratios* since they are in general easier to interpret than raw or normalized scores. In addition, logistic regression provides a framework to fuse the scores of different systems, obtaining a single *well-calibrated likelihood ratio* [Brümmer *et al.*, 2007].

3.2 Experimental Setup

The following section presents the datasets considered to perform the development, enrollment and testing stages for speaker verification and also describes the speaker verification system considered in this thesis. A evaluation setup is presented with the goal of determining the degradation of the system when the enrollment and test segments contain two speakers and no speaker diarization is considered to separate them.

3.2.1 Databases

In this study, we consider the datasets provided by the NIST for Speaker Recognition Evaluations (SRE) that NIST organizes periodically since 1997. The NIST SRE [NIST, 2010c] sets a common evaluation framework to compare different approaches for speaker verification, providing databases and additional information and setting an evaluation protocol and a desired operating point common to all the participants.

The datasets considered in this thesis are subsets extracted from the NIST SRE 2004, 2005, 2006, 2008 and 2010 corpora. These corpora are composed of recordings of telephone conversations and interviews, the former recorded over the telephone line or far-field microphones and the later recorded only over far-field microphones. In this thesis only telephone conversations recorded over the telephone line are considered.

Every NIST SRE corpus is divided into different conditions, depending on the purpose within the SRE. Conditions differ from each other in the way the data have been recorded, or the amount of data available for the enrollment and testing stages. From the NIST SRE 2004, 2005 and 2006, the n-conv conditions are considered (1-conv, 3-conv and 8-conv conditions, when available). These conditions are composed of excerpts extracted from telephone conversations recorded over the telephone line, with a duration of around five minutes each. These excerpts are recorded in stereo: each speaker in the conversation is recorded in a different channel. The recordings extracted from these conditions will be used for the development stage.

From the NIST SRE 2008, subsets extracted from the *short2* and *short3* conditions and the whole *summed* condition are considered. The subset extracted from the *short2* condition is composed of a set of five minute length telephone conversations recorded in stereo over the telephone line. This condition is intended for speaker enrollment in the NIST SRE 2008 and it is used for the same purpose in this work. The subset extracted from the *short3* condition is also composed of telephone conversations of five minute length recorded in the same way as the *short2* condition. The only difference is that the *short3* condition is intended for testing in the evaluation and it is used for the same purpose in this thesis. Finally the *summed* condition is obtained simply summing both channels of the conversations extracted from the *short3* condition. Thus, the *summed* condition is composed of the same telephone conversations as the *short3* condition, but every conversation is a mono recording that contains two speakers. The *short3* and *summed* datasets are considered to compare results obtained when diarization is and is not needed.

From the NIST SRE 2010, only the *summed* conditions are considered. The recordings in these conditions are obtained in the same fashion as in the NIST SRE 2008 *summed* condition. The NIST SRE 2010 *summed* conditions are only intended to evaluate approaches to speaker diarization that need some calibration procedure as the use of confidence measures presented in Chapter 7 or validating a clustering procedure (see Chapter 6).

In all cases, the data has been recorded with a sample rate of 8 KHz and stored in 8-bit μ-law format.

Table 3.1 summarizes the statistics for the datasets considered, and the purpose of each dataset.

Dataset	condition	Conv.	Rec. sides	Spks	Purpose
NIST SRE 2004 - 06	*n-conv*	16588	1566	1566	Development
NIST SRE 2008	*short2*	1644	1788	1304	Enrollment
NIST SRE 2008	*short3*	2213	2573	1030	Testing (stereo)
NIST SRE 2008	*summed*	2213	2213	1040	Testing (mono)
NIST SRE 2008	*all*	3857+2213	4361+2213	1319	Enrol. and Test.
NIST SRE 2010	*summed*	7130	7130	1703	Validation

Table 3.1: Statistics of the datasets considered for speaker verification

In those datasets composed of stereo and mono recordings, the number of stereo and mono recordings are separated (3857+2213 = 3857 stereo + 2213 mono conversations). In the stereo recordings, only the recording sides considered are accounted to obtain the number of speakers. Note that speakers are repeated across different NIST SRE 2008 conditions, so the total number of speakers in the NIST SRE data is not the sum of the speakers present in all conditions considered. Note also that the number of conversations in the *short3* and *summed* conditions are the same. In fact, both datasets contain the same conversations. The number of speakers for these two conditions differs since for the *short3* condition only 2573 sides out of the 4426 are considered. Not all the speakers present in the *summed* condition are labeled by NIST, some of them are unknown, but are known to be different to those that are labeled.

Along with the recordings, the NIST provides ASR transcriptions obtained for every side of the telephone conversation. These transcriptions can be used to extract high level features for speaker verification. In addition, the transcriptions provide time marks for every word, which can be useful as speech/non-speech labels. The NIST provides these labels to encourage the participants to focus on the speaker verification task rather than on data conditioning and frame selection. In this work, these transcriptions are considered in order to extract the reference labels for speaker diarization, since the transcriptions are available for stereo recordings.

3.2.2 System Description

The system considered for speaker verification in this thesis is an i-vector [Dehak *et al.*, 2010] gaussian PLDA system [Kenny, 2010]. The main blocks are described next:

The front end extracts 18 MFCC not including C0 and first delta features are computed, obtaining feature vectors of dimension 36. As speech/non-speech labels the NIST ASR speech/non-speech segmentation labels are considered (see Section 3.2.1). The front end may include a diarization system in order to separate the frames belonging to different speakers when more than one speaker is present in the recording. Finally, the frames are gaussianized using feature warping [Pelecanos and Sridharan, 2001].

The UBM is a 1024-component GMM, trained on NIST SRE 2004, 2005 and 2006 data. Thus, the GMM-sv space has a dimension of $1024 \times 36 = 36864$. A total variability subspace of 400 dimensions is estimated using also NIST SRE 2004, 2005 and 2006 data. This total variability subspace enables us to model every single speaker in a recording session with an i-vector. Then a PLDA model is trained on all the i-vectors extracted for the n-conv sessions from NIST SRE 2004, 2005 and 2006. The PLDA model considers a low-rank eigenvoice matrix that spans a subspace of dimension 100, and a full-rank channel space, as presented in eq. (3.5).

3.2.3 Evaluation Setup

In order to evaluate the described speaker verification system, a flexible evaluation setup is proposed. The aim of this evaluation setup is to analyze the accuracy of the speaker verification system when it uses recordings containing a single speaker for enrollment and testing, and how the presence of multiple speakers in a single recording and the use of speaker diarization affect this accuracy.

Two main speaker verification tasks are defined. The first one is the traditional speaker verification task where a target speaker model is given and the test segment is a mono recording or a side of a stereo recording that contains a single speaker. The objective is to determine whether or not the speaker present in the test segment is the target speaker. We refer to this task as 1 : 1 (*testing segments : models*), since every trial involves a single speaker for both testing and enrollment. The second task is introduced to analyze the importance of speaker diarization in these systems. In this task, a target speaker model and a two speaker conversation are given, and the objective is to determine whether or not the target speaker is *involved* in the test conversation. The conversation may be a mono recording containing two speakers (diarization is needed) or a stereo recording containing a single speaker on each side (perfect speaker separation for comparison). Thus, both sides of the conversation need to be analyzed to detect the target speaker. We refer to this task as 2 : 1, since every trial involves a single speaker for enrollment, but two speakers for testing.

These two task enable us to evaluate the accuracy of the speaker verification in four different scenarios:

- The enrollment and testing segments are extracted from conversations recorded in stereo, where the speakers are easily separable. In this case, no diarization is needed. The telephone conversations considered for training and testing are recorded in stereo and echo cancellation techniques can be used to ensure that each side contains only a single speaker. This is the most favorable case to perform speaker verification, and it is evaluated considering both proposed tasks (1 : 1 and 2 : 1). This way, the degradation of considering a complete conversation rather than a single side of the conversation as test input can be analyzed. We refer to this scenario as *stereo-stereo*.

- The enrollment segment is extracted from a mono recording that contains two speakers, and the test segment is extracted from a stereo recording. In this case, diarization may be used to separate the speakers in the enrollment recording. It is assumed that only the hypothetical speaker obtained by the diariation that best matches the actual target speaker is considered for enrollment. This is realistic, since in most applications the enrollment stage is performed offline. It is usually very costly to perform manual diarization, but it is possible to listen to the segmented conversation and select the fragment that best matches the desired speaker which is known during the enrollment stage. This scenario enables us to analyze the importance of diarization for enrollment when the enrollment segment contains multiple speakers. This scenario will be considered for both tasks (1 : 1 and 2 : 1). We refer to this scenario as *mono-stereo*.

- The enrollment segment is extracted from a stereo conversation, while the testing segment is extracted from a mono conversation containing two speakers. Again, diarization may be used to separate the speakers in the testing recording, but unlike the *mono-stereo* scenario, in this case it is not possible to listen to the segmented conversation since many applications deal with many testing segments that are processed online. Therefore, in case of performing any diarization on the test segment, all the hypothetical speakers need to be evaluated. This scenario is only compatible with the 2 : 1 task. We refer to this scenario as *stereo-mono*.

- The enrollment and testing segments are extracted from mono recordings that contain two speakers. Diarization may be performed in enrollment and testing stages, and during the enrollment stage, the best matching hypothetical speaker can be selected, discarding the other. Again, this cannot be done during the testing stage. This scenario is only compatible with the 2 : 1 task. We refer to this scenario as *mono-mono*.

Tasks	Scenarios	Speaker models (conversations)	Testing segments (conversations)
1 : 1	*stereo-stereo*	1458 (1359 stereo)	2559 (2203 stereo)
	mono-stereo	1458 (1359 mono)	2559 (2203 stereo)
2 : 1	*stereo-stereo*	1458 (1359 stereo)	4406 (2203 stereo)
	mono-stereo	1458 (1359 mono)	4406 (2203 stereo)
	stereo-mono	1458 (1359 stereo)	2203 (2203 mono)
	mono-mono	1458 (1359 mono)	2203 (2203 mono)

Table 3.2: Proposed tasks and scenarios for speaker verification

Table 3.2 summarizes the different scenarios and the tasks that are evaluated. A total of 1458 speaker models are trained, extracted from 1359 two speaker conversations. The test segments are extracted from 2203 two speaker conversations, obtaining a total of 2559 test segments. The two speaker conversations for train and testing sides are selected to ensure that there exists at least a target trial for all of them, so any model subset or test segment subset has target trials. All possible trials are evaluated, up to 3731022 trials (1458 × 2559) for the 1 : 1 task and 3211974 trials (1458 × 2203) for the 2 : 1 task. In the case of the 2 : 1 task, when the testing conversations are recorded in stereo, every trial is composed of a speaker model and two testing segments, since every model must be compared with both sides of the conversation. This gives a total of 4406 (2203 × 2) testing segments.

The speaker models trained during the enrollment stage are obtained from the NIST SRE 2008 *short2* condition. In the *stereo-stereo* and *stereo-mono* scenarios, the 1458 speaker models are trained on segments extracted from a subset of the *short2* condition. This subset is composed of 1359 telephone conversations recorded in stereo. For the *mono-stereo* and *mono-mono* scenarios, a dataset composed of mono conversations is built by simply merging both sides of the stereo conversation in a single channel. This is done for the 1644 conversations available in the *short2* condition, but only 1359 are considered for the evaluation of the speaker verification system. We will refer to this new dataset for enrollment composed of mono conversations as the *summed-short2* dataset. This dataset is used to train the 1458 speaker models. If a diarization system is available, diarization can be performed on every conversation to separate the two speakers. Then, the hypothetical speaker that best matches the actual desired speaker is used for training and the other hypothetical speaker discarded.

The testing segments considered during the testing stage are obtained from the NIST SRE 2008 *short3* and *summed* conditions. In the *stereo-stereo* and *mono-stereo* scenarios, the 2559 testing segments for the 1 : 1 task and the 4406 testing segments for the 2 : 1 task are extracted from a subset of the *short3* condition composed of 2203 telephone conversations recorded in stereo. For the *stereo-mono* and *mono-mono* scenarios, the same subset of 2203 conversations is extracted from the *summed* condition. If a speaker diarization system is available, the two speakers present in the conversation can be separated, but both need to be evaluated.

The accuracy of the speaker recognition system will be evaluated using DET curves. Also, the EER and the normalized version of the minimum of the detection cost function defined in eq. (3.8), with the parameters considered in the NIST evaluations will be obtained in every scenario and task. The parameter values (see eq. (3.9)) considered in the NIST evaluations are: $c_{miss} = 10$, $c_{fa} = 1$, $P_{target} = 0.01$.

3.3 Speaker Recognition without Speaker Diarization

In this section we evaluate the proposed speaker recognition system for the tasks and scenarios described in Section 3.2.3. Firstly, the *stereo-stereo* scenario is evaluated, setting the best results that can be achieved in any other scenario. Then the remaining scenarios are analyzed to study the degradation introduced by the presence of undesired speakers in the enrollment and testing stages. In the following experiments, we assume that there is no method available to get rid of the undesired speaker. This is the worst case, but is the case we face if no speaker diarization, either manual or automatic, is considered.

3.3.1 *Stereo-Stereo* Scenario

The *stereo-stereo* scenario is the one where the input data for the speaker verification system is best conditioned, with respect to the presence of undesired speakers. In this scenario, the enrollment and testing segments are extracted from conversations recorded in stereo. In most cases the enrollment and testing segments will contain only the desired speaker, with the exception of really unusual cases where background speakers appear in the desired side of the conversation. One case that is common in telephone conversation is the presence of cross-talk or echo in the different channels. This effect can introduce the speaker from the undesired side of the conversation into the desired side of the conversation. However, this effect is easy to remove when both sides of the conversation are available. In fact the stereo recordings provided by NIST in the datasets considered are supposed to be processed by an echo cancellation system.

Therefore, we can assume that the desired side of the conversation contains only the desired speaker. It means that during the enrollment stage, only the desired side of the conversation is considered. During the testing stage, for the 1 : 1 task, only the desired side of the conversation is considered. Thus a single speaker verification score is obtained for every trial. However, for the 2 : 1 task, both sides of the conversation are considered in the testing stage. Two scores are obtained for every trial, and only the maximum score is kept as final trial score. In any case, every conversation side considered in the testing stage contains a single speaker.

Task	EER	$min(C_{norm})$
1 : 1	3.12%	0.1529
2 : 1	4.02%	0.1929

Table 3.3: EER and $min(C_{norm})$ for the 1 : 1 and 2 : 1 tasks in the *stereo-stereo* scenario.

Figure 3.2 shows the DET curves obtained for the 1 : 1 and 2 : 1 tasks considering the *stereo-stereo* scenario, and Table 3.3 shows the accuracy of the speaker verification task in terms of EER and minimum of the C_{norm}. It can be seen that the system obtains reasonably good results (for contrastive results, the results for the *short2-short3* condition 6 obtained in the NIST SRE 2008 Workshop can be checked [NIST, 2010a]).

The results for both tasks are comparable, but there is a significant degradation in the accuracy obtained for the 2 : 1 task compared to that obtained for the 1 : 1 task. This degradation is due to the fact that the 2 : 1 task performed over stereo testing recordings obtains two scores for every trial, and it keeps only the maximum score. This procedure should obtain a score distribution for the *target* trials identical to that obtained for the 1 : 1

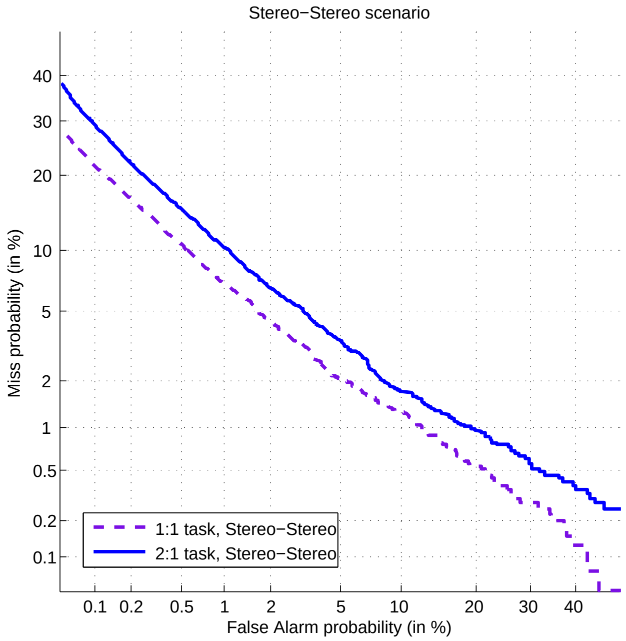


Figure 3.2: *DET curves for the* 1 : 1 *and* 2 : 1 *tasks in the stereo-stereo scenario.*

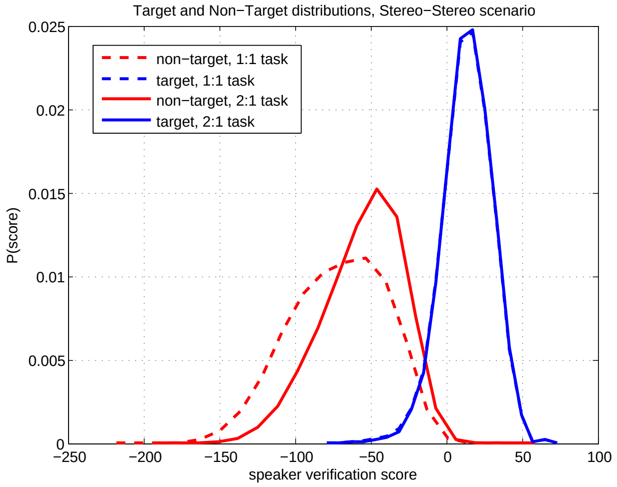


Figure 3.3: *Normalized target and non-target score distributions for the* 1 : 1 *and* 2 : 1 *tasks in the stereo-stereo scenario.*

task (assuming that in the *target* trials, the side obtaining higher score is the one where the target speaker is present, which is reasonable). However, the score distribution for the *non-target* trials will be biased with respect to that obtained for the 1 : 1 task. In effect, for a *non-target* trial, instead of analyzing a single side, we analyze both sides and keep the maximum score, so the final score in the 2 : 1 task is always equal or greater than that obtained in the 1 : 1 task.

The normalized distributions for the *target* and *non-target* scores for the 1 : 1 and 2 : 1 tasks are represented in Figure 3.3. Note that both *target* score distributions are almost identical: in the figure they seem completely overlapped. However, the *non-target* score distribution for the 2 : 1 task is shifted to the right and its variance is reduced, compared to the *non-target* score distribution for the 1 : 1 task. This shift increases the confusion

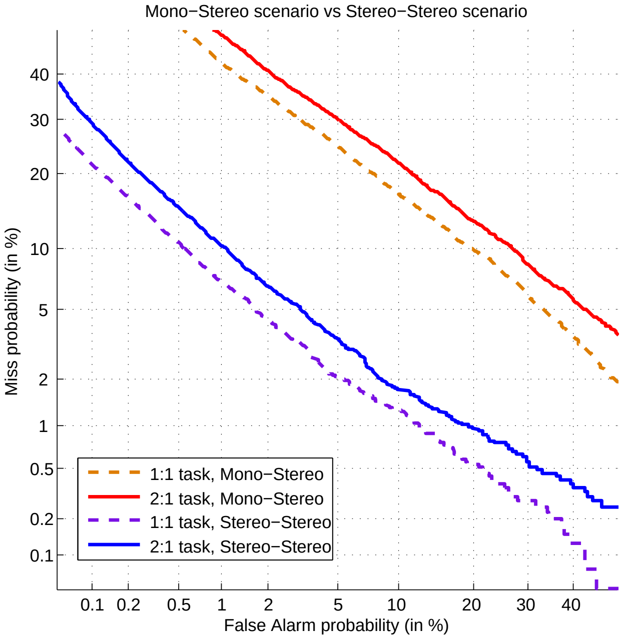


Figure 3.4: *DET curves for the* $1:1$ *and* $2:1$ *tasks in the mono-stereo and stereo-stereo scenarios.*

between *target* and *non-target* score distributions for the $2:1$ task, degrading the accuracy of the speaker verification system.

3.3.2 *Mono-Stereo* Scenario

In the *mono-stereo* scenario the testing segments are extracted from conversations recorded in stereo, but the enrollment segments must be extracted from conversations recorded in mono. Thus, the enrollment audio signals will contain the target speaker but also an undesired speaker. Assuming that there is no diarization system available (neither manual nor automatic), the whole audio signal, containing both the target and the undesired speakers, is utilized during the enrollment stage to train the target speaker model. This methodology is not usual in real applications, since in most cases the enrollment stage is performed off-line so the speakers can be separated manually. However, there are speaker recognition applications that deal with huge datasets for the enrollment of many target speakers, and manual diarization is not feasible. The analysis of this scenario will give an idea of the degradation that is obtained in this sort of applications when no automatic diarization is performed.

As in the *mono-stereo* scenario, both $1:1$ and $2:1$ tasks are evaluated. The only difference is that in this case, the target speaker model is contaminated with an undesired speaker.

Task	EER	$min(C_{norm})$
$1:1$	13.58%	0.5203
$2:1$	15.76%	0.5865

Table 3.4: EER and $min(C_{norm})$ for the $1:1$ and $2:1$ tasks in the *mono-stereo* scenario.

The DET curves obtained for the 1 : 1 and 2 : 1 tasks considering the *mono-stereo* scenario are displayed in Figure 3.4. For comparison, the curves for the *stereo-stereo* scenario are also shown. In addition, Table 3.4 shows the accuracy of the speaker verification task in terms of EER and minimum of the C_{norm} for both tasks in the *mono-stereo* scenario.

It can be seen in Figure 3.4 and in Table 3.4 that the accuracy of the speaker verification system in the *mono-stereo* scenario suffers a dramatic degradation compared to the accuracy obtained in the *stereo-stereo* scenario. The EER increases from 3.12% to 13.58% for the 1 : 1 task and from 4.02% to 15.76% for the 2 : 1 task. This degradation is kept for all operating points as we can observe in Figure 3.4, including the point of minimum C_{norm}. Such a degradation is unacceptable for most speaker recognition applications.

This huge degradation is only produced by the presence of an undesired speaker in the audio segment considered to train the target speaker models. Therefore, in the *mono-stereo* scenario, the use of some methodology to remove the undesired speaker (i.e. a diariation system) during the enrollment stage is mandatory.

It is also interesting to observe that the degradation for the 2 : 1 task compared to the 1 : 1 task observed in the *stereo-stereo* scenario is kept in the *mono-stereo* scenario. Again, selecting the highest score between two evaluations for every trial shifts the non-target score distribution degrading the accuracy of the speaker recognition system.

From now on, only the 2 : 1 task is considered, for two reasons. Firstly, the degradation due to evaluating both sides of the testing conversation for every trial (2 : 1) with respect to evaluating just one side (1 : 1) seems to introduce a constant shift in the DET curves, so there is only need to evaluate one task. Secondly, in any scenario where the testing conversation is recorded in mono, it is necessary to evaluate both speakers present. Thus, a comparison with an 1 : 1 task is not fair. Note that in some unusual situations the accuracy in the 2 : 1 task for mono testing recordings should be compared to that obtained in the 1 : 1 task for stereo recordings. For example, during the deployment of an interception system, it may be necessary to decide between an expensive stereo interception system and a cheap mono interception system.

3.3.3 *Stereo-Mono* Scenario

In the *stereo-mono* scenario the enrollment segments are extracted from conversations recorded in stereo, but the testing segments must be extracted from conversations recorded in mono. Thus, the segments considered to train the target speaker models only contain the desired speaker, while the testing segments contain two speakers. In this scenario it is only possible to evaluate the 2 : 1 task since the information of which speaker is in which side of the telephone conversation is missed. Again, if we work under the assumption that there is no speaker diarization system available, the two speakers present in the testing segment cannot be separated. Then, every trial is reduced to a single evaluation that compares the target speaker model with the complete testing segment containing two speakers. This is not very problematic for *non-target* trials, since neither of the speakers present in the test segment is the target speaker. However, for *target* trials, the target speaker in the testing segment is contaminated with an undesired speaker. Without any diarization method to separate both speakers, we can expect the scores obtained for the *target* trials to be significantly reduced, increasing the confusion between *target* and *non-target* trials.

The DET curves obtained for the 2 : 1 task considering the *stereo-mono* scenario are displayed in Figure 3.4. For comparison, the curves for the same task considering the

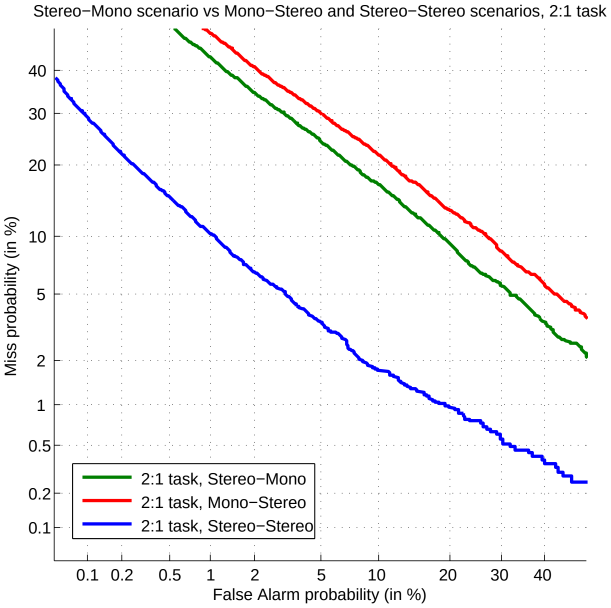


Figure 3.5: *DET curves for the* 2 : 1 *task in the stereo-mono, mono-stereo and stereo-stereo scenarios.*

Task	EER	$min(C_{norm})$
2 : 1	13.32%	0.5235

Table 3.5: EER and $min(C_{norm})$ for the 2 : 1 task in the *stereo-mono* scenario.

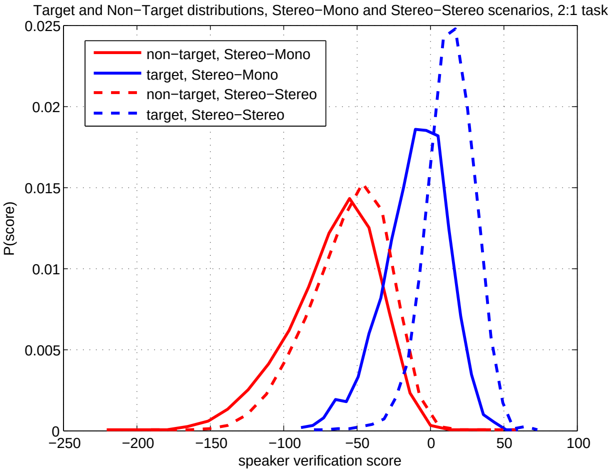


Figure 3.6: *Normalized target and non-target score distributions for the* 2 : 1 *task in the stereo-mono and stereo-stereo scenarios.*

stereo-stereo and *mono-stereo* scenarios are also shown. Table 3.5 shows the accuracy of the speaker verification task in terms of EER and minimum of the C_{norm} for the 2 : 1 tasks in the *mono-stereo* scenario.

Again, the results show a dramatic degradation in the accuracy of the speaker verification system in the *stereo-mono* scenario, compared to the results obtained in the *stereo-stereo* scenario. The degradation is kept for all operating points, including the EER point and the point of minimum C_{norm}. Such a degradation is unacceptable for most speaker recognition applications.

This degradation is only produced by the presence of two speakers in the testing segment. As commented before, in *non-target* trials, the presence of the two speakers to evaluate merged in the testing segment is not very problematic. However, in *target* trials, the target speaker segment for testing contains also an undesired speaker, so the scores for the *target* trials will be reduced.

Figure 3.6 shows the score distributions for *target* and *non-target* trials, in the *stereo-mono* and *stereo-stereo* scenarios, considering the 2 : 1 task. In effect, it can be observed that the score distributions for the *non-target* trials are similar in both scenarios. However, the score distribution for the *target* trials in the *stereo-mono* scenario is shifted to the left and its variance is increased, compared to the score distribution for the *target* trials in the *stereo-stereo* scenario.

Therefore, in the *stereo-mono* scenario, the use of some diarization approach to separate both speakers present in the testing segment during the testing stage is mandatory. Note that in most applications, the amount of data to process and the response time requirements (sometimes responses must be on-line), make impossible to use supervised diarization in this scenario, so an unsupervised or semi-supervised automatic diarization system is required.

Going back to Figure 3.4, it is also interesting to compare the DET curves for the *stereo-mono* and *mono-stereo* scenarios. It can be seen that the speaker verification system obtains better results when the conversation in mono is considered in the testing stage rather than in the enrollment stage. This effect is not due to the fact that the audio segments considered for enrollment are more critical than those considered for testing in our speaker verification approach. In fact, the scoring is symmetric in the sense that for a given trial that compares two recordings, it does not matter which recording is considered for enrollment and which one is considered for testing. Both recordings are processed in the same way and the score equation eq. (3.6) produces the same result no matter which recording is considered for enrollment or testing.

The difference in the DET curves for the *mono-stereo* and *stereo-mono* is again due to the effect of the 2 : 1 task. In the *mono-stereo* scenario, we are using a mono recording containing two speakers for enrollment, and for every trial we evaluate the two sides of the testing conversation, keeping the highest score. On the other hand, in the *stereo-mono* scenario, we are using only one side of a stereo conversation to enroll a speaker, and for every trial we only perform a single comparison between the mentioned side and a mono conversation containing two speakers. Since it does not matter which recording is considered for training or testing, the *stereo-mono* can be reversed so that it is assumed that a mono conversation containing two speakers is considered for speaker enrollment, and only one side of a stereo conversation is evaluated in every trial. This scenario is then a *mono-stereo* scenario, but the task is not a 2 : 1 anymore, but a 1 : 1 task, since the side of the stereo conversation to evaluate during the testing stage is known. In effect, it can be observed that the DET curve obtained in the *stereo-mono* scenario represented in Figure 3.5 is very similar to that obtained in the *mono-stereo* scenario for the 1 : 1 task, depicted in Figure 3.4.

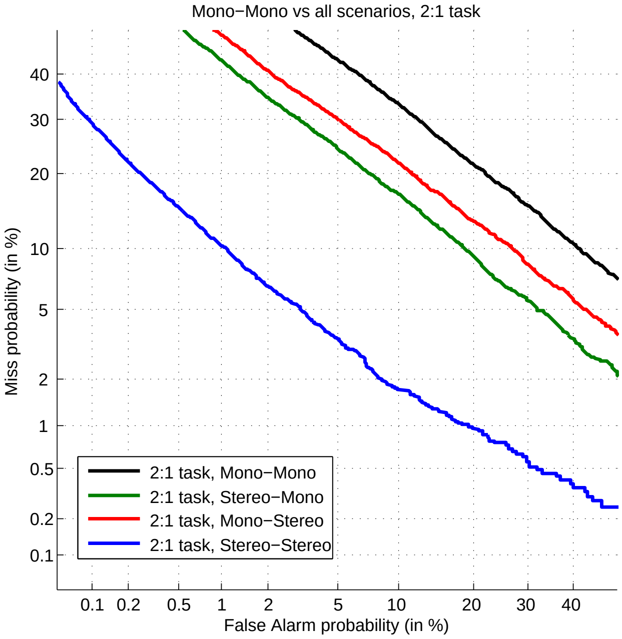


Figure 3.7: *DET curves for the* $2:1$ *task in the mono-mono, stereo-mono, mono-stereo and stereo-stereo scenarios.*

3.3.4 *Mono-Mono* Scenario

In the *mono-mono* scenario the enrollment and testing segments are extracted from conversations recorded in mono. Thus, the segments considered to train the target speaker models contain the target speaker and an undesired speaker, and the testing segments contain two speakers. As in the *stereo-mono* scenario, in this case it is only possible to evaluate the $2:1$ task since the information of which speaker is in which side of the telephone conversation is missed. If we work under the assumption that there is no speaker diarization system available during enrollment or testing, the target speaker models will be contaminated with an undesired speaker, and the two speakers present in every testing segment cannot be separated. Therefore, every trial is reduced to a single evaluation that compares the contaminated target speaker model with the complete testing segment containing two speakers. This is the worst scenario a speaker verification system can face regarding the presence of undesired speakers, assuming that the number of speakers in every conversation is limited to two (of course, the larger the number of speakers the harder the task).

Task	EER	$min(C_{norm})$
$2:1$	20.76%	0.7231

Table 3.6: EER and $min(C_{norm})$ for the $2:1$ task in the *mono-mono* scenario.

The DET curves obtained for the $2:1$ task considering the *mono-mono* scenario are displayed in Figure 3.4. For comparison, the curves for the same task considering all the other scenarios are also shown. Table 3.6 shows the accuracy of the speaker verification task in terms of EER and minimum of the C_{norm} for the $2:1$ tasks in the *mono-mono* scenario.

In the *mono-mono* scenario, the degradation in the accuracy of the speaker verification

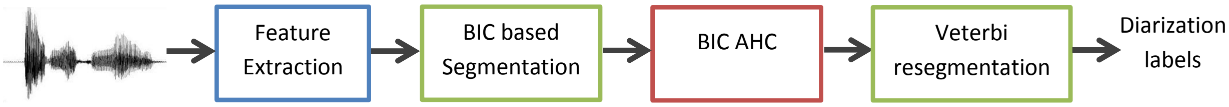


Figure 3.8: *Block Diagram of the Baseline Diarization System.*

system is significant when compared to any other scenario, specially when compared to the *stereo-stereo* scenario, the best possible scenario. Again, the degradation is kept for all operating points, including the EER point and the point of minimum C_{norm}. Such a degradation is unacceptable for most speaker recognition applications.

This huge degradation is only due to the presence of an undesired speaker in the enrollment segment, and to the presence of two speakers in the testing segment. We have studied the effect of this situations separately, and they both have shown a dramatic degradation. This scenario which combines both, suffers a degradation that makes a very accurate speaker verification system useless for most applications. Therefore, in the *mono-mono* scenario, the use of some diarization approach to remove the undesired speaker during the enrollment stage and to separate both speakers present in the testing segment during the testing stage is mandatory.

3.4 Speaker Diarization for Speaker Recognition

The importance of the use of some speaker diarization approach as a support technology for speaker verification in any scenario that involves mono conversations containing more than one speaker has been demonstrated. Now, the question is how accurate the diarization system must be to mitigate the effect of having multiple speakers in mono recordings. In this section we try to answer this question, analyzing the impact that a traditional speaker diarization system has in the accuracy of the speaker verification system. Every scenario involving mono conversations is analyzed, in order to find out the level of accuracy in speaker diarization that will be needed to obtain speaker verification results comparable to those obtained in the *stereo-stereo* scenario. Unless specified, the 2 : 1 task is considered.

3.4.1 A Traditional Speaker Diarization System

As baseline speaker diarization system, we consider a traditional BIC based AHC system [Tranter and Reynolds, 2006]. We select this approach for speaker diarization as baseline for two main reasons. Firstly, it is a state-of-the-art technique that has demonstrated very good performance in broadcast news and meeting environments, so we can expect a good behavior in telephone conversations, which is an easier task. Secondly, although there may be better approaches in the literature to solve this diarization problem, the objective is to find out the diarization accuracy needed for every one of the scenarios described in Section 3.3. For this purpose we need a system that introduces diarization errors in a significant subset of the datasets considered in this analysis to analyze the degradation in the speaker verification task.

Figure 3.8 show the components of the baseline diarization system considered. It is composed of the traditional modules needed to build a diarization system, as presented in Chapter 2, but it combines two different segmentation techniques: a metric based (BIC) for the first pass, and a model-based (Viterbi) for the final resegmentation pass. The modules depicted in Figure 3.8 are deeply described next:

- **Feature Extraction**: A feature vector composed of 12 MFCC including c0 is extracted every 10 ms, over a window of 25 ms. Features are processed rawly, no normalization or compensation technique is considered. Only the speech frames, as determined by the NIST ASR labels (see Section 3.2.1) are considered for the following modules, although the original mark times (including non-speech fragments) are considered for the segmentation algorithms.

- **BIC Segmentation**: A first segmentation pass is performed using a metric-based segmentation system (see Section 2.3). The segmentation system makes use of a sliding window of 6 second length (see Figure 2.2 in Section 2.3.1), and evaluates as candidate boundary only the point located in the middle of the window. The window advancement is 250 ms. Thus, for a given window position, two hypotheses are evaluated: H_1 states that the whole belong to the same speaker, and H_2 states that it exists a speaker boundary in the middle of the window. To do so, every hypothetical segment is modeled with a Gaussian distribution with full covariance and BIC is considered as distance metric. So, for every window location t, the $\Delta BIC(t)$ value is computed (see Section 2.3.1).

  Decisions are not taken until the whole recording has been processed. Therefore, the complete sequence of $\Delta BIC(t)$ values is obtained for the whole recording, and then the local maxima are selected as speaker boundaries. Several restrictions apply when selecting the speaker boundaries. Firstly, both fragments obtained from the sliding window at a location t when splitting the window at its middle point must contain at least 2 seconds of net speech. If this restriction is not fulfilled, the candidate boundary is kept but the fragment or fragments that do not fulfill the restriction are increased (and thus the window length is increased too) until the restriction is fulfilled and a $\Delta BIC(t)$ value is obtained. If this is not possible, the corresponding candidate boundary is discarded.

  In addition, there must be at least 3 seconds between two consecutive speaker boundaries. This restriction is considered to avoid oscillations in the $\Delta BIC(t)$ values that may produce several candidate boundaries to be selected. Thus, local maxima are also estimated over a 6 second length window. Also, a maximum number of boundaries can be selected for a given recording. This number of obtained as a function of the length of the recording: the average duration of a speaker turn is expected to be over 5 seconds, so depending on the duration of the recording T in seconds, the maximum number of speaker boundaries to select is set to $\frac{T}{5}$. In the case of NIST recordings, the length is always 5 minutes, so no more than 60 speaker boundaries are selected.

  Finally, $\Delta BIC(t)$ must be over the threshold $\theta = 0$ to select a speaker boundary, as it is usual in BIC based segmentation systems. The complexity penalty factor for the BIC computation is set to $\lambda = 1.0$.

- **BIC AHC**: Once the initial hypothetical speaker boundaries are obtained, the resulting fragments are agglomerated using again BIC as distance metric. A bottom-up clustering approach is considered, and since the number of speakers is priorly known, the stopping criterion for the clustering algorithm is fulfilled when we obtain as many cluster as the desired number of speakers (2 in this case).

- **Viterbi Resegmentation**: Finally, a speaker model is built for every cluster obtained as output of the BIC AHC module, and the frames of the recording are reassigned by means of Viterbi decoding. This step refines the speaker boundaries and enables the system to retrieve short speaker turns that were not properly segmented by the BIC Segmentation. As speaker models, 32 component GMMs are considered, and for every speaker, an HMM with 10 tied states [Levinson, 1986] whose observation distribution is given by the corresponding speaker GMM is built. The use of tied-states smoothes the speaker turns by adjusting the speaker turn duration distribution to be more realistic. Since the non-speech segments are priorly given, the decoding algorithm is forced to enter in a single non-speech state during the non-speech frames.

3.4.2 Impact of Diarization Errors on Speaker Recognition

The proposed baseline system for speaker diarization serves to study the importance of speaker diarization for speaker characterization. In addition, it is interesting to analyze the requirements in terms of diarization accuracy for every one of the three scenarios previously defined that involve mono conversations. For this purpose, the diarization accuracy for the *summed* and *summed-short2* datasets is measured in terms of DER, and then, the datasets are divided into several subsets depending on the DER values and the degradation in terms of accuracy of the speaker verification system is analyzed for every subset separately.

Dataset (subset)	DER
summed-short2 (all)	4.86%
summed-short2 (enrollment)	4.92%
summed (all)	5.21%
summed (testing)	5.20%

Table 3.7: DER for the baseline speaker diarization system evaluated on the *summed-short2* and *summed* datasets.

Table 3.7 shows the DER obtained for the *summed* and *summed-short2* datasets, and also for the subsets of these datasets considered in the speaker verification task. Note that DER is obtained considering that the speech/non-speech labels are given and overlapped speech segments are not evaluated. Note also that the results are obtained for the complete datasets and also for the subsets considered in the speaker verification tasks. Since the subsets considered for speaker verification are composed of most of the recordings of the original subsets, the results does not change significantly.

3.4.2.1 *Mono-Stereo* Scenario

We consider the baseline speaker diarization system to separate the two speakers from the *summed-short2* subset considered from enrollment in the *mono-stereo* scenario. The system gives an overall DER of 4.92%in this dataset (see Table 3.7). Our purpose is to analyze whether or not diarization is helping speaker verification in this scenario. Also, it is

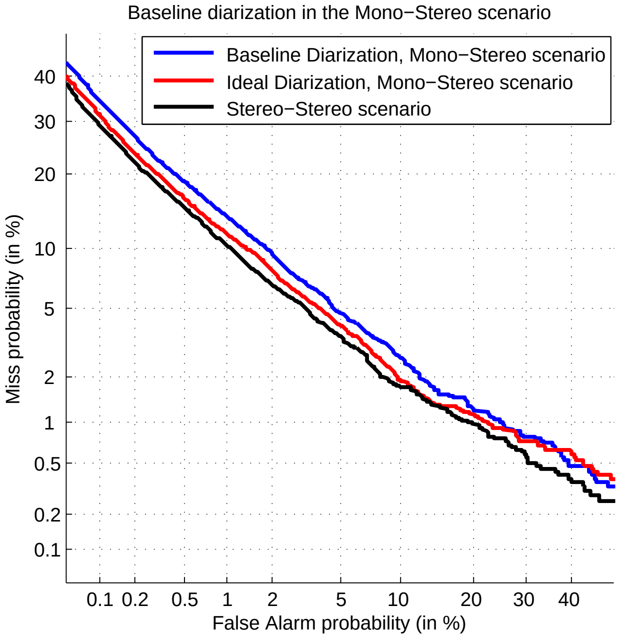


Figure 3.9: *DET curves considering the baseline and the ideal diarization systems in the mono-stereo scenario. The DET curve obtained in the stereo-stereo scenario is shown for comparison.*

interesting to determine a threshold in the DER so that no degradation due to diarization is obtained in the speaker verification task.

Note that, once the enrollment recordings extracted from the *summed-short2* dataset are processed by the diarization system, it is assumed that the target speaker to be trained is known. Thus, only the hypothetical speaker given by the diarization output that best matches the desired speaker is considered, and the remaining speaker is discarded.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.40% (0.00%)	0.2042 (0.00%)
Baseline	4.76% (8.18%)	0.2295 (12.39%)

Table 3.8: EER and minimum C_{norm} considering the baseline and the ideal speaker diarization systems in the *mono-stereo* scenario. The degradation with respect to the ideal diarization system is shown.

Figure 3.9 shows the DET curve obtained in the *mono-stereo* scenario considering the baseline speaker diariation during the enrollment. For comparison, the DET curves obtained in the *stereo-stereo* scenario and in the *mono-stereo* scenario when considering an ideal speaker diarization system, are also shown. To simulate the ideal speaker diarization system, the same NIST ASR labels considered as reference for speaker diarization evaluation are used as ideal speaker diarization labels. In addition, results in terms of EER and minimum C_{norm} are displayed in Table 3.8. In all cases, the task $2:1$ is considered.

It can be seen that the DET curve obtained considering ideal diarization is slightly above the curve obtained in the *stereo-stereo* scenario. This effect is mainly due to the presence of overlapped speech, which is not a problem in the *stereo-stereo* scenario, but in a mono recording containing more than one speaker, the fragments of overlapped speech may not be helpful for speaker characterization. In the ideal system, these fragments are assigned to both speakers, but this is not necessarily the best we can do. It is interesting to be aware of this, since it means that we will never be able to reach the DET curve obtained for the *stereo-stereo* scenario. The best that can be done is given by the DET curve obtained considering ideal diarization, unless overlapped speech is processed further.

If we focus on the DET curve obtained considering our baseline diarization system, we can see that it is slightly above the other two curves, but not very far from the ideal diarization system or the *stereo-stereo* scenario. Compared to the DET curve obtained for the *mono-stereo* scenario presented in Figure 3.4 (task 2 : 1), most of the degradation has been compensated, but there is still a small gap between the results obtained with our baseline diarization system and with an ideal diariation system.

Subset	Conversations (%)	Speaker models (%)	Trials
$DER < 2\%$	719 (52.91%)	774 (53.09%)	1705122
$2\% \leq DER < 5\%$	302 (22.22%)	328 (22.50%)	722584
$5\% \leq DER < 10\%$	142 (10.45%)	148 (10.15%)	326044
$DER \geq 10\%$	196 (14.42%)	208 (14.26%)	458224

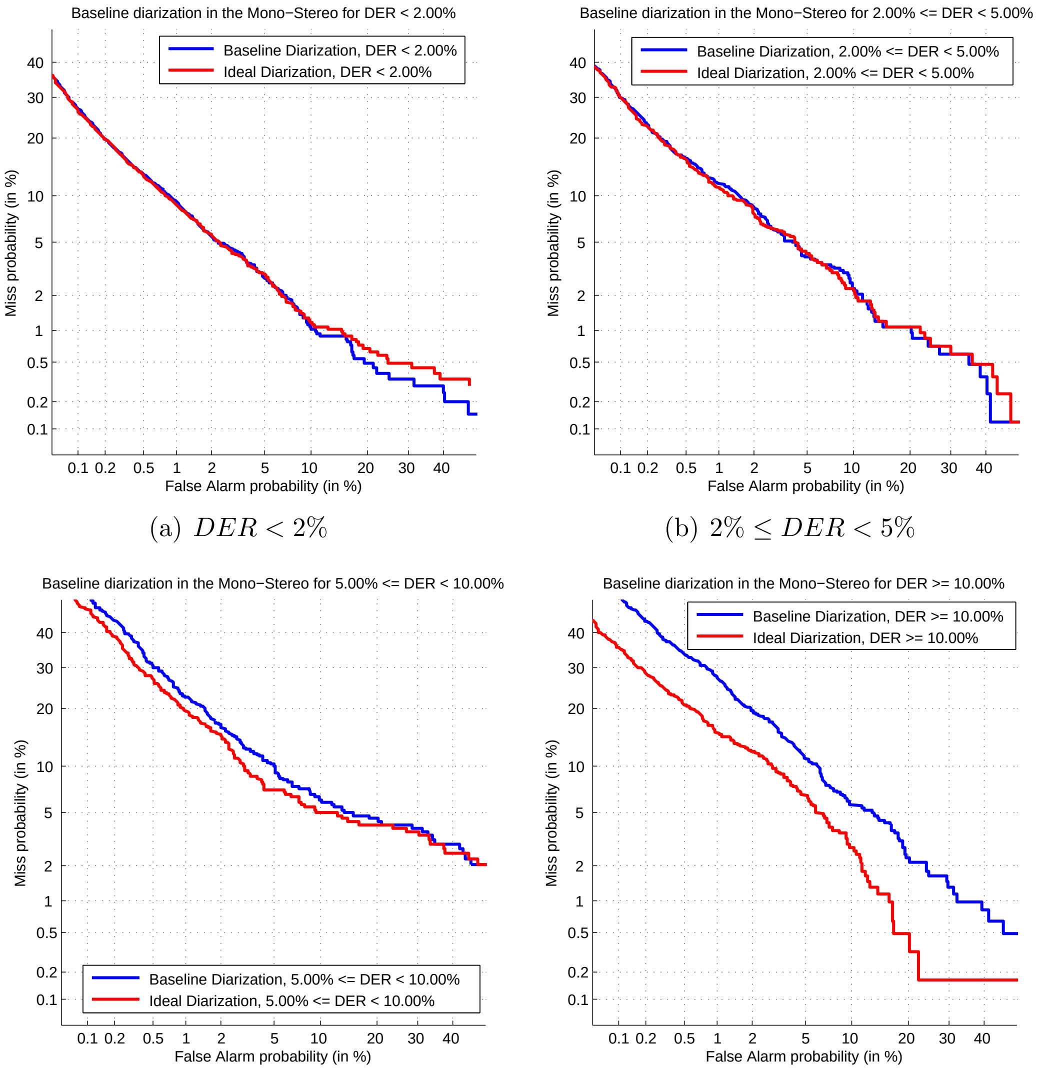
Table 3.9: Statistics of the DER dependent subsets of the *summed-short2* dataset in the *mono-stereo* scenario.

This gap becomes important when we analyze the degradation as a function of the diarization error. For this purpose, we fragment the *summed-short2* dataset considered for enrollment in this scenario into four subsets. Thus, the enrollment recordings are classified depending on the DER the baseline diarization system obtains on them. The statistics for each one of the four subsets are presented in Table 3.9. It can be seen that the baseline diarization system obtains a DER below 2% for most of the conversations. Note also that some conversations make use of both speakers involved to enroll speaker models and others do not, so the number and percentage of conversations and speaker models in a given subset may differ.

DER considered	Ideal Diarization		Baseline Diarization	
for Enrollment	EER	$min(C_{norm})$	EER (degrad.)	$min(C_{norm})$ (degrad.)
$DER < 2\%$	3.58%	0.1746	3.63% (1.40%)	0.1778 (1.83%)
$2\% \leq DER < 5\%$	4.44%	0.1973	4.31% (-2.93%)	0.2036 (3.19%)
$5\% \leq DER < 10\%$	6.22%	0.2915	7.24% (16.40%)	0.3215 (10.29%)
$DER \geq 10\%$	5.54%	0.2486	7.32% (32.13%)	0.3648 (46.74%)

Table 3.10: EER and minimum C_{norm} considering the baseline and the ideal speaker diarization systems in the *mono-stereo* scenario, for several subsets depending on the DER obtained by the baseline diarization system.

Figure 3.10 and Table 3.10 compares the accuracy of the speaker verification system considering the baseline and ideal diarization systems for every defined subset. Note that the speaker verification system obtains very similar results considering both diarization systems for the first ($DER < 2\%$) and second ($2\% \leq DER < 5\%$) subsets. The differences are not significant for these two subsets. For the third ($5\% \leq DER < 10\%$) subset, a slight degradation when considering the baseline system compared to the ideal system can



(a) $DER < 2\%$

(b) $2\% \leq DER < 5\%$

(c) $5\% \leq DER < 10\%$

(d) $DER \geq 10\%$

Figure 3.10: *DET curves considering the baseline and the ideal diarization systems in the mono-stereo scenario, for several subsets depending on the DER obtained by the baseline diarization system.*

be observed. Finally, for the fourth ($DER \geq 10\%$) and last subset, the use of the baseline diarization system show an important degradation when compared to the ideal diarization system. Note also that the accuracy of the speaker verification task when considering the ideal diarization system degrades as we consider a subset that obtains higher DER when processed with the baseline system. This effect shows that the task of speaker diarization and characterization are related in the sense that a subset that obtains poor accuracy in one task will probably obtain poor accuracy in the other.

Given these results we can affirm that diarization errors in recordings considered for enrollment introduce a degradation in speaker verification once the DER exceeds 5%, and the degradation is severe when the DER exceeds 10%. A total of $774 + 328 = 1102$ out of the 1458 speaker models are trained on recordings obtaining a DER below 5%. Thus, the degradation observed in Figure 3.9 and Table 3.8 is only due to a relatively small set

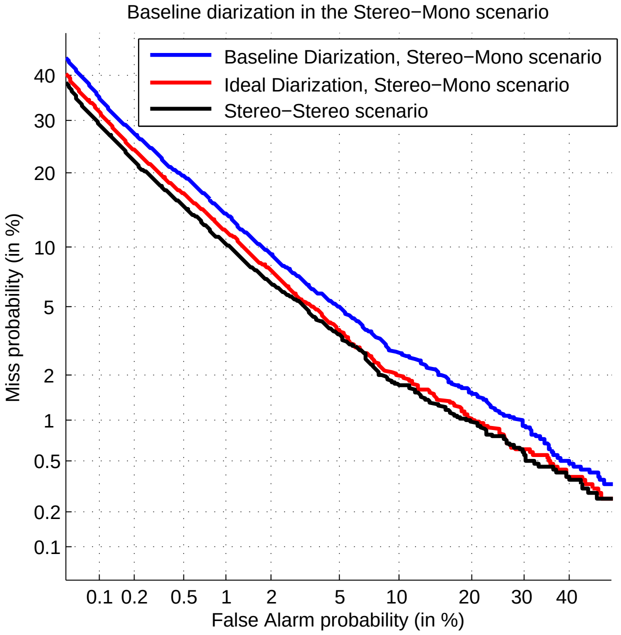


Figure 3.11: *DET curves considering the baseline and the ideal diarization systems in the stereo-mono scenario. The DET curve obtained in the stereo-stereo scenario is shown for comparison.*

of 356 speaker models (24.42% of the speaker models) trained on recordings that obtain a DER over 5%, and specially due to the subset of 208 models (14.26% of the speaker models) trained on recordings that obtain a DER over 10%.

Therefore, an improvement in the speaker diarization system will not show much better overall results on the speaker verification task. However, for certain speaker models (those trained on recordings with high DER when processed by the baseline diarization system), the accuracy of the speaker verification system can be improved significantly considering better approaches for speaker diarization.

3.4.2.2 *Stereo-Mono* Scenario

We analyze the degradation that introduces the baseline diarization system on the speaker verification task in the *stereo-mono* scenario. In this case the diarization system is considered to separate the two speakers from the *summed* subset used for testing. The system gives an overall DER of 5.20% in this dataset (see Table 3.7). Our purpose is again to analyze whether or not diarization is helping speaker verification in this scenario and also to determine a threshold in the DER so that no degradation due to diarization is obtained in the speaker verification task.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.23% (0.00%)	0.2102 (0.00%)
Baseline	4.94% (16.78%)	0.2334 (11.04%)

Table 3.11: EER and minimum C_{norm} considering the baseline and the ideal speaker diarization systems in the *stereo-mono* scenario.

Figure 3.11 shows the DET curves obtained in the *stereo-mono* scenario considering the baseline and the ideal speaker diariation systems. For comparison, the DET curves obtained in the *stereo-stereo* scenario is also shown. In addition, results in terms of EER and minimum C_{norm} are displayed in Table 3.11.

As it was observed in the *mono-stereo* scenario, in this scenario the DET curve obtained considering ideal diarization is again slightly above the curve obtained in the *stereo-stereo* scenario, due to the presence of overlapped speech assigned to both speakers in the testing stage. Also, the DET curve obtained considering our baseline diarization system is again above the other two curves. Still, compared to the DET curve obtained for the *stereo-mono* scenario presented in Figure 3.5, most of the degradation has been compensated. However, compared to the *mono-stereo* scenario, in this scenario the gap between the ideal and baseline DET curves seems to be wider.

This effect can also be seen comparing Tables 3.8 and 3.11. The degradation introduced by the baseline system in terms of $min(C_{norm})$ is similar in both scenarios, but in terms of EER is doubled in the *stereo-mono* scenario with respect to the *mono-stereo* scenario. Since we are evaluating the 2 : 1, and for such a task, the diarization is more critical during the enrollment stage (see Sections 3.3.2 and 3.3.2), we expected the opposite behavior, but in this case, the higher degradation in the *stereo-mono* scenario is probably due to a higher overall DER in the *summed* subset than in the *summed-short2* subset.

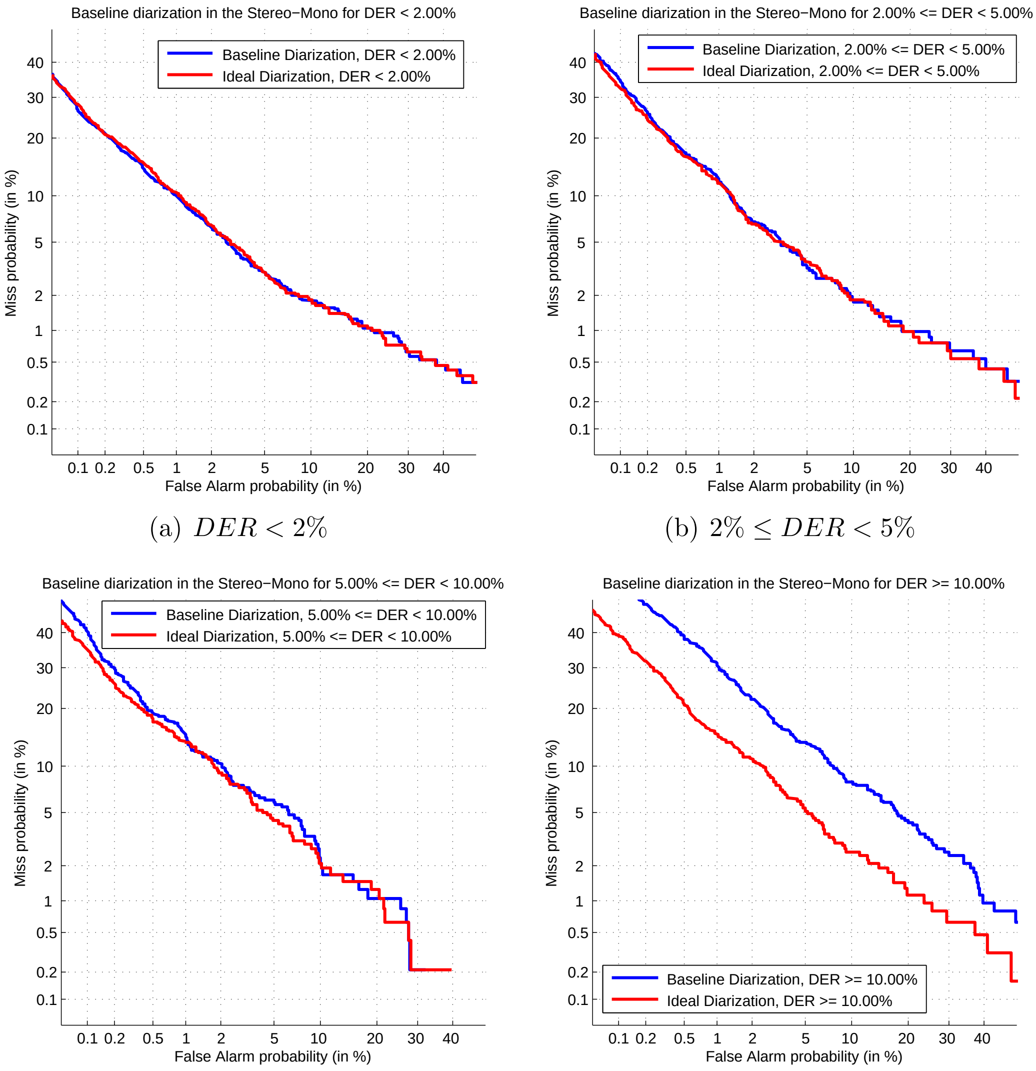
Subset	Conversations (%)	Trials
$DER < 2\%$	1124 (51.02%)	1638792
$2\% \leq DER < 5\%$	496 (22.51%)	723168
$5\% \leq DER < 10\%$	253 (11.48%)	368874
$DER \geq 10\%$	330 (14.98%)	481140

Table 3.12: Statistics of the DER dependent subsets of the *summed* dataset in the *stereo-mono* scenario.

We analyze the degradation introduced by the baseline diarization system depending on the DER obtained for every recording, as it has been done previously in the *mono-stereo* scenario. In this case, the testing recordings are classified in four subsets depending on the DER the baseline diarization system obtains on them. The statistics for each one of the four subsets are presented in Table 3.12. Again, it can be seen that the baseline diarization system obtains a DER below 2% for most of the conversations, but the rate of recordings obtaining 5% or higher DER is slightly greater than in the previous scenario.

DER considered for Testing	Ideal Diarization		Baseline Diarization	
	EER	$min(C_{norm})$	EER (degrad.)	$min(C_{norm})$ (degrad.)
$DER < 2\%$	3.81%	0.1888	3.67% (-3.67%)	0.1829 (-3.12%)
$2\% \leq DER < 5\%$	4.12%	0.2081	4.05% (-1.70%)	0.2130 (2.35%)
$5\% \leq DER < 10\%$	4.61%	0.2227	5.28% (14.53%)	0.2345 (5.30%)
$DER \geq 10\%$	5.04%	0.2397	8.66% (71.83%)	0.3962 (65.29%)

Table 3.13: EER and minimum C_{norm} considering the baseline and the ideal speaker diarization systems in the *stereo-mono* scenario, for several subsets depending on the DER obtained by the baseline diarization system.



(a) $DER < 2\%$

(b) $2\% \leq DER < 5\%$

(c) $5\% \leq DER < 10\%$

(d) $DER \geq 10\%$

Figure 3.12: *DET curves considering the baseline and the ideal diarization systems in the stereo-mono scenario, for several subsets depending on the DER obtained by the baseline diarization system.*

Figure 3.12 and Table 3.13 compares the accuracy of the speaker verification system considering the baseline and ideal diarization systems for every defined subset. As in the previous scenario, there are no significant differences in the speaker verification results considering both diarization systems for the first ($DER < 2\%$) and second ($2\% \leq DER < 5\%$) subsets. For the third ($5\% \leq DER < 10\%$) subset, a slight degradation when considering the baseline system compared to the ideal system can be observed. Finally, for the fourth ($DER \geq 10\%$) and last subset, the use of the baseline diarization system show an important degradation when compared to the ideal diarization system. In this case the degradation is even more accused than before for $DER \geq 10\%$.

Given these results we can conclude that diarization errors in conversations considered for testing introduce a degradation in speaker verification when the DER exceeds 5%, and the degradation is severe when the DER exceeds 10%. In this case, a total of 1620 out of the

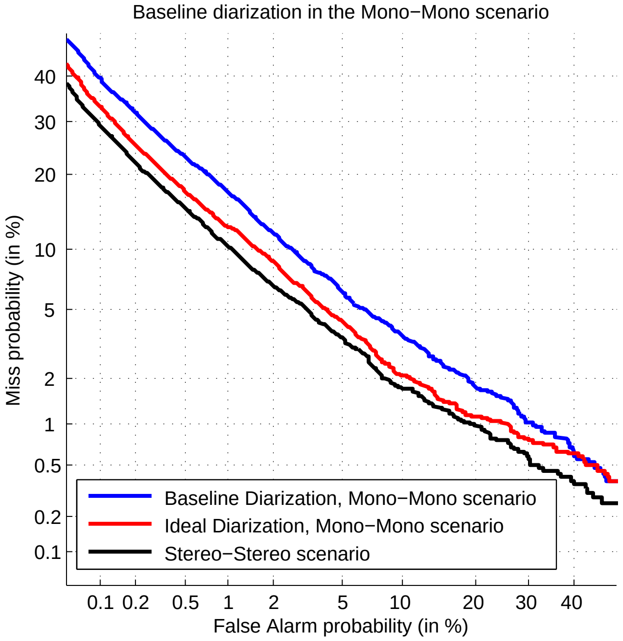


Figure 3.13: *DET curves considering the baseline and the ideal diarization systems in the mono-mono scenario. The DET curve obtained in the stereo-stereo scenario is shown for comparison.*

2203 conversations considered for testing obtain a DER below 5%. Thus, the degradation observed in Figure 3.9 and Table 3.8 is only due to a relatively small set of 583 conversations that obtain a DER over 5%, and specially due to the subset of 330 conversations that obtain a DER over 10%.

Thus, for the *stereo-mono* scenario, we can extract the same conclusions as for the mono-stereo scenario. Improving speaker diarization will not let us obtain much better overall results in the speaker verification task. However, if we focus on the specific conversations that obtain high diarization errors (DER above 5%, or specially over 10%), the accuracy of the speaker verification system can be improved significantly considering better approaches for speaker diarization.

3.4.2.3 *Mono-Mono* Scenario

Finally, the degradation that introduces the baseline diarization system on the speaker verification task in the *mono-mono* scenario is analyzed. In this case the diarization system is considered for processing both the *summed-short2* subset in the enrollment stage and the *summed* subset in the testing stage. Again, we want to analyze whether or not diarization is helping speaker verification in this scenario and also to determine a threshold in the DER for both enrollment and testing so that no degradation due to diarization is obtained in the speaker verification task.

Figure 3.13 shows the DET curves obtained in the *mono-mono* scenario considering the baseline and the ideal speaker diariation systems. For comparison, the DET curves obtained in the *stereo-stereo* scenario is also shown. In addition, results in terms of EER and minimum C_{norm} are displayed in Table 3.8.

As it was observed in the previous scenarios, the DET curve obtained considering ideal

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.54% (0.00%)	0.2157 (0.00%)
Baseline	5.53% (21.81%)	0.2695 (24.94%)

Table 3.14: EER and minimum C_{norm} considering the baseline and the ideal speaker diarization systems in the mono-mono scenario.

diarization is again slightly above the curve obtained in the *stereo-stereo* scenario. Also, the DET curve obtained considering our baseline diarization system is above the other two curves. However, in this case, the gaps between the DET curves are wider than before, specially the gap between the DET curves obtained considering the ideal and baseline diarization systems. This can be also appreciated in Table 3.14. The degradation in terms of EER and minimum C_{norm} is over 20%, much more severe than in the previous scenarios.

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	869976 (27.09%)	383904 (11.95%)	195822 (6.10%)	255420 (7.95%)
$2\% \leq DER < 5\%$	368672 (11.48%)	162688 (5.07%)	82984 (2.58%)	108240 (3.37%)
$5\% \leq DER < 10\%$	166352 (5.18%)	73408 (2.29%)	37444 (1.17%)	48840 (1.52%)
$DER \geq 10\%$	233792 (7.28%)	103168 (3.21%)	52624 (1.64%)	68640 (2.14%)

Table 3.15: Number and rate of trials for every DER dependent subset combination from the *summed-short2* and *summed* subsets in the *mono-mono* scenario.

As in the previous scenarios, we analyze the degradation that the baseline diarization system introduces in the speaker verification task with respect to the ideal diarization system, depending on the DER obtained for the enrollment and testing recordings. Table 3.15 shows the number of trials considered to obtain the accuracy of the speaker verification task for every combination of the DER dependent subsets considered for enrollment and testing. We can see that most of the trials involve recordings for enrollment and testing that obtain a DER below 5%.

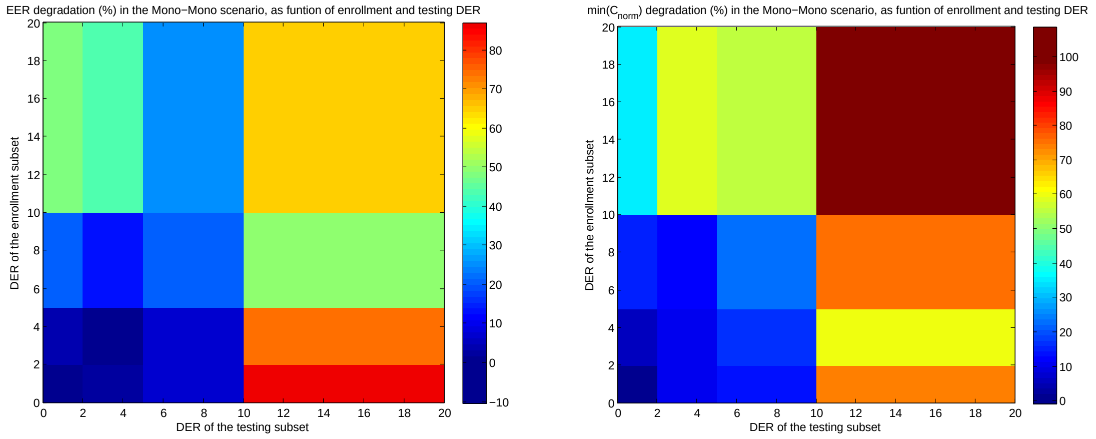
DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	0.22%	3.09%	7.54%	86.97%
$2\% \leq DER < 5\%$	3.37%	-10.45%	6.50%	74.19%
$5\% \leq DER < 10\%$	20.82%	13.60%	21.07%	49.08%
$DER \geq 10\%$	47.58%	42.98%	26.02%	65.14%

Table 3.16: EER degradation introduced by the the baseline diarization system depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario. The degradation is measured with respect to the EER obtained for ideal diarization system for the corresponding subsets.

Tables 3.16 and 3.17 show the degradation that the accuracy of speaker verification system suffers when considering the baseline diarization system with respect to the ideal diarization system in the *mono-mono* scenario. The degradation in terms of EER and $min(C_{norm})$ is presented depending on the DER obtained in the enrollment and testing

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	-0.81%	11.38%	14.19%	74.51%
$2\% \leq DER < 5\%$	6.18%	10.58%	16.84%	60.32%
$5\% \leq DER < 10\%$	16.06%	12.00%	23.84%	75.84%
$DER \geq 10\%$	35.09%	59.28%	54.92%	108.69%

Table 3.17: $min(C_{norm})$ degradation introduced by the baseline diarization system depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario. The degradation is measured with respect to the $min(C_{norm})$ obtained for ideal diarization system for the corresponding subsets.



(a) *EER degradation (%)*

(b) $min(C_{norm})$ *degradation (%)*

Figure 3.14: *EER and* $min(C_{norm})$ *degradation introduced by the baseline diarization system depending on the DER obtained for every enrollment and testing subset in the mono-mono scenario.*

conversations. The absolute EER and $min(C_{norm})$ values obtained considering the baseline and ideal diarization systems as function of the DER obtained in enrollment and testing are presented in Appendix A, along with a DET curve comparison for every DER dependent subset combination presented in Tables 3.16 and 3.17. For easier interpretation of these tables, the degradation is represented using a colormap in Figure 3.14. For clarity sake, the graphs in Figure 3.14 only represent the degradation up to a DER of 20%. The regions with DER between 10% and 20% actually represent the degradation considering those recordings with $DER \geq 10\%$.

It can be seen that considering only those recordings obtaining a DER below 5%, the degradation in terms of EER is always below 10%, and the degradation in terms of $min(C_{norm})$ is around 10% in the worst cases. This degradation is not very significant. In fact, in some cases, the baseline system obtain better results than the ideal diarization system. This effect can be due to the overlapped speech, since assigning it to both speakers is not always the best we can do. In some cases, a speaker model having little amount of speech data to be trained can benefit from using the overlapped speech, and a speaker model having enough clean speech data may be degraded adding overlapped speech. In any case, the improvement obtained by the baseline system is not significant.

The degradation in terms of $min(C_{norm})$ starts to be significant when the DER obtained either for enrollment or for testing recordings is above 5% . In terms of EER, as far as the DER obtained for enrollment is below 5%, it seems that the DER obtained for testing can be in the range [0%, 10%). In any case, the degradation in terms of EER and $min(C_{norm})$ is dramatic whenever the DER obtained for the enrollment or testing recordings is above 10%. Note also that the presence of diarization errors both in enrollment and testing increases the degradation further than having diarization errors only in one side. Since the speaker verification system produces symmetric scores, diarization errors in enrollment and testing have similar impact in the accuracy of the speaker verification system.

Given these results we can conclude that diarization errors in conversations considered either for enrollment or testing introduce a degradation in speaker verification when the DER exceeds 5%, and the degradation is severe when the DER exceeds 10%. In this scenario, since there may be diarization errors in both enrollment and testing, the overall degradation introduced by the baseline system is more severe than in the previous scenarios as it can be seen in Figure 3.13.

Therefore, in the *mono-mono* scenario, the accuracy of the speaker verification system can be improved significantly considering better approaches for speaker diarization.

Part II

Improving Diarization Accuracy

4

Accurate Diarization

In Chapter 3, it has been shown that in any scenario involving conversations with two speakers recorded in mono, the use of speaker diarization is mandatory. In addition, the impact of speaker diarization errors in the accuracy of a speaker verification system has been analyzed, and it has been shown that keeping a DER below 5% for every recording will not produce significant degradation in the accuracy of a speaker verification system, and that the degradation introduced by diarization errors is severe when the DER is above 10%.

In this chapter we present a new approach for speaker diarization that make use of successful techniques recently developed for speaker recognition. The objective is to obtain as many recordings with a DER below 5%, or at least below 10%, as possible.

4.1 Speaker Variability Modeling for Diarization

The study and modeling of inter-speaker variability, that is, the variability present among different speakers, has shown to be very successful in the field of speaker recognition [Kenny *et al.*, 2008]. Consequently, in the last years, many approaches for speaker segmentation based on inter-speaker variability modeling have been proposed [Castaldo *et al.*, 2008], [Reynolds *et al.*, 2009], [Kenny *et al.*, 2010], [Vaquero *et al.*, 2010a]. Most of these approaches build a factor analysis model using prior knowledge on inter-speaker variability to obtain a compact representation of a single speaker. This compact representation is usually a low dimension vector y, whose components are known as speaker factors. Such a representation has the advantage that, compact as it is, does not need much data to be estimated.

Most of the mentioned approaches share the way speakers are modeled. Assume that a set of T feature vectors $\chi = \{x(1), x(2), ...x(T)\}$ (for example, MFCC) of dimension D has been extracted from a recording or set of recordings that belongs to a single speaker s, and that a Universal Background Model (UBM), trained on a large and rich dataset (containing a wide variety of speakers), is available. The UBM is a GMM of C components whose component mean vectors and covariance matrices can be represented with the pair (M_{UBM}, Σ_{UBM}), were M_{UBM} is the UBM GMM-supervector, obtained concatenating all its component means, and its associated covariance matrix Σ_{UBM} is a diagonal matrix whose diagonal blocks are the diagonal covariance matrices of the GMM components. Then, every speaker is modeled using a Gaussian Mixture Model (GMM) whose means are adapted from the UBM, using an eigenvoice approach [Kuhn *et al.*, 2000] [Kenny *et al.*, 2008], according to:

$$M_s = M_{UBM} + Vy_s, \tag{4.1}$$

where M_s is the speaker s GMM supervector of dimension CD, y_s is a set of R speaker factors that represent the speaker s, and V is a $CD \times R$ low rank eigenvoice matrix that models inter-speaker variability, capturing those directions of maximum variability among different speakers. This model can be seen as a factor analysis model, that tries to describe this variability among different speakers using a small set of variables, i.e. speaker factors y, that follow a Normal Standard distribution $\mathcal{N}(y|0, I)$ a priori.

In order to train this factor analysis model, we need to obtain the parameters $\{M_{UBM}, \Sigma_{UBM}, V\}$. We usually estimate the pair $\{M_{UBM}, \Sigma_{UBM}\}$, training the UBM GMM using a large dataset and assume these estimations are good enough so we will not reestimate them. Once the UBM is obtained, the eigenvoice matrix V is trained using the factor analysis paradigm described in [Kenny *et al.*, 2008].

To perform speaker segmentation, given a recording that may contain speech from different speakers, we estimate the posterior distribution of $y(i)$ for N_w small overlapped segments $i = 1, .., N_w$, according to the factor analysis model presented in [Kenny *et al.*, 2008]. Only one speaker is assumed to be active in every segment i, and that speaker will be represented as a point estimate given by the mean of the posterior of $y(i)$, $m_{y(i)}$. Therefore, $m_{y(i)}$ will be a compact representation of the speaker present in every segment i. This way, the problem of speaker segmentation reduces to a clustering problem, where the speaker factors associated to the same speaker should be clustered together. Since we know that for a given speaker s the posterior distribution of y_s is normal, the problem of two-speaker segmentation reduces to finding the two Gaussian models that generated the obtained stream of speaker factors $Y = m_{y(1)}, ..., m_{y(N_w)}$.

Note that the point estimate of $y(i)$ considered is actually the MAP estimation of the speaker factor vector that represents the speaker present in the segment i. In fact, the point estimate selected is the one that maximizes the likelihood on the posterior distribution of $y(i)$. Since this distribution is Gaussian, its mode is equal to its mean $m_{y(i)}$.

Figure 4.1 shows the stages and modules that are involved in our approach for speaker diarization, that has been presented in [Vaquero *et al.*, 2010a]. These stages and modules are described in the following sections.

4.1.1 Front End

The first stage in the diarization system presented in Figure 4.1 is the Front End and it includes two modules: Feature Extraction and Speaker Factor Extraction. Given a recording, the Feature Extraction module extracts feature vectors traditionally considered for speaker diarization. In this work, we consider MFCC as features, but any other features reviewed in Section 2.2 could be considered as well.

Once a sequence of T feature vectors χ with dimension D is obtained, the Speaker Factor Extraction module computes speaker factor vectors of dimension R over a sliding window on the sequence χ, as shown in Figure 4.2. To compute the speaker factor vector for a given window w_i, the available (speech and non-speech) frames in the window are considered to estimate the posterior distribution of $y(i)$, whose prior is a Normal Standard distribution $\mathcal{N}(y|0, I)$. As explained before, we only consider the mean of the posterior distribution obtained as a point estimate of the speaker present in the window w_i.

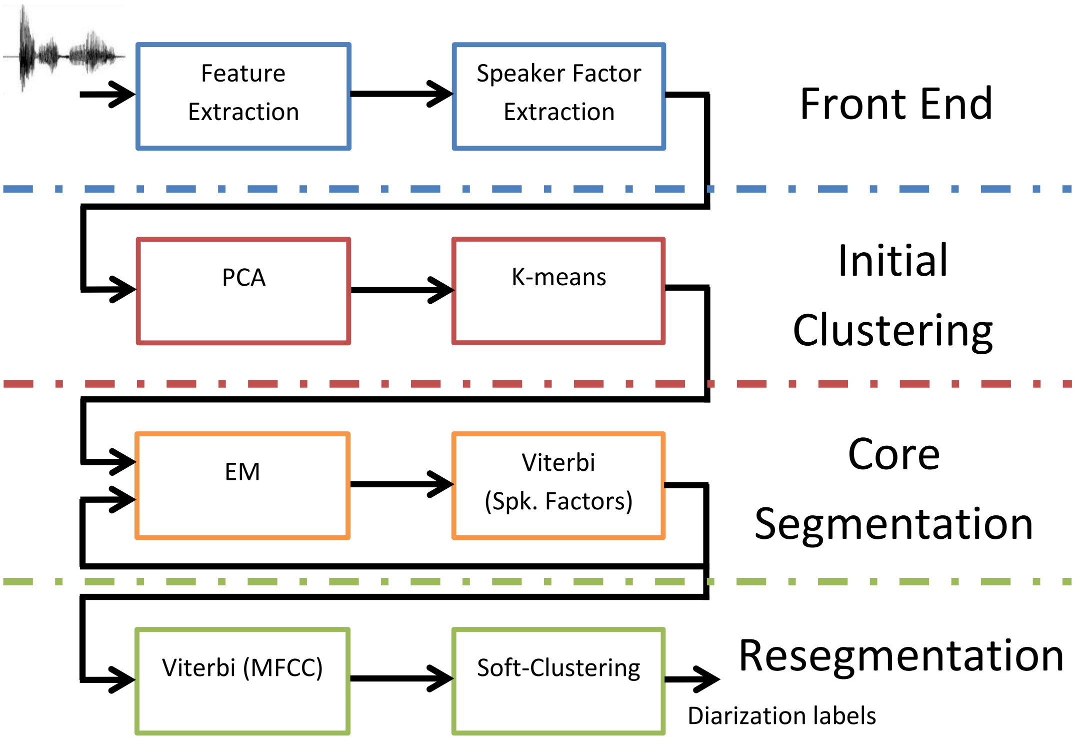


Figure 4.1: *Block Diagram of the proposed diarization system*

The sliding window is defined by two parameters: the window length L and the window step M. Both parameters are constant. The window length L defines the number of frames (including speech and non-speech frames) that will be considered to estimate the speaker factors. The window step M is the number of frames that the window is advanced every time a new speaker factor is to be computed. In this study we consider $M = 1$ in all cases, since it has the advantage that we obtain a sequence of speaker factors Y whose length is the same as the original sequence of feature vectors, and we can consider the speaker factor vectors as feature vectors that directly represent a frame of the original sequence. Considering $M > 1$ has other advantages, as computational cost reduction, that will not be explored in this work.

Setting $M = 1$ and forcing to extract a speaker factor vector for every frame simplifies the notation since the current window w_i and number of windows N_w reduce to $w_i = w_t$ and $N_w = T$. From now on, t and T will be used to refer either to feature vectors or speaker factor vectors.

The length of the window L is a critical parameter that sets a trade-off between accuracy in the estimation of $m_{y(t)}$ and accuracy in the segmentation process. Shorter windows will produce inaccurate estimations of $m_{y(t)}$, but the sequence of speaker factors will be less smooth and speaker boundaries can be determined more accurately since the temporal resolution is increased. Also the probability for including two speakers in a single window will be reduced. Longer windows will produce accurate estimations of $m_{y(t)}$, but the sequence of speaker factors will be smoother and the resolution reduced. In addition, the probability for including two speakers in a single window will increase.

In Section 4.2.2, L and other parameters, including C (number of components in the UBM), D and R(dimension of the feature and speaker factor vectors) are studied, analyzing

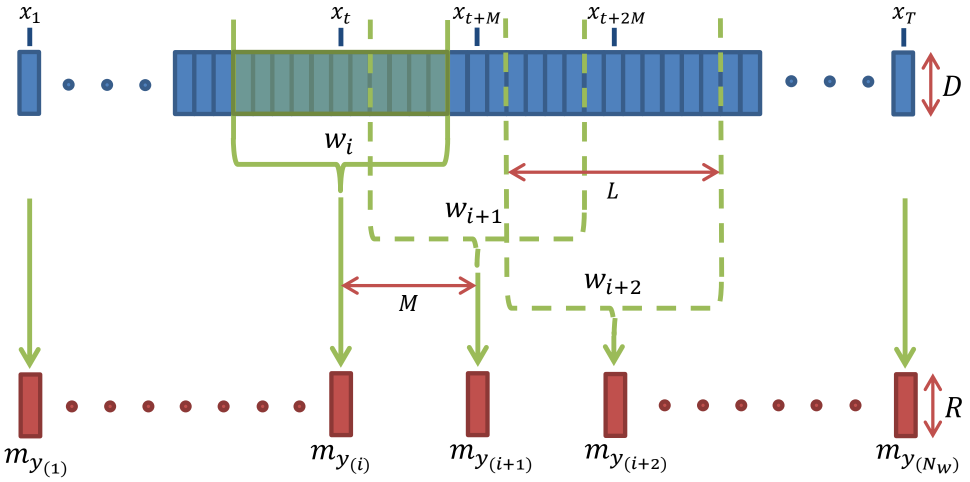


Figure 4.2: *Speaker Factor Extraction over a sliding window given the sequence of feature vectors*

their impact on the accuracy of the diarization system and determining their optimal values.

4.1.2 Initial Clustering

Once the speaker factor sequence Y is extracted, the second stage is to obtain an initial diarization hypothesis. Since the speaker factors are compact speaker representations, based on an inter-speaker variability model, we can expect that two speaker factor vectors obtained from segments (or windows) that contain the same speaker will be close to each other. On the other hand, these speaker factor vectors will be far from those extracted from segments containing a different speaker. Thus, an initial diarization hypothesis can be determined performing a simple clustering algorithm.

We can use prior knowledge about the speaker factors to enhance our clustering algorithm. It is assumed that, a priori, the speaker factor vectors are distributed as $\mathcal{N}(y|0, I)$. In fact, during the training stage, the inter-speaker variability model is forced to fulfill this assumption by means of Minimum Divergence Estimation [Kenny *et al.*, 2008]. When computing the posterior distribution of $y(t)$ given a segment t, we are considering a small set of frames (L, from tens to one or two hundreds of frames) compared to the number of frames that is usually accounted for a robust posterior estimation (the complete recording, tens of thousands of frames in most cases).

With such a small set of frames to estimate the posterior distribution of $y(t)$, we can expect this distribution to have a mean $m_{y(t)}$ that depends on the speaker present in the segment t and on the segment itself, and a covariance close to the identity matrix I. Thus, the posterior distribution of $y(t)$ is given by $\mathcal{N}\left(y(t)|m_{y(t)}, \Sigma_{y(t)} \approx I\right)$. Considering point estimates to model the segment t, given by the mean $m_{y(t)}$, we can see $m_{y(t)}$ as samples of the true speaker dependent posterior $\mathcal{N}(y_s|m_{y_s}, \Sigma_{y_s} \approx I)$ estimated over short segments. We expect the true speaker dependent posterior to have again a covariance close to I, and a mean m_{y_s} that only depends on the speaker s present in the segments, and not on the segments considered.

Following the previous assumptions, given a set of T segments containing a single speaker s, we can model the speaker s in the speaker factor space with a Normal distribution

$\mathcal{N}(y_s|m_{y_s}, \Sigma_{y_s} \approx I)$, whose mean m_{y_s} and covariance Σ_{y_s} can be estimated from the sequence speaker factor vectors $Y = m_{y(1)}, ..., m_{y(T)}$ extracted from the T available segments. Although we can in general estimate Σ_{y_s} from the available sequence of speaker factors, we expect it to be close to I.

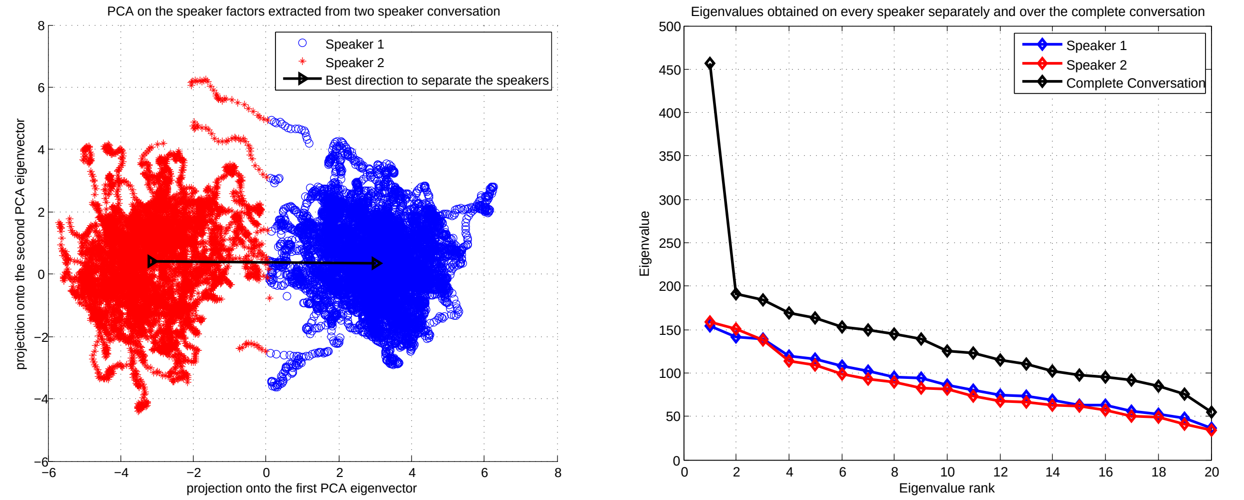
When a recording contains more that one speaker, for example, two in the case of a telephone conversation recorded in mono, Y is actually generated by two random processes, each one corresponding to a unique speaker. The distribution that governs the speaker factor vectors corresponding to each one of the two speakers is unknown, but under the previous assumptions, both distributions are Normal with different means and covariance close to I. Therefore, the speaker factors can be easily clustered into two sequences considering the two modules depicted in figure 4.1.

Firstly, PCA is applied in order to find the best direction to separate the two speakers present. In effect, the direction of maximum variability among the random samples generated by two Normal distributions with identical and spherical covariance matrix and different means is the same as the one determined by the vector obtained as difference between the means of the Normal distributions, assuming that enough samples for each distribution are available. Then, the speaker factors are projected onto the single dimension subspace generated by the first eigenvector given by the PCA algorithm (the one that we expect to provide the best direction to separate the speakers), and K-means is used to cluster the scalar values obtained into two clusters.

The clusters provided by the PCA module are used to initialize the centroids for the K-means algorithm (K-means module), but this time, all dimensions in the speaker factor space are considered. Since the distribution of the speaker factors for every speaker are assumed to have the same spherical covariance matrix (I), the K-means algorithm, correctly initialized should provide a good initial diarization hypothesis. K-means is performed until convergence, obtaining the clusters that are the output of the Initial Clustering stage. Note that two K-means are performed: the first one is considered as part of the PCA intialization (inside the PCA module), it works with a single dimension, and it is aimed at obtaining some first rough cluster labels to initialize the second K-means. The second K-means is performed in the K-means module and its objective is to refine the rough clusters provided by the PCA module.

Figure 4.3(a) shows the speaker factor vectors extracted for an audio from the *summed* dataset, projected onto the two first directions of maximum variability determined by PCA. For clarity sake, the speaker factors corresponding to non-speech and overlapped speech frames are not represented. The direction of maximum variability is the one determined by the x-axis. The best direction to separate the speakers under the assumption that both have identical and spherical covariance is depicted with a black line. It can be seen that this line is aligned with the direction of maximum variability, so the first direction given by PCA is a good estimation of the best direction to separate the speakers. Note also that the speaker factors extracted for two different speakers are easily separable.

Figure 4.3(b) shows the eigenvalues obtained by PCA considering every speaker separately and the complete conversation. The eigenvalues are sorted by value. It can be seen that the first eigenvalue for the complete conversation is much higher than the remaining eigenvalues and also than the eigenvalues computed for every speaker separately. This first eigenvalue is the one that corresponds to the direction selected to separate both speakers. Also, note that the eigenvalues for every speaker decrease smoothly. The eigenvalue spread for every speaker, defined as:



(a) *Speaker factors projected onto the two first directions of maximum variability given by PCA. The direction of maximum variability (x-axis) is the best to separate the speakers*

(b) *Eigenvalues obtained applying PCA to the speaker factors considering every speaker separately and the complete conversation. The eigenvalues are ranked by their values.*

Figure 4.3: *PCA operation in the speaker factor space, for the recording "gbzfe".*

$$spread(\lambda) = \frac{\lambda_{max}}{\lambda_{min}}, \tag{4.2}$$

is around 4 on both cases, so the covariance matrix is not spherical. However, the assumption of spherical covariance can be valid for K-means clustering since the variability between both speakers (related first eigenvalue for the complete conversation) is much higher than the variability within every single speaker (related to the first eigenvalue for every speaker).

4.1.3 Core Segmentation

As input to the Core Segmentation stage we have a sequence of speaker factors, grouped into two clusters, that represent two different speakers. For each one of the clusters, we train a Gaussian speaker model. Every one of these Gaussian models is an initial estimation of the posterior distribution of the speaker factors y_s for the speaker s, when estimated over short windows. At this point, since enough samples for every speaker are available, both the speaker mean and the full covariance matrix can be estimated for every speaker model, and we do not need to assume that $\Sigma_{y_s} = I$ anymore.

The Gaussian speaker models are used to initialize a two-component GMM and then several Expectation-Maximization (EM) iterations are run in order to obtain the two Gaussians that best fit the sequence of speaker factors. We do not expect these Gaussians to be different from the initial Gaussian speaker models, but EM provides robustness when the Gaussian distributions of the speaker factors are overlapped, or the covariance is far from the identity I.

From the GMM obtained after several EM iterations, the two components are extracted and considered as Gaussian speaker models in the speaker factor space. Then, a Hidden Markov Model (HMM) is built for every one of the two speakers. Both HMMs are composed of a left-to-right sequence of tied states [Levinson, 1986] (states that share the same PDF for the observations), and all states make use of the corresponding Gaussian speaker model as observation distribution. Tied states are considered in order to obtain a more realistic distribution for the speaker turn duration, avoiding the geometric distribution (obtained with a single state) that cannot accurately model real speaker turn durations. The distribution of the speaker turn duration is studied further in Section 4.2.4.

To model non-speech frames, a Markov Chain with a single state is considered. No PDF is needed for the silence since it is assumed that a speech/non-speech detection system has provided non-speech labels previously. Thus, the decoding algorithm is forced to go through the non-speech state for all the frames that the speech/non-speech detection system detected as non-speech.

We consider the Viterbi algorithm to find the most likely sequence of states, and the speaker factor vectors are reassigned to the two speakers according to this sequence. This frame reassignment enables us to obtain new Gaussian speaker models, that can be refined again building a GMM and performing EM iterations, and a new reassignment can be obtained by Viterbi decoding considering the new Gaussian speaker models. This process is done iteratively, until convergence. Convergence is reached whenever two consecutive iterations obtain the same speaker labels for the sequence of frames.

4.1.4 Resegmentation

The output of the Core Segmentation System gives accurate speaker labels in most cases, but the Core Segmentation may not be very accurate depending on the length of the window L considered to estimate the speaker factors. We mentioned in Section 4.1.1 that larger L values will produce better estimates of the speaker factors for every window, but the speaker factors will be smoother and the location of the speaker boundaries will be less accurate. In addition, the longer the window, the longer the speaker turns that may be missed, since short speaker turns may not be long enough compared to the window to produce clean speaker factors, specially if the short speaker turn is surrounded by the other speaker.

To avoid these problems, a resegmentation stage is introduced. This stage performs segmentation passes to refine the speaker boundaries and to retrieve short speaker turns that were not properly segmented by the previous stages. In this case we consider MFCC as features since they are estimated over shorter windows.

As speaker models, GMMs are considered, and for every speaker, an HMM whose observation distribution is given by the corresponding speaker GMM is built, considering the same number of tied states as in the Core Segmentation. In a first resegmentation pass, the frames are reassigned using Viterbi decoding, as in the baseline diarization system. Then, new speaker GMMs and HMMs are built according to the frame reassignment, and a Soft-clustering pass is performed. Soft-clustering is a resegmentation technique firstly presented in [Reynolds *et al.*, 2009], that comprises two steps: firstly, a forward-backward pass is run in order to perform a soft reassignment of the frames to the two speakers. Then, GMM models are retrained according to the soft reassignment and new HMMs are built. A final Viterbi resegmentation is performed considering these last HMMs.

As in the Core Segmentation, the decoding processes (Viterbi and forward-backward) are forced to go through a non-speech state for all the frames that the speech/non-speech detection system detected as non-speech.

4.2 System Configuration

In this section we analyze the influence of the most relevant parameters of the proposed diarization system on its the accuracy. This diarization system is intended for speaker characterization, so the final objective is to obtain as many recordings as possible whose DER is below the threshold values determined in Chapter 3. These threshold values are 5% to avoid degradation in speaker verification, and 10% to keep a low degradation when the diarization system is compared to a ideal diarization system.

The complete NIST SRE 2008 *summed* dataset (see Section 3.2.1), which contains 2213 two-speaker conversations of 5 minute length, is considered to study the speaker factor based diarization system. The speech/non-speech and reference (ideal diarization) labels are extracted from the ASR transcripts provided by NIST as in the baseline diarization system described in Section 3.4.1. Results are evaluated in terms of DER, and also in terms of the percentage of recordings in the dataset that obtains a DER below 5% and 10%.

Firstly, the default configuration is presented, showing the selected values for the main parameters of the diarization system. Then, this configuration is validated, stage by stage, analyzing every parameter separately. This way, the influence of every parameter on the overall accuracy can be easily studied.

4.2.1 Default Configurations

Two different configurations are proposed for the speaker factor based diarization system. Both configurations only differ in the front end, specifically, in the number of features considered, the size of the UBM and the number of speaker factors extracted. The first configuration is referred as *light-weight*, since it uses a small set of features and a small UBM to extract low dimension speaker factor vectors. The second is referred as *heavy-weight*, since it considers a larger set of features and a larger UBM to extract speaker factor vectors with higher dimensionality. The names are selected according to the computational cost. The speaker factor extraction is one of the most costly steps in the proposed approach for speaker diarization and its computational cost is of $O(CDR + R^2)$, where D is the dimension of the feature vectors, C the number of components of the UBM, and R the number of speaker factors. Therefore, the *light-weight* configuration, which is expected to be faster and less accurate, is intended for applications where computational cost is critical, and the *heavy-weight* configuration, which is expected to be slower but more accurate, is intended for applications where computational cost is not important.

The *light-weight* configuration for the proposed diarization system uses 12 MFCC including C0 as features, computed every 10 ms over a 25 ms window, with no delta features and without any sort of compensation or normalization. The C0 is included since it is known that it helps for diarization purposes. The features are not normalized since normalization techniques aims at compensating for variability mostly introduced by the channel and environment conditions. This variability may characterize the speakers within a single session, so it may help the diarization system. The features considered here are identical to those extracted in the baseline system presented in 3.4.1. Thus, the comparison between both systems is fair in terms of information extracted from the speech signal. A gender-independent UBM of $C = 256$ components is trained on the NIST SRE 2004, 2005 and 2006 databases. All components in the UBM GMM have diagonal covariance matrix. An eigenvoice matrix V of rank $R = 20$ is trained on the same databases, and for every input recording we estimate the speaker factors frame by frame (every 10 ms) over a window of 100 frames (1 second length). Thus, two consecutive windows have 99% overlap. The speaker supervectors M_s have a dimension of $CD = 256 \times 12 = 3072$, and the speaker factor vectors of the obtained sequence $Y = m_{y(1)}, ..., m_{y(T)}$ have a dimension of $R = 20$.

On the other hand, the *heavy-weight* configuration for the proposed diarization system makes use of 19 MFCC including C0 plus delta as features ($D = 38$), computed every 10 ms over a 25 ms window, without any sort of compensation or normalization. A gender-independent UBM of $C = 1024$ components is trained on NIST SRE 2004, 2005 and 2006 databases. All components in the UBM GMM have diagonal covariance matrix. An eigenvoice matrix V of rank $R = 50$ is trained on the same databases, and for every input recording we estimate again the speaker factors frame by frame (every 10 ms) over a window of 100 frames (1 second length, 99% overlap). The speaker supervectors M_s have a dimension of $CD = 1024 \times 38 = 38912$, and the speaker factor vectors of the obtained sequence $Y = m_{y(1)}, ..., m_{y(T)}$ have a dimension of $R = 50$.

The Initial Clustering, Core Segmentation and Resegmentation stages share the same configuration for both the *light-weight* and *heavy-weight* configurations. In the Core Segmentation stage, a total of 8 EM iterations are performed to find the Gaussian speaker models that best fit the sequence Y, and every speaker HMM is composed of 10 tied states, whose observation distributions are the Gaussian speaker models in the speaker factor space. Since Y varies smoothly (the overlap is of 99%), the transition probability for all states is set to 0.1. Thus the stay probability is set to 0.9 in all states of the HMM.

The Resegmentation stage is performed using again 12 MFCC including c0, with no delta features. In this step, every speaker is modeled again using an HMM with 10 tied states, but the observation distributions are now 32-component GMM speaker models in the MFCC space. This time the transition probability is set to 10^{-3} since MFCC are estimated with much less overlap that speaker factors.

System module	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
PCA	20.26%	377 (17.04%)	676 (30.55%)
+K-means	4.79%	1708 (77.18%)	1936 (87.48%)
Core seg	3.02%	1889 (85.36%)	2052 (92.72%)
+Viterbi reseg	2.24%	2003 (90.51%)	2100 (94.89%)
+Soft-clustering	**2.12%**	**2014 (91.01%)**	**2107 (95.21%)**

Table 4.1: Accuracy of the proposed diarization system with the *light-weight* configuration in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Tables 4.1 and 4.2 show the accuracy of the described diarization system, analyzed step by step (see Fig. 4.1), for the two proposed configurations. It can be seen that the complete system obtains very low DER: 2.12% for the *light-weight* and 1.77% for the *heavy-weight* configuration. As expected, the *heavy-weight* configuration outperforms the *light-weight* one. It is interesting to notice that just using the PCA+K-means initialization the system output is very accurate, and at that point, frames are assigned to one speaker or the other assuming statistical independence, no context or temporal information is used. Note also that the relative improvement introduced by the Core Segmentation stage is lower for the *heavy-weight* than for the *light-weight* configuration. Thus, increasing D and C in the Front End enables the Initial Clustering to better separate the speakers, and reduces the improvement

System module	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
PCA	18.12%	524 (23.68%)	849 (38.36%)
+K-means	2.88%	1930 (87.21%)	2079 (93.94%)
Core seg	2.37%	1996 (90.19%)	2093 (94.58%)
+Viterbi reseg	1.85%	2052 (92.72%)	2112 (95.44%)
+Soft-clustering	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**

Table 4.2: Accuracy of the proposed diarization system with the *heavy-weight* configuration in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

margin of the Core Segmentation stage. In addition, it is interesting to observe that the Resegmentation stage introduces similar relative improvement is both cases. Finally, note that every stage and step consistently improves the results in terms of DER and percentage of recordings below the thresholds determined for the DER.

Comparing the overall DER in Tables 4.1 and 4.2 to that obtained in Table 3.7 for the baseline system, that was 5.20%, it can be observed the improvement in accuracy introduced by this approach. These results are further compared to the baseline system in Section 4.3.

4.2.2 Front End Configuration

There are several parameters to set in the Front End that affect the accuracy of the diarization system. Among them, some of the most interesting are the dimension of the feature vector (D), the number of components in the UBM (C), the length of the window considered to extract speaker factors (L), and the dimension of the speaker factor vector R.

The parameters C and D are closely related since they are set with the objective of obtaining a speaker model space of high dimensionality where the speaker supervectors M_s are easily separable (see eq. (4.1)). In fact, since we use a Gaussian distribution to model M_s in that space, we expect the speaker supervectors M_s to present linear separability. For this purpose, it is reasonable to expect that the higher the dimension of this space, which is given by $C \times D$, the better the separability. Thus increasing C and D will increase the accuracy of the proposed speaker diarization system. In fact, comparing the accuracy of the *light-weight* and the *heavy-weight* configurations, it can be seen that this is true. However, we have to be careful since there is a third parameter that is also increased in the *heavy-weight* configuration: the dimension of the speaker factor vector R. In this section, the influence of this parameter is studied exhaustively for both configurations, and the improvement due to the difference of $C \times D$ between both configurations is fairly analyzed.

The research community has explored the influence of C and D in the task of speaker verification. In the latest NIST evaluations [NIST, 2010b], the best performing systems were considering UBMs with up to $C = 2048$, but some systems with smaller UBMs $C = 512, C = 1024$ obtained competitive results. As features, more than 20 MFCC plus delta and double delta features are hardy ever considered. It is not our intention to present an exhaustive study regarding C and D parameters, since in the recent NIST evaluations these parameters have been studied and optimized exhaustively for the task of speaker verification and the conclusions extracted there apply also for this task. However, the conclusions extracted in the NIST evaluations regarding the dimension of the speaker factor vectors R does not apply here, since in the diarizatrion task we estimate the speaker factors over shorter segments, so we expect to obtain less robust estimations. Thus, in this work, the focus is on the R and L parameters and the trade-off existing between robust estimation of speaker factors and accuracy in the speaker boundary detection.

Window Length	Core Seg.	Complete system		
	DER	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$L = 0.1s$	24.43%	12.22%	1175 (53.10%)	1361 (61.50%)
$L = 0.2s$	10.69%	5.28%	1722 (77.81%)	1859 (84.00%)
$L = 0.5s$	3.71%	2.38%	1979 (89.43%)	2080 (94.58%)
$L = 1.0s$	**3.01%**	**2.12%**	**2014 (91.01%)**	**2107 (95.21%)**
$L = 2.0s$	4.92%	2.53%	1984 (89.65%)	2074 (93.71%)

Table 4.3: Accuracy of the proposed diarization system with the *light-weight* configuration for several window length (L) values. The accuracy is measured in terms of DER at the output of the Core Segmentation stage and in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system.

The first parameter to analyze is the window length L considered to estimate the speaker factors. Table 4.3 show the accuracy of the speaker diarization system as a function of the window length, considering the *light-weight* configuration. The results presented are obtained at the output of the Core Segmentation stage (since the Resegmentation stage is not directly affected by the value of L) and at the output of the complete diarization system. It can be seen that the optimal value for L among all tested values is around $L = 1second$. In fact, as the L value gets further from this optimal value, the degradation is dramatic.

As it has been mentioned previously, there is a trade-off when setting the L parameter value. Very short windows do not provide enough frames to perform a robust estimation of the speaker factors, but very long windows do not provide resolution to detect short speaker turns, and the accuracy in the speaker boundary detection is reduced. The value of $L = 1second$ seems reasonable: it provides 100 frames for speaker factor estimation, and over one second, there are usually enough frames to capture phonetic variation and obtain an estimation of the speaker rather than of a particular phoneme. Shorter windows will present high variability in the speaker factor estimation for a single speaker since the window will contain a small set of phonemes, or even a single phoneme. On the other hand, longer windows may merge two speakers very often and the number of speaker factor vectors estimated on a single speaker will be reduced. Thus, the distributions obtained for every speaker will not be close to a Gaussian distribution anymore, and the presented approach will not operate properly.

It is interesting to notice the capability of the Resegmentation stage to correctly reassign the frames even when the output of the Core Segmentation was not very accurate. After the Core Segmentation, it is clear that the optimal value for L is $L = 1s$, but the L values surrounding this optimal value become competitive after Resegmentation. Note the improvement for $L = 2s$: although the accuracy output of the Core Segmentation is far from the output obtained for $L = 1s$, after Resegmentation both values get much closer. This is due to the higher resolution provided by the MFCC features considered in the Resegmentation stage.

We can expect that a higher number of speaker factors, for example, $R = 50$ as considered in the *heavy-weight* configuration, will require a longer window to be estimated properly.

However, given the that longer windows cannot segregate short speaker turns and will merge both speakers in several windows, introducing significant degradation (60% relative at the Core Segmentation Output in Table 4.3), and given the good accuracy obtained with the *heavy-weight* configuration in Table 4.2, we expect this $L = 1second$ to be optimal independently from the value of R.

The second and last parameter to analyze in the Front End stage is actually the dimension of the speaker factor vectors R. This parameter is studied for both configurations, since we expect that a higher dimension of the M_s space will enable us to find a higher number of directions of inter-speaker variability, and thus a higher optimal R value.

Number of speaker factors	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$R = 20$	**2.12%**	2014 (91.01%)	**2107 (95.21%)**
$R = 50$	2.14%	**2024 (91.46%)**	2100 (94.89%)
$R = 100$	2.16%	2023 (91.41%)	2101 (94.94%)

Table 4.4: Accuracy of the proposed diarization system with the *light-weight* configuration for several values of the speaker factor vector dimension. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system.

Number of speaker factors	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$R = 20$	1.89%	2032 (91.82%)	2116 (95.62%)
$R = 50$	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**
$R = 100$	1.87%	2052 (92.72%)	2122 (95.89%)

Table 4.5: Accuracy of the proposed diarization system with the *heavy-weight* configuration for several values of the speaker factor vector dimension. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system.

Tables 4.4 and 4.5 show the accuracy of the proposed speaker diarization system for $R = 20, 50, 100$, considering the *light-weight* and the *heavy-weight* configurations respectively. It can be seen that increasing the number of speaker factors does not always enable us to increase the accuracy of the diarization system. In fact, for the *light-weight* configuration, the results do not improve when increasing the dimension of the speaker factor vectors over $R = 20$, the accuracy seems to saturate at $R = 20$. For the *heavy-weight*, the best results are obtained for $R = 50$. As expected, having a higher dimension in the speaker supervector space ($C \times R$) enables the system to benefit from increasing R, but, again, increasing R over certain value does not give any improvement. In fact, in the case of the *heavy-weight* configuration, the system is less accurate for $R = 100$ than for $R = 50$. This due to the fact that the speaker factor vectors are estimated over a window that contains a fixed amount of data (length $L = 1s$). Thus, the higher number of speaker factors to estimate, the poorer the estimation. However, as it can be seen in Tables 4.4 and 4.5, considering a number of speaker factors higher than the optimal value does not introduce significant degradation.

Tables 4.4 and 4.5 also enables us to compare the improvement in the accuracy introduced by increasing the dimension of the speaker supervector space, considering the same number of speaker factors. Considering the *light-weight* configuration, the space where the speaker supervectors lie has a dimension of $CD = 256 \times 12 = 3072$, while for the *heavy-weight* configuration, the dimension of that space is of $CD = 1024 \times 38 = 38912$. For a fixed value of R, the *heavy-weight* configuration obtains higher accuracy. In addition, and as mentioned previously, the *heavy-weight* configuration can benefit from increasing the value of R.

4.2.3 Initial Clustering Configuration

One of the main contributions of the proposed speaker diarization system is the use of PCA+K-means approach for Initial Clustering, that was introduced in [Vaquero *et al.*, 2010a]. In the following study, we show the importance of the initialization (Initial Clustering stage) for this approach for speaker diarization and why the described PCA+K-means strategy is selected. This study is done considering the *heavy-weight* configuration, but the conclusions extracted apply for any other front-end configuration, including the *light-weight* configuration.

Firstly, we analyze the importance of the K-means module. For this purpose, we compare the best possible or Perfect initialization with a Random initialization for the speaker diarization system, analyzing how the use of K-means clustering affects the accuracy of the system. For the Perfect initialization, it is assumed that the actual diarization labels are available at the Initial Clustering stage. For the Random initialization, the speech frames are randomly assigned to two clusters, and these clusters are considered as initialization. To obtain results considering the Random initialization, several experiments are considered and the results averaged.

Initial Clust.	Init. Clust.	Complete system		
strategy	DER	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
Perfect	0.00%	0.53%	2196 (99.23%)	2211 (99.91%)
Random	49.65%	10.60%	1188 (53.68%)	1455 (65.70%)
Perfect + K-means	2.56%	1.55%	2082 (94.08%)	2143 (96.84%)
Random + K-means	2.96%	1.86%	2057 (92.95%)	2118 (95.71%)

Table 4.6: Accuracy of the diarization system considering Perfect and Random initializations. Accuracy is measured in terms of DER at the output of the Initial Clustering stage and in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system.

Table 4.6 shows the importance of using K-means in the initialization. If we do not consider K-means in the Initial Clustering stage, the Perfect initialization obtains 0% DER at the output of this stage while the Random initialization obtains 49.65% DER. The remaining stages slightly degrade the accuracy in the case of Perfect initialization, but increase the accuracy significantly in the case of Random initialization. Still the difference for both initializations is huge. Thus, the Initial Clustering stage is critical for the correct operation of the proposed diarization system.

The addition of K-means at the end of the Initial Clustering stage shows very interesting results. As it can be seen, feeding either the Perfect or a Random initialization into the

K-means algorithm gives very good accuracy at the output of the Initial Clustering stage. Now the difference in performance between Random and Perfect initializations has reduced significantly after K-means, and the subsequent stages will keep reducing the difference between both initializations.

So K-means gives a robust initialization, independently of how we initialize the algorithm itself. However, there is also an undesired behavior. K-means introduces some error at the initialization, even when considering Perfect initialization, and this error propagates to the output of the complete diarization system, obtaining a performance significantly worse when introducing K-means after the Perfect initialization than that obtained feeding the Perfect initialization directly into the Core Segmentation stage.

Nevertheless, the performance of the system when using K-means is still very good, and the robustness provided by such algorithm against poor initializations (such as Random initialization) ensures a good behavior of the system in most cases. So now the goal is to obtain an initialization for the K-means that gives an output accuracy as close as possible to that obtained using Perfect + K-means at the Initial Clustering stage.

For this purpose, we consider the following approaches for initialization:

- BIC AHC on the MFCC : A sliding window is used to segment the sequence of MFCC into small pure chunks, considering BIC as distance metric. These small chunks are agglomerated lately, using BIC as merge criterion, until two clusters are obtained. This procedure is identical to that considered in the baseline diarization system 3.4.1, but in this case no resegmentation is performed (it will be performed later, in the Resegmentation stage).

- BIC AHC on the speaker factors: the same procedure as explained before is performed on the sequence of speaker factors instead of on the MFCC vectors.

- Minimum Kurtosis direction: Since we assume that the speaker factors belonging to a single speaker follow a Gaussian distribution, the distribution of the speaker factors from two speakers will follow a bimodal distribution that in general will not be Gaussian, and more accurately will be platykurtic. So, finding the direction of minimum Kurtosis will give the best direction to separate both speakers. Once such direction is obtained, the speaker factors are projected onto that direction, obtaining a scalar for every frame, and K-means is used to obtain two initial clusters in the one-dimensional space, that will be fed into the K-means applied on the fully-dimensional speaker factors.

- PCA, maximum variability direction: Since it is assumed that the distributions of the speaker factor vectors for two different speakers have spherical and identical covariance matrices but different means, and the good performance of the K-means clustering seems to support this assumption, we can exploit it further. Given a sequence of speaker factors containing two different speakers, with the previous assumption, we can expect that the direction of maximum variability will be the one that best separates both speakers. So using PCA to obtain such direction and projecting the data onto that direction we obtain a sequence of scalars that are clustered lately using K-means into two clusters. These two clusters will serve as initialization for the K-means algorithm. The idea behind this approach has been shown graphically in Figure 4.3(a).

Initial Clust. strategy	Initial Clust.	Complete system		
	DER	DER	$\%_{DER<th}$	$DU(\Omega)$
BIC MFCC + K-means	**2.79%**	**1.77%**	**2063 (93.22%)**	2122 (95.89%)
BIC spk fact. + K-means	3.03%	1.89%	2051 (92.68%)	2116 (95.62%)
Min Kurtosis + K-means	3.06%	1.91%	2046 (92.45%)	2119 (95.75%)
PCA + K-means	2.88%	**1.77%**	2056 (92.91%)	**2125 (96.02%)**

Table 4.7: Accuracy of the diarization system considering different initializations for the K-means algorithm. Accuracy is measured in terms of DER at the output of the Initial Clustering stage and in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system.

Table 4.7 show the accuracy of the presented diarization system measured at the output of the Initial Clustering stage and at the output of the complete system, for the approaches proposed to initialize the K-means clustering. It can be seen that all approaches obtain similar accuracy, due to the robustness that the K-means algorithm introduces in the Initial Clustering stage. The initialization based on BIC and AHC on the speaker factors and the minimum Kurtosis direction do not seem to work: The accuracy at the output of the Initial Clustering stage is below that obtained considering Random initialization.

The two remaining approaches for initialization obtain similar performance. In some measures the BIC AHC initialization on the MFCC slightly outperforms the PCA intialization and vice-versa. However the differences are not significant to decide which one should be selected. The advantage of the BIC AHC initialization on the MFCC is its robustness. this initialization does not rely on a inter-speaker variability model, so if this model is poor, the initialization will not be degraded. However, the rest of the diarization system will not be able to benefit from a good initialization: K-means and Core Segmentation may suffer severe degradation if the inter-speaker variability model is poor.

In general we can expect to have enough data to train a robust inter-speaker variability model, specially when the diarization system is developed to help a speaker characterization system, that usually needs a huge amount of data to be trained properly. Thus, the PCA initialization seems to be better in this case, since further research and improvements in the inter-speaker variability model will translate into improvements in the initialization, and at some point, it may clearly outperform the BIC AHC initialization on the MFCC. In fact, in Chapter 5, different types of variability are studied in order to increase the separability between speaker factors belonging to different speakers. PCA can benefit from this, but the BIC AHC initialization on the MFCC cannot.

4.2.4 Core Segmentation Configuration

In this section the Core Segmentation stage is analyzed, focusing on two aspects of this stage: the importance of the iterative modeling and decoding process and the importance of temporal information in order to perform a better reassignment of the speaker factor vectors.

Table 4.8 shows the accuracy of the speaker diarization system considering the *light-weight* configuration. It can be seen that a single iteration in the Core Segmentation stage increases significantly the accuracy. Then, the iteratively process improves the results further, but the improvement is not that significant. The improvement introduced by the

Stage	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
Initial Clustering	4.79%	1708 (77.18%)	1936 (87.48%)
Core Seg. (1 iteration)	3.21%	**1889 (85.36%)**	2032 (91.82%)
Core Seg. until convergence	**3.02%**	**1889 (85.36%)**	**2052 (92.72%)**

Table 4.8: Accuracy of the proposed diarization system with the *light-weight* configuration, measured at the output of the Initial Clustering stage, after a single iteration in the Core Segmentation stage and at the output of the Core Segmentation stage. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Stage	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
Initial Clustering	2.88%	1930 (87.21%)	2079 (93.94%)
Core Seg. (1 iteration)	2.39%	**1996 (90.19%)**	**2096 (94.71%)**
Core Seg. until convergence	**2.37%**	**1996 (90.19%)**	2093 (94.58%)

Table 4.9: Accuracy of the proposed diarization system with the *heavy-weight* configuration, measured at the output of the Initial Clustering stage, after a single iteration in the Core Segmentation stage and at the output of the Core Segmentation stage. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Core Segmentation is mainly due to two reasons: first, the assumption of identity covariance considered in the Initial Clustering stage is relaxed, and in the Core Segmentation stage, a full covariance matrix is estimated for every Gaussian speaker model in the speaker factor space. The second reason is related to the introduction of temporal information in the decoding process. The HMM considered to model every speaker are intended to smooth the output labels of the Core Segmentation stage, removing those excessively short speaker turns that the Initial Clustering stage may introduce.

The behavior observed in Table 4.8 also applies to the *heavy-weight* configuration. For this configuration, the DER is reduced from 2.88% at the output of the Initial Clustering stage to 2.39% after a single iteration in the Core Segmentation stage and to 2.37% after convergence, as shown in Table 4.9. The relative improvement introduced by the Core Segmentation is lower in this case since the separability of the speaker factor vectors is much higher than for the *light-weight* configuration, and the Initial Clustering obtains very good performance on its own.

The improvement introduced by iterating in the Core Segmentation stage until convergence for both configurations is small or even non-significant in the case of the *heavy-weight* configuration. This is due to the fact that for most of the recordings, the Core Segmentation converges very fast (in 5 or 6 iterations). This means that the output obtained after the first iteration is usually very close to the output obtained after convergence. In fact, for the *heavy-weight* configuration more than half of the recordings converge before the sixth iteration.

To analyze the importance of the temporal information in the decoding process of the Core Segmentation, the topology of the speaker dependent HMM is analyzed. It is not our objective to be very exhaustive adjusting the topology and the parameters of the HMM so, in all cases, the probability of staying in a given state is 0.9, and thus the probability of leaving the state is 0.1. This probabilities have been selected since the sequence of speaker factors is very smoothed due to the 99% overlap in two consecutive windows. Therefore,

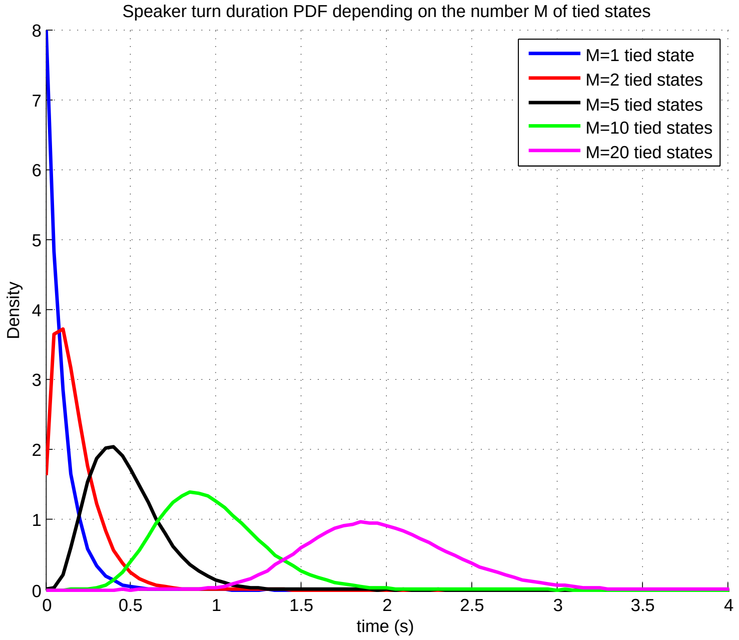


Figure 4.4: *Speaker turn duration PDF as funtion of the number of tied states (M). The transition probability considered is 0.1 for all states.*

only the number of tied-states in the HMM is analyzed.

Number of tied states	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$M = 1$	**2.37%**	1996 (90.19%)	**2093 (94.58%)**
$M = 2$	**2.37%**	1995 (90.15%)	**2093 (94.58%)**
$M = 5$	**2.37%**	1996 (90.19%)	**2093 (94.58%)**
$M = 10$	**2.37%**	1996 (90.19%)	**2093 (94.58%)**
$M = 20$	**2.37%**	**1998 (90.28%)**	2092 (94.53%)

Table 4.10: Accuracy of the proposed diarization system depending on the number of tied states (M) considered in the Core Segmentation stage. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the Core Segmentation stage, considering the *heavy-weight* configuration.

Table 4.10 shows the accuracy of the speaker diarization system considering several values for the number of tied states M, measured at the output of the Core Segmentation stage. Note that there is not a significant variation (in most cases, there is no variation at all) in the accuracy when modifying this parameter. This is because of the smoothness in the sequence of speaker factor vectors, due to the high overlap present in two consecutive estimation windows. Given these results, we can select any number of tied-states to model the HMM.

However, if we focus on the distribution of the speaker turn duration, not all topologies are reasonable. Figure 4.4 show the distributions of the speaker turn duration depending on the number of tied states for the transition probability considered. Note that for a single state, the distribution is geometric, which is not reasonable, since every speaker turn is expected to last more than a single frame (10 ms), and the mode of the distribution for $M = 1$ is a single frame. As the number of states increases, the speaker turn distribution becomes more realistic. However, M can not be increased as much as desired, since setting a value for M forces the decoding algorithm to stay in the same speaker M frames. If this value it high, short speaker turns cannot be modeled. As limit to this value we select $M = 20$, since we do not expect to obtain speaker turns with duration below 200 ms, but the duration of a short speaker turn may be close to 200 ms, for example in the case of a short turn to confirm that the conversation is being followed, which is typical in telephone conversations.

4.2.5 Resegmentation Configuration

The last step to analyze in the proposed diarization system is the Resegmentation stage. First, the dimension of the MFCC feature vectors is studied. Both the *light-weight* and *heavy-weight* configurations consider MFCC feature vectors of dimension 12 without delta features in the Resegmentation stage, but they consider different dimension of the MFCC feature vectors for the extraction of speaker factors. The *light-weight* configuration considers also 12 MFCC without delta features for speaker factor extraction, but the *heavy-weight* considers 19 MFCC plus delta features. The question is whether the additional MFCC and delta features considered in the *heavy-weight* configuration that has shown to be helpful for the extraction of speaker factors is useful for the Resegmentation stage.

Dimension of the MFCC feature vector	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
12 MFCC	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**
12 MFCC + Δ	2.09%	2030 (91.73%)	2111 (95.39%)
19 MFCC +	2.39%	1975 (89.25%)	2095 (94.67%)
19 MFCC + Δ	2.67%	1937 (87.53%)	2088 (94.35%)

Table 4.11: Accuracy of the proposed diarization system depending on the dimension of the MFCC feature vectors in the Resegmentation stage. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system, considering the *heavy-weight* configuration.

Table 4.11 compares the accuracy of the speaker diarization system considering the *heavy-weight* configuration, in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$, for four different configurations of the feature vectors in the Resegmentation stage: 12 MFCC, 12 MFCC plus delta features, 19 MFCC and 19 MFCC plus delta features. It is interesting to notice that increasing the number of MFCC from 12 to 19 degrades the results and so does the inclusion of delta features. Thus, among these configurations, the one providing the best accuracy is also the simplest and fastest: 12 MFCC without delta features.

But between 12 and 19 MFCC there is a wide range of dimensions for the feature vectors that may obtain higher accuracy than considering 12 MFCC. Figure 4.5 shows the overall DER obtained by the diarization system as a function of the dimension of the MFCC feature vector (delta features are not considered). It can be seen that the DER does not vary significantly for values of the dimension of the MFCC feature vectors between 9 and 12. From 13 MFCC, the DER increases as the number of MFCC increases. For 15 MFCC and more, the degradation is significant. We select the value of 12 for the dimension of the MFCC feature vectors since we do not expect a lower number of MFCC to obtain good accuracy in all conditions.

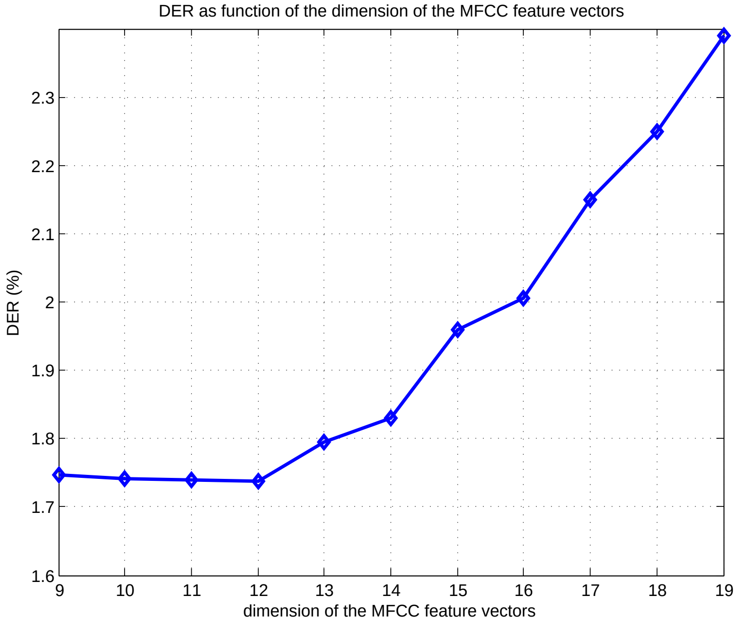


Figure 4.5: *DER obtained by the diarization system as function of the dimension of the MFCC feature vectors considered in the Resegmentation stage, for the heavy-weight configuration*

Since the Resegmentation stage is not directly affected by the number of speaker factors considered in the system, these conclusions extracted apply to any other configuration of the previous stages (Front End, Initial Clustering or Core Segmentation), including the *light-weight* configuration.

In the Resegmentation stage, the topology of the speaker dependent HMM plays an important role, as in the Core Segmentation stage. However, in this case, the sequence to be modeled by the HMM is composed of MFCC feature vectors, which evolve more abruptly than the speaker factors, since they are estimated over smaller windows (25 ms) and the estimation windows are less overlapped (60%). Thus, the probability of staying in a single state should be increased to avoid fast speaker changes.

In this study, we set the probability of staying to 0.999 for every state considered in the Resegmentation stage, and thus the probability of leaving a state is 10^{-3}. Again, we focus on determining the best topology for this task, only modifying the number of tied states considered.

Number of tied states	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$M = 1$	23.23%	7 (0.32%)	99 (4.47%)
$M = 2$	3.67%	1793 (81.02%)	2071 (93.58%)
$M = 5$	2.12%	2024 (91.46%)	2119 (95.75%)
$M = 10$	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**
$M = 20$	1.90%	2054 (92.82%)	2116 (95.52%)

Table 4.12: Accuracy of the proposed diarization system depending on the number of tied states (M) considered in the Resegmentation stage. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ at the output of the complete diarization system, considering the *heavy-weight* configuration.

Table 4.12 shows the accuracy of the complete diarization system depending on the number of tied states considered in the Resegmentation stage, considering the *heavy-weight* configuration. The number of tied states considered in the Core Segmentation stage is the same as the one considered in the Resegmentation stage, for the sake of coherence (in fact, Table 4.10 shows that for the values considered for M in this study, the accuracy at the output of the Core Segmentation does not vary significantly).

In this case, the number of tied states M considered affects significantly the accuracy of the diarization system. Small values of M ($M = 1, 2$) enables the decoding process to introduce very small speaker turns, and this is likely to happen when MFCC are considered as features, because they vary much more rapidly than speaker factors. This effect produces false short speaker turns that degrade the accuracy of the system. On the other hand, the need to increase the probability of staying in a state to mitigate this effect, penalizes severely those topologies containing many tied states. The cost of introducing short speaker turns is higher as M increases, and the actual speaker turns can be missed. In fact, for $M = 20$, the degradation is due to the missed short speaker turns. Thus, the value of $M = 10$ is selected.

Note that $M = 5$ or $M = 20$ could be good candidates if the probability of staying were adjusted for these topologies. In fact, the optimal probability of staying in every state for every value of M could be obtained, but this requires an exhaustive analysis that is not our intention, and the obtained configurations may be adapted to the development data. The value of $M = 10$ is selected because for a reasonable value of the staying probability provides the best accuracy. In addition, the decoding process is forced to stay in every speaker at least $M = 10$ states, which means that the speaker turn duration should be longer than 100 ms, which is reasonable, as mentioned in Section 4.2.4.

4.3 Evaluation

In Section 4.1, an approach for speaker diarization that makes use of a inter-speaker variability model to increase separability among speakers has been presented. This approach has been deeply studied, proposing and validating two configurations in Section 4.2. In this Section, the proposed speaker diarization system is compared to the traditional BIC AHC system described in Section 3.4.1. In addition, the speaker factor based diarization system is evaluated in a speaker verification task, considering the three scenarios that involve mono recordings containing two speakers, previously presented in Section 3.2.3.

4.3.1 Speaker Factor vs BIC AHC Diarization

The speaker factor based diarization system is compared to the traditional diarization system presented as baseline in Section 3.4.1. To compare both systems, the complete *summed* condition from the NIST SRE 2008 is evaluated. Both proposed configurations of the speaker factor based diarization system are considered for the evaluation.

Table 4.13 compares the accuracy obtained with the two diarization systems studied, including the two proposed configurations for the speaker factor based diarization system. It can be seen that the proposed approach for speaker diarization clearly outperforms the traditional BIC AHC in terms of DER. Therefore, the number of recordings obtaining a DER below the proposed values (5% and 10%) are also much higher. Note the difference

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	5.21%	1627 (73.52%)	1880 (84.95%)
Spk factors *light-weight* conf	2.12%	2014 (91.01%)	2107 (95.21%)
Spk factors *heavy-weight* conf	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**

Table 4.13: Comparison of the accuracy obtained with a traditional BIC AHC speaker diarization system and the proposed approach based on speaker factors. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$ obtained in the *summed* dataset from the NIST SRE 2008.

in the percentage of recordings with $DER < 10\%$, the BIC AHC system cannot obtain a DER below 10% for around 15% of the recordings in the dataset. The proposed system considering the heavy configuration, reduces this number to less than 4%. This means that an additional 11% of the dataset obtains a $DER < 10\%$. The impact of this fact in a speaker characterization task is analyzed in the next section.

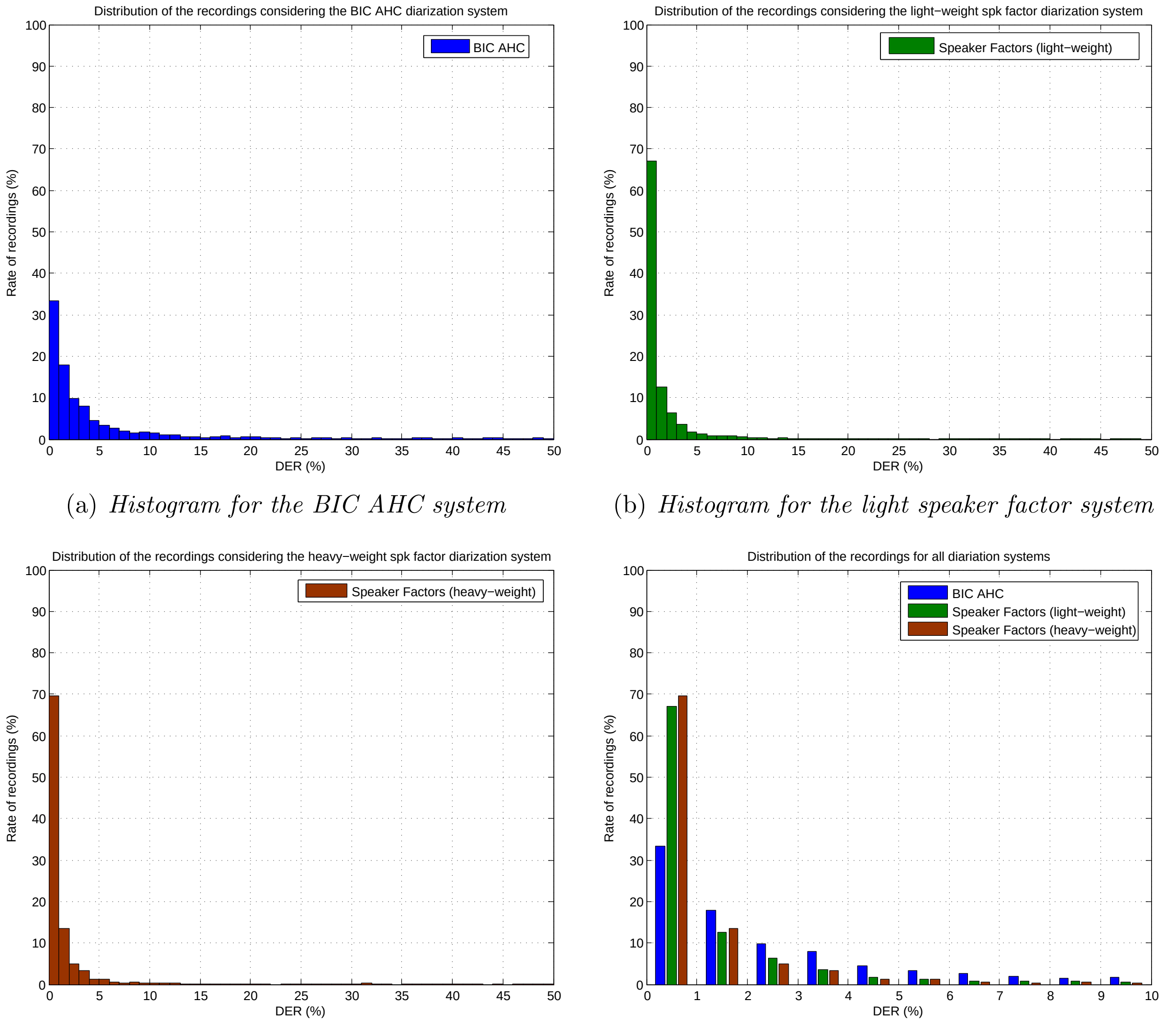
It is also interesting to study the distribution of the recordings according to the DER obtained, for the systems under analysis. Figure 4.6 shows the histogram of the recordings depending on their DER value, for each one of the systems and configurations considered. All histograms have a similar shape. Most recordings concentrate on the low DER values. In all cases the peak value is below 1% DER and as the DER increases, the number of recordings for the corresponding interval decreases. The decrease gets smoother as the DER increases. It can be seen that the number of recordings decreases faster for both configurations of the speaker factor based diarization system that for the baseline system. Also the concentration of recordings on the low DER values is much higher for the speaker factor system. It is interesting to note that both configurations of the proposed system obtain a DER below 1% for more that half of the recordings. Note also that the *heavy-weight* configuration obtains a slightly higher concentration of recordings below 1% and 2% than the *light-weight* configuration and the baseline system, but the concentration decreases faster as the DER increases, and for DER values over 3%, the concentration of recordings is higher for the baseline system than for both configurations of the proposed system.

4.3.2 Speaker Factor Diarization for Speaker Characterization

The improvement obtained by the speaker factor system with respect to the baseline system in terms of diarization should be reflected in a speaker characterization task that makes use of the diarization hypotheses generated by the former system. To analyze the improvement that the proposed diarization system introduces in a speaker characterization task, we use the same speaker verification task considered in Chapter 3 to analyze the baseline system. The experimental setup is described in Section 3.2, and from all tasks and scenarios proposed in that Section, only the 2 : 1 task in the three scenarios that involve mono conversations in either enrollment or testing sides are considered. These scenarios are the *mono-stereo*, *stereo-mono* and the *mono-mono*.

4.3.2.1 *Mono-Stereo* Scenario

Table 4.14 compares the accuracy obtained by the baseline, the *light-weight* and the *heavy-weight* speaker factor diarization systems on the subset of the *summed-short2* condition of



(a) *Histogram for the BIC AHC system*

(b) *Histogram for the light speaker factor system*

(c) *Histogram for the heavy speaker factor system*

(d) *Comparison of histograms: detail ($DER < 10\%$)*

Figure 4.6: *Distribution of the recordings according to the DER value obtained by each one of the systems under analysis.*

the NIST SRE 2008 considered for enrollment in the *mono-stereo* scenario. This subset is composed of 1359 recordings (see Table 3.2). Again it can be seen that the speaker factor diarization system is more accurate than the BIC AHC one, and that the *heavy-weight* configuration obtains higher accuracy than the *light-weight* configuration. Note that the degradation introduced by diarization errors in this scenario was accounted to be over 30% relative in terms of EER and over 40% in terms of $min(C_{norm})$ (see Table 3.10) for the set of recordings with $DER > 10\%$. Compared to the baseline, the heavy speaker factor system reduces the number of recordings obtaining a $DER > 10\%$ from 14.42% to 3.38%, which is more than a 70% relative reduction.

Figure 4.7 shows the DET curves obtained in the *mono-stereo* scenario considering the light and heavy configurations of the speaker factor system to diarize the enrollment dataset. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. It can be seen that both DET curves obtained considering the

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	4.92%	1021 (75.13%)	1163 (85.58%)
Spk factors *light-weight* conf	1.83%	1246 (91.69%)	1298 (95.51%)
Spk factors *heavy-weight* conf	**1.57%**	**1269 (93.38%)**	**1313 (96.62%)**

Table 4.14: Comparison of the accuracy obtained with a traditional BIC AHC speaker diarization system and the proposed approach based on speaker factors, evaluated on the enrollment subset of the *summed-short2* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

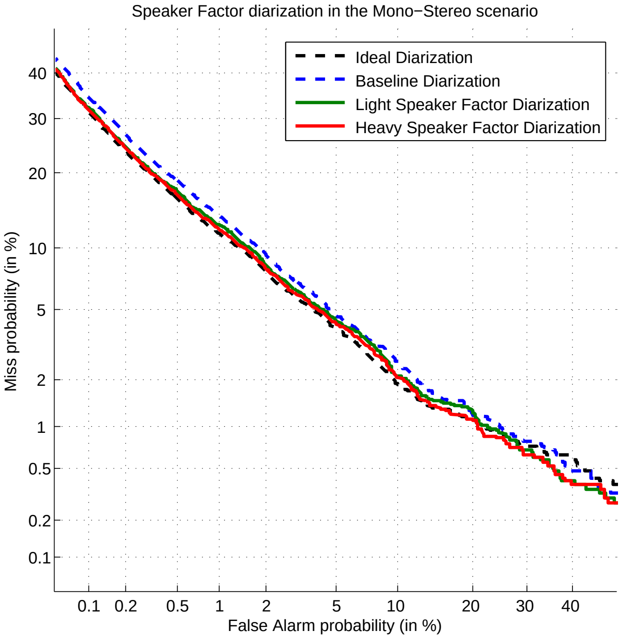


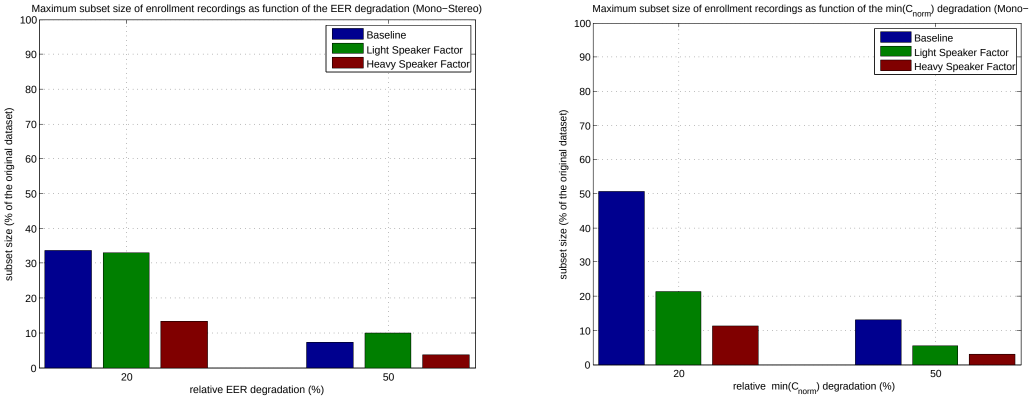
Figure 4.7: *DET curves considering the speaker factor diarization system with the light-weight and heavy-weight configurations in the mono-stereo scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

speaker factor diarization system are closer to the DET curve obtained with the ideal diarization system than that obtained with the baseline system. The curves obtained with the speaker factor system are specially close in the low false alarm region. As we expected, the speaker factor diarization system improves the accuracy of the speaker verification task with respect to the baseline system. This improvement and the small degradation in terms of $min(C_{norm})$ can also be observed in Table 4.15. However the improvement is not impressive. This is mainly because most of the dataset considered seems "easy" to diarize, and thus even the baseline diarization system obtains low DER for most recordings as it can be seen in Figure 4.6(a). Therefore, the accuracy of the speaker verification task considering the baseline system is close to that obtained with the ideal diarization system and there is not much margin to improve.

However, other environments may produce datasets mostly composed of recordings that are "hard" to diarize. Table 4.14, shows that the percentage of recordings with $DER < 10\%$ is increased from 85.58% considering the baseline system to 96.62% considering the *heavy-weight* speaker factor diarization system. This means that a 76.56% relative reduction of

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.40% (0.00%)	0.2042 (0.00%)
Baseline	4.76% (8.18%)	0.2295 (12.39%)
Light Speaker Factor	4.64% (5.45%)	0.2122 (3.76%)
Heavy Speaker Factor	**4.53% (2.95%)**	**0.2095 (2.59%)**

Table 4.15: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization systems with the *light-weight* and *heavy-weight* configurations in the *mono-stereo* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of* $min(C_{norm})$

Figure 4.8: *Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, in the mono-stereo scenario.*

recordings with $DER > 10\%$ is achieved (from 14.42% to 3.38%). Assuming a scenario where for most of the recordings, the baseline obtains a $DER > 10\%$, reducing this number a 76.56% will increase dramatically the accuracy of the speaker verification task.

Therefore, the improvement in the accuracy of speaker diarization enables us to retrieve more recordings from the given dataset that can be considered for the speaker verification task, as if they were correctly diarized. To show how this improvement in diarization affects the task of speaker verification in this scenario more clearly, the following analysis is proposed. For each of the studied diarization systems, the enrollment recordings are ranked according to their DER values. Then, the maximum size of a subset of the enrollment recordings with highest DER that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value is analyzed. As the subset size increases, it includes recordings with lower DER, and thus the degradation is reduced. The lower the maximum number of recordings that can be accounted to keep the degradation over certain value, the better performance, since the overall degradation will be less affected.

Figure 4.8 and Tables 4.16 and 4.17 show the percentage of recordings with highest DER of the enrollment dataset that can be accounted to keep the degradation in the EER (4.8(a)) and $min(C_{norm})$ (4.8(b)) with respect to the ideal diarization system over certain value. It is interesting to notice that the heavy speaker factor diarization system

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	33.61%	7.20%
Light Speaker Factor	32.92%	9.95%
Heavy Speaker Factor	**13.37%**	**3.77%**

Table 4.16: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the mono-stereo scenario.

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	50.75%	13.03%
Light Speaker Factor	21.26%	5.49%
Heavy Speaker Factor	**11.32%**	**3.09%**

Table 4.17: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the mono-stereo scenario.

accounts much fewer recordings that obtain a considerable degradation than the BIC AHC baseline and the light speaker factor system. The light speaker factor system obtains similar performance to the baseline in the EER operating point, and outperforms the baseline system in the $min(C_{norm})$ operating point. Note that these numbers can be interpreted as follows: assuming that the speaker verification application can tolerate a maximum degradation in terms of EER or $min(C_{norm})$, the results displayed in Tables 4.16 and 4.17 show the percentage of recordings in the enrollment dataset that should not be considered, which is related to the percentage of target speakers that will not work properly in the system. Therefore, assuming that the application works in the $min(C_{norm})$ operating point, the *heavy-weight* speaker factor diarization system miss around 4 times fewer enrollment recordings than the baseline diarization system when the maximum relative degradation allowed is of 20% or 50%.

4.3.2.2 *Stereo-Mono* Scenario

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	5.20%	1620 (73.54%)	1873 (85.02%)
Spk factors *light-weight* conf	2.13%	2004 (90.97%)	2097 (95.19%)
Spk factors *heavy-weight* conf	**1.78%**	**2046 (92.87%)**	**2115 (96.01%)**

Table 4.18: Comparison of the accuracy obtained with a traditional BIC AHC speaker diarization system and the proposed approach based on speaker factors, evaluated on the testing subset of the *summed* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Table 4.18 compares the accuracy obtained by the baseline, the *light-weight* and the *heavy-weight* speaker factor diarization systems on the subset of the *summed* condition of the NIST SRE 2008 considered for testing in the *stereo-mono* scenario. Once more, it can be seen that the speaker factor diarization system is more accurate than the BIC

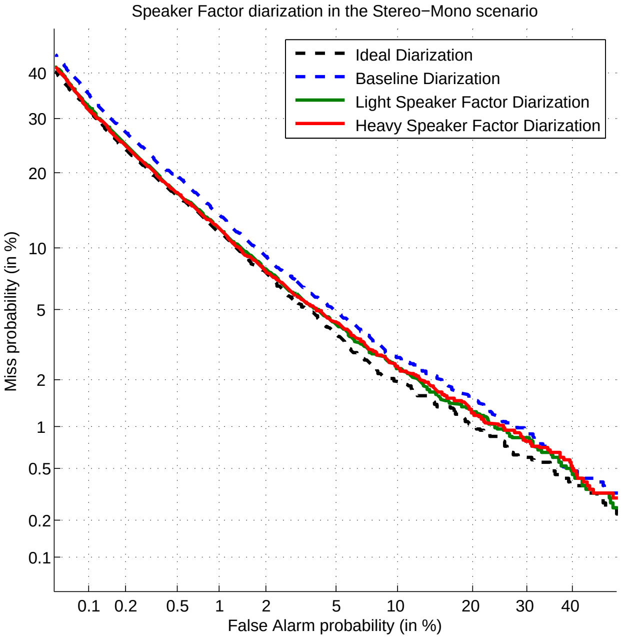


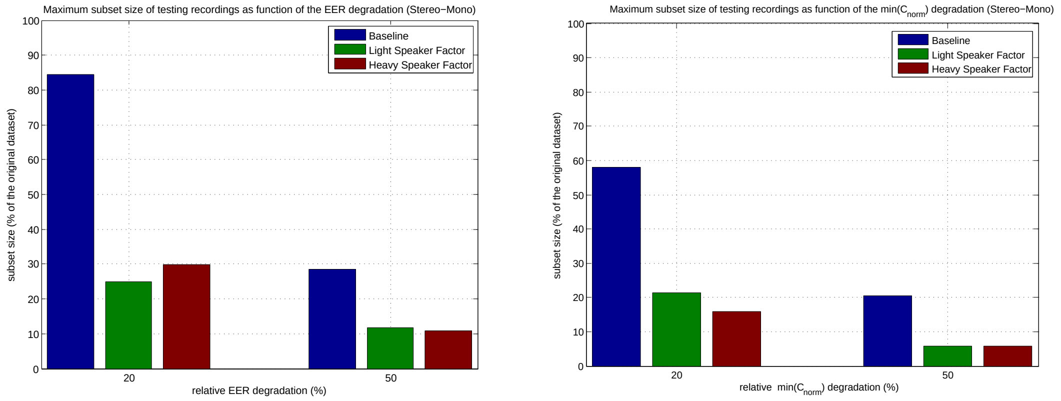
Figure 4.9: *DET curves considering the speaker factor diarization system with the light-weight and heavy-weight configurations in the stereo-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.23% (0.00%)	0.2102 (0.00%)
Baseline	4.94% (16.78%)	0.2334 (11.04%)
Light Speaker Factor	4.52% (6.86%)	0.2148 (2.19%)
Heavy Speaker Factor	**4.51% (6.62%)**	**0.2125 (1.09%)**

Table 4.19: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization systems with the *light-weight* and *heavy-weight* configurations in the *stereo-mono* scenario. The degradation with respect to the ideal diarization system is shown.

AHC one, and that the *heavy-weight* configuration obtains higher accuracy than the *light-weight* configuration. The degradation introduced by diarization errors in this scenario was accounted to be over 70% relative in terms of EER and over 60% in terms of $min(C_{norm})$ (see Table 3.13) for those recordings with $DER > 10\%$. The *heavy-weight* speaker factor system reduces the number of recordings obtaining a $DER > 10\%$ from 14.98% to 3.99%, which is more than a 70% relative reduction.

Figure 4.9 shows the DET curves obtained in the *stereo-mono* scenario considering the light and heavy configurations of the speaker factor system to diarize the testing dataset. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. Table 4.19 shows the accuracy of the speaker verification system in terms of EER and $min(C_{norm})$ for the four diarization systems, in the *stereo-mono* scenario. In this scenario we can extract similar conclusions to those obtained in the *mono-stereo* scenario, because of the symmetry in the score of the PLDA speaker verification system. Again the speaker factor diarization systems improve the results compared to the BIC AHC system, but the absolute improvement is not impressive since there is not much margin to



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of* $min(C_{norm})$

Figure 4.10: *Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, in the stereo-mono scenario.*

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	84.43%	28.60%
Light Speaker Factor	**24.97%**	11.80%
Heavy Speaker Factor	29.96%	**10.89%**

Table 4.20: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

improve. In this case, the curves are slightly more separated. Also, it can be seen again that the speaker factor diarization systems are very close to the ideal system in the false alarm region and a little further around the EER region. In fact, the degradation in terms of $min(C_{norm})$ of the heavy speaker factor system compared to the ideal system is insignificant (a 1.09% relative).

Figure 4.10 and Tables 4.20 and 4.21 show the percentage of recordings with highest DER of the testing dataset that can be accounted to keep the degradation in the EER (4.10(a)) and $min(C_{norm})$ (4.10(b)) with respect to the ideal diarization over certain value. In this scenario, both the *light-weight* and the *heavy-weight* configurations for the speaker factor diarization system accounts similar number of recordings that obtain a considerable degradation. Both systems account much fewer recordings than the BIC AHC baseline system. Thus, assuming that the speaker verification application works in the $min(C_{norm})$ operating point, both speaker factor systems can make use of three fourths of the testing recordings that the baseline diarization system would miss when the maximum relative degradation allowed is of 20% or 50%.

4.3.2.3 *Mono-Mono* Scenario

Finally, the *mono-mono* scenario is analyzed. This scenario make use of the *summed-short2* enrollment and *summed* testing subsets. The accuracy of the diarization systems evaluated on these datasets are shown in Tables 4.14 and 4.18. In this scenario, the differences in the

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	58.10%	20.43%
Light Speaker Factor	21.33%	**5.90%**
Heavy Speaker Factor	**15.89%**	**5.90%**

Table 4.21: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

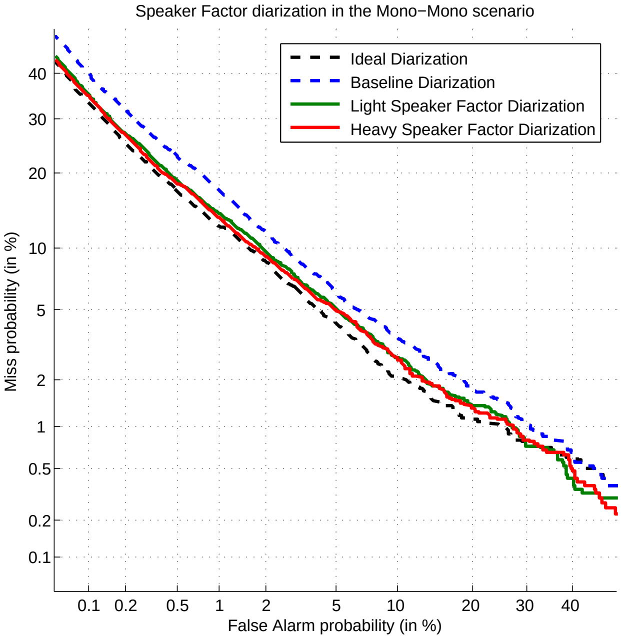


Figure 4.11: *DET curves considering the speaker factor diarization system with the light-weight and heavy-weight configurations in the mono-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

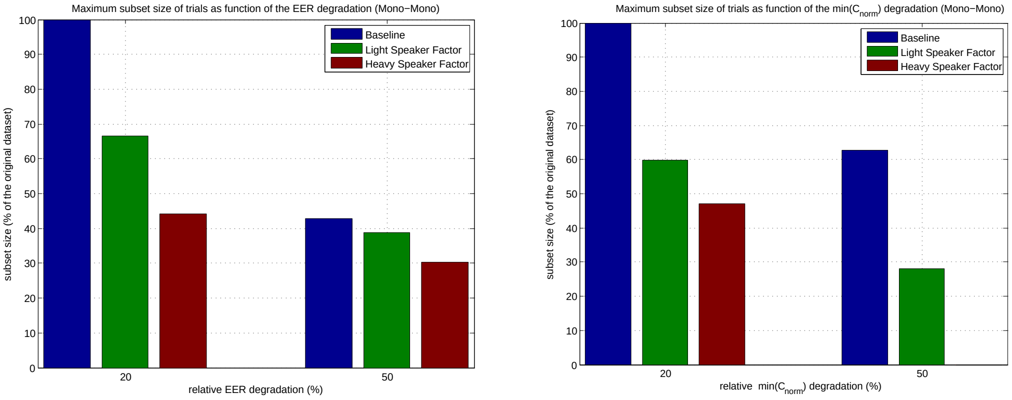
accuracy of the speaker diarization systems should be reflected more clearly in the speaker verification task, since diarization errors are introduced in both enrollment and testing sides.

Figure 4.11 shows the DET curves obtained in the *mono-mono* scenario considering the light and heavy configurations of the speaker factor system to diarize both the enrollment and testing datasets. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. Table 4.22 shows the accuracy of the speaker verification system in terms of EER and $min(C_{norm})$ for the four diarization systems, in the *mono-mono* scenario. In this case, the DET curves are more separated than in the previous scenarios, as it was expected, since diarization errors affects both enrollment and testing stages. Because of this, the speaker factor diarization system introduces more absolute improvement with respect to the baseline than in previous scenarios, but the relative improvement is similar to that obtained previously.

In the previous scenarios, the percentage of recordings with highest DER of a dataset that can be accounted to keep the degradation in the EER and $min(C_{norm})$ with respect to the ideal diarization over certain value has been analyzed. In this case, since both enrollment and testing recordings are processed by the diarization systems, we study the percentage

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.54% (0.00%)	0.2157 (0.00%)
Baseline	5.53% (21.81%)	0.2695 (24.94%)
Light Speaker Factor	5.03% (10.80%)	0.2318 (7.46%)
Heavy Speaker Factor	**4.99% (9.91%)**	**0.2289 (6.12%)**

Table 4.22: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization systems with the *light-weight* and *heavy-weight* configurations in the *mono-mono* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum number of trials depending on the degradation in terms of EER*

(b) *Maximum number of trials depending on the degradation in terms of $min(C_{norm})$*

Figure 4.12: *Percentage of trials in the mono-mono scenario that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value, considering the enrollment and testing recordings with highest DER, in the mono-mono scenario.*

of trials in the speaker verification task that involve recordings with high DER in both enrollment and testing sides. For this purpose, all the recording considered for enrollment and testing are pooled together and sorted according to their DER values. The subsets of trials are built considering only those recordings having the highest DER, and the DER threshold is reduced to increase the size of the subset of trials. Again, as the subset is bigger, the degradation will be reduced since recordings with lower DER are considered in the evaluation. Note that considering the percentage of trials enables us to compare the results in the previous scenarios with those obtained in this scenario, since in previous scenarios, the size of the subset considered for enrollment or testing is related to the number of trials in the speaker verification task.

Figure 4.12 and Tables 4.23 and 4.24 show the percentage of trials that can be accounted to keep the degradation in the EER (4.12(a)) and $min(C_{norm})$ (4.12(b)) with respect to the ideal diarization over certain value, considering those recordings with highest DER for enrollment and testing. In this scenario, it can be observed that the *heavy-weight* configuration for the speaker factor diarization system accounts fewer number of trials that obtain a considerable degradation than the *light-weight* configuration, and the later much fewer trials than the baseline system.

Note that, considering the heavy speaker factor diarization system, it is not possible to

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	100.00%	42.78%
Light Speaker Factor	66.50%	38.92%
Heavy Speaker Factor	**44.23%**	**30.49%**

Table 4.23: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	100.00%	62.80%
Light Speaker Factor	59.89%	**28.12%**
Heavy Speaker Factor	**47.02%**	**0.00%**

Table 4.24: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

find a subset of trials, even a small one, that obtains a degradation in terms of $min(C_{norm})$ over 50% relative when compared to the ideal diarization system. This surprising result is due to the fact that those recordings that are not correctly diarized by the heavy speaker factor system are also “hard” recordings for the speaker verification task. Thus, the diarization errors do not degrade dramatically the accuracy of the speaker verification task, since it was poor previously. This effect of correlation between the speaker verification and speaker diarization performance, that was also observed in Section 3.4.2, is specially strong in the speaker factor diarization system. This is probably due to the fact that the speaker factor diarization system uses approaches extracted from spekaer verification techniques, that are in fact similar to those techniques considered in this speaker verification task. This effect can be useful to predict the accuracy of a speaker verification task guessing the accuracy of a speaker diarization task. This is further studied in Chapter 7.

5

Variability in Speaker Diarization

In this Chapter we analyze the different types of variability involved in the speaker diarization process, aiming at compensating for the harmful sources of variability. In this study, variability is compensated in the space of speaker factors rather than in the MFCC space, since in the former space the sources of variability are easier to separate.

5.1 Types of variability

The speaker factor based diarization approach presented in Chapter 4 makes use of an inter-speaker variability model to obtain a representation of the speakers that enables easier separability among them. However, there exist other sources of variability, such as the channel, the evolution of the conversation, or the speaker mood. The variability introduced by these and other sources that are not directly related to the presence of different speakers in a recording or a dataset are usually referred as intra-speaker variability.

The intra-speaker variability comprises a wide range of sources of variability. In this work, we classify the intra-speaker variability into two types of variability: inter-session and intra-session variability. According to this classification, three types of variability may affect a speaker diarization system. inter-speaker, inter-session and intra-session variability.

To explain these types of variability, let us assume that a dataset of recordings is available. The dataset contains several speakers, and for every speaker, there are several recordings available. Every recording contains a single speaker. The sources of variability present is this dataset fall into the three mentioned types of variability, which are displayed in Fig. 5.1 and described below.

5.1.1 Inter-Speaker Variability

Inter-Speaker Variability refers to the variability present among several recordings or sets of recordings that contain different speakers (see Figure 5.1). Modeling this type of variability enables us to obtain compact speaker representations (speaker factors), which are suitable for speaker diarization, as presented in Chapter 4. In fact, the proposed speaker factor diarization system makes use solely of this type of variability obtaining very satisfactory results.

Nevertheless, the other types of variability can affect the accuracy of a speaker diarization system, if they are not modeled or compensated properly.

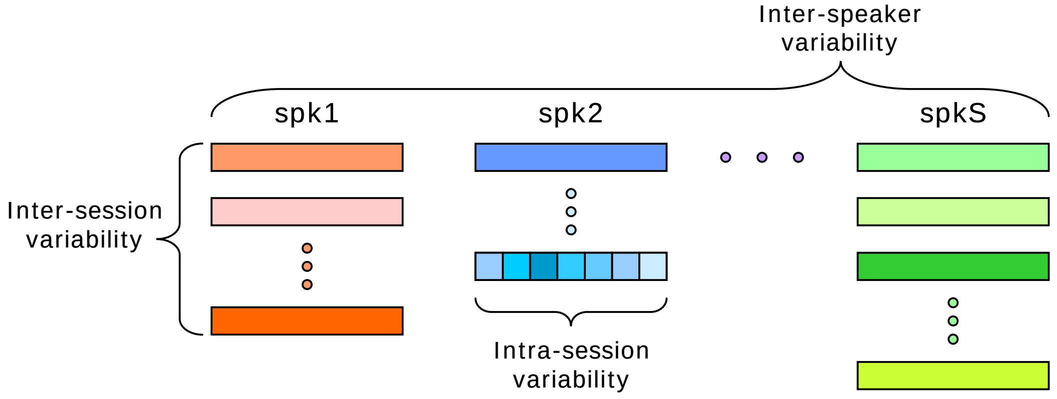


Figure 5.1: Types of variability

5.1.2 Inter-session Variability

In speaker recognition systems, one of the hardest problems is to deal with the variability present in a speaker recorded over different sessions. This is known as inter-session variability (see Figure 5.1) and includes variability due to the speaker, since her/his speech may vary along different recording sessions, as well as variability due to the recording environment.

There are several techniques to model this variability. Some of the more recent and successful approaches have been Nuissance Attribute Projection (NAP) for SVM-GMM speaker recognition systems [Campbell *et al.*, 2006], Eigenchannel modeling and JFA [Kenny *et al.*, 2007]. Most of these techniques assume that the speaker is modeled by a supervector in a high dimension space (usually a GMM supervector, as in our factor analysis model for speaker diarization) and different sessions for a given speaker produce different estimations of the speaker supervector. The variability in these estimations or inter-session variability is assumed to lie in a low dimension subspace, so all inter-session variability compensation techniques try to estimate the component of the speaker session in such space and remove it to obtain a session independent speaker supervector.

The question is whether inter-session variability compensation is useful for speaker diarization. Speaker diarization systems aim at answering the question “Who spoke when?” in an unsupervised fashion. In other words, no prior knowledge of the speakers involved in a conversation is available. Thus, it is not possible to find the same speaker over different sessions. Therefore, the compensation for inter-session variability is not expected to help the task of speaker diarization.

On the other hand, in many scenarios, session variability models may enhance diarization accuracy since different speakers may use different communication channels. This is the case of telephone conversations or meetings in a room where the speakers remain static.

5.1.3 Intra-Session Variability

In addition, a single speaker can present variability during a single session when we process such session in small segments. We will refer to this variability as intra-session variability (see Figure 5.1). Some examples of this variability includes the mood or excitement of the speaker as the conversation evolves, or the unbalanced phonetic load present that may appear in small segments as those considered in the speaker factor diarization system (1 second length).

Some works [Vogt *et al.*, 2009] have studied the importance of Intra-session or *Within-session* variability for speaker recognition, showing that compensating for it properly can provide robustness when the utterance lengths are short and vary. However, intra-session variability is not usually taken into account for speaker recognition, since state of the art systems usually integrate over all observations of a given speaker obtaining an average model, which may differ from session to session. In such case intra-session variability modeling and compensation will only be useful as far as it is related to inter-session variability. Actually, both intra and inter-session variability share many sources of variability, but some of them are more critical than others. For example, channel is a source of inter-session variability that in general does not introduce intra-session variability (however, if a speaker is recorded in a room with a far field microphone and he moves as he talks, channel will introduce intra-session variability). On the other hand, unbalanced phonetic load will be more critical for intra-session variability modeling, specially as the segments to analyze in a given session become smaller.

Although it is not usual to compensate for intra-session variability in speaker recognition problems, this type of variability is very important and should be taken into account in the task of speaker diarization. Most speaker diarization systems (including the one proposed in this thesis) analyze small and pure segments and then try to agglomerate them to obtain pure clusters that should belong to a single speaker. All the existing variability among these segments that is not due to the presence of different speakers is undesired and may mislead the clustering process. Thus intra-session variability should be compensated to ensure the correct behavior of the clustering algorithm. The importance of compensating for intra-session variability in speaker diarization has been studied in [Vaquero *et al.*, 2011a].

5.2 Intra-session Variability Compensation

Between inter-session and intra-session variability, the later is probably the one that affect a speaker diarization system the most. In this section, two methods to compensate for this type of variability when performing speaker diarization are proposed. These methods assume that the speaker factor diarization system is considered.

Let us assume that a set of N recordings is available and each recording contains a single speaker. Thus, we can obtain a sequence $Y_n = \{m_{y_n(1)}, ..., m_{y_n(T_n)}\}$ of T_n speaker factor vectors for every recording session n. Note that $m_{y_n(t)}$ is the point estimate of $y_n(t)$ (see Section 4.1). The speaker factor vectors obtained from a session belong to the same class (same speaker in the same session), so we can study the inter-session and inter-speaker variability as between-class variance and the intra-session variability as within-class variance. This approach is similar to the one presented in [Dehak *et al.*, 2010], but in that case it was used for speaker recognition and inter-session variability compensation.

Note that we are assuming that every class is not a single speaker but a single speaker in a fixed recording session. This means that we try to keep inter-session variability in our speaker representations, i.e. we do not compensate for inter-session variability, and we try to remove or compensate only for intra-session variability. This procedure is done according to our previous discussion, where we put forward that inter-session variability may help to separate speakers in a conversation. In fact, the information of the environment and the channel that will be present in the features considered for speaker diarization can be helpful to separate speakers since, in general, different speakers will use different communication channels.

Turning inter-session and inter-speaker variability into between class variance and intra-session variability into within class variance enables us to consider well known techniques to enhance class separability, such as Linear Discriminant Analysis or Within Class Covariance Normalization, which are described below.

5.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a technique for dimensionality reduction that, given a set of features belonging to different classes, finds the orthogonal basis to represent the features that enables better discrimination between different classes by maximizing between-class variance and minimizing within class variance. Linear discriminant analysis assumes that the observations belonging to each class are normally distributed and that within class covariance is kept across different classes. The speaker factor vectors satisfy the first assumption, while the second is expected to be satisfied since we expect the posterior covariance of $y(t)$ to be close to the prior (identity matrix I) as explained in Chapter 4.

In our problem we estimate between-class covariance (S_b) and within class covariance (S_w) as:

$$S_b = \frac{1}{N-1}\sum_{n=1}^{N}(\mu_n - \mu)(\mu_n - \mu)' \tag{5.1}$$

$$S_w = \frac{1}{N-1}\sum_{n=1}^{N}\frac{1}{T_n-1}\sum_{t=1}^{T_n}(m_{y_n(t)} - \mu_n)(m_{y_n(t)} - \mu_n)' \tag{5.2}$$

$$\mu_n = \frac{1}{T_n}\sum_{n=1}^{T_n} m_{y_n(t)} \tag{5.3}$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N}\mu_n \tag{5.4}$$

and thus the problem reduces to finding the matrix A of eigenvectors that satisfies:

$$S_b A = \lambda S_w A, \tag{5.5}$$

and project the speaker factors onto A or onto a low rank matrix U obtained selecting those eigenvectors of A having higher eigenvalues, for dimensionality reduction.

5.2.2 Within Class Covariance Normalization

Within Class Covariance Normalization (WCCN) is a normalization method that enables us to obtain a linear transformation for a given set of features belonging to different classes so that the within class covariance matrix S_w is equal to the identity matrix I. Again this technique assumes that all classes share the same covariance matrix, so a single linear transformation can turn this covariance matrix into the identity I for all the classes.

To obtain the linear transformation we apply Cholesky decomposition to S_w^{-1}, so the transformed speaker factors m'_y will follow this expression:

$$m'_y = W m_y \tag{5.6}$$

$$S_w^{-1} = W^T W \tag{5.7}$$

where W is the upper triangular matrix obtained by Cholesky decomposition.

5.2.3 Performance Evaluation

In order to analyze the improvement introduced by compensating for intra-session variability, the speaker factor diarization system described in Chapter 4 is considered. Both LDA and WCCN are evaluated on the *heavy-weight* configuration described in Section 4.2.1. In addition, different dimensions of the speaker factor vectors are considered for each configuration, in order to analyze the improvement introduced by LDA when applying dimensional reduction. As experimental dataset, the complete NIST SRE 2008 *summed* dataset (see Section 3.2.1) is utilized.

Intra-session compensation	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
$R = 20$	1.89%	2032 (91.82%)	2116 (95.62%)
$R = 50$	1.77%	2056 (92.91%)	2125 (96.02%)
$R = 100$	1.87%	2052 (92.72%)	2122 (95.89%)
LDA 50 → 20	1.74%	2044 (92.36%)	2131 (96.29%)
WCCN 20	1.65%	2062 (93.18%)	2137 (96.57%)
LDA 50 → 20 + WCCN	1.54%	2080 (93.99%)	2145 (96.93%)
LDA 100 → 50	1.51%	2088 (94.35%)	2145 (96.93%)
WCCN 50	1.38%	**2113 (95.48%)**	2155 (97.38%)
LDA 100 → 50 + WCCN	**1.31%**	2108 (95.26%)	**2158 (97.51%)**
WCCN 100	1.40%	2106 (95.16%)	2150 (97.15%)

Table 5.1: Accuracy of the *heavy-weight* speaker factor diarization system using intra-session variability compensation. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Table 5.1 analyzes the proposed intra-session variability compensation techniques on the *heavy-weight* speaker factor diarization system with $R = 20$, $R = 50$ and $R = 100$. Several interesting conclusions can be extracted from this results. Firstly, as previously observed in Section 4.2.2, we cannot improve the accuracy by simply increasing the number of speaker factors R. If we do not consider intra-session variability compensation, the results are better for $R = 50$ than for $R = 20$ but also than for $R = 100$, probably because we use a small fixed window (1 second length) to estimate the R speaker factors so the estimation will be noisier as R increases.

Analyzing the results obtained when using intra-session variability compensation, both LDA and WCCN increase the accuracy of the diarization system. Using LDA to reduce the dimensionality from $R = 100$ to $R = 50$ reduces the DER to 1.51%, a value that is lower than those obtained for any value of R without intra-session variability compensation. Using WCCN reduces the DER in all cases. This technique enables us to reduce the DER to 1.38% when $R = 50$. From this results, we can conclude that WCCN obtains better performance than LDA for the proposed diarization system.

The use of LDA for dimensionality reduction also shows an increase in accuracy in all cases, but it is always better to apply WCCN after dimensionality reduction. Actually, it seems that the best strategy is to estimate a large number of speaker factors, reduce to the desired dimensionality using LDA and then apply WCCN. This is in all cases better than estimating directly the speaker factors with the desired dimensionality and applying WCCN to them. It is interesting to note that the best configuration is actually to extract $R = 100$ speaker factors, use LDA to reduce the dimensionality to 50 and finally apply WCCN, obtaining a DER of 1.31%, even though the baseline for $R = 100$ obtained worse performance than for $R = 50$.

Summing up, we have seen that using WCCN for intra-session variability compensation always improves the performance of the system, when applied directly on the speaker factors or after performing LDA dimensionality reduction. This is probably due to the fact that WCCN is quite suitable for our speaker diarization system, since it transforms the speaker factors so they are closer to fulfill the assumption described in Section 4.1.2: it is assumed that, for a given speaker, the speaker factors follow a Normal distribution with the identity as covariance matrix.

On the other hand, we have seen that LDA can be useful when high dimension speaker factors are extracted, since they may be noisy before performing dimensionality reduction, but after dimensionality reduction, the additional information provided by increasing R is kept, and the noise removed, obtaining improved performance. However, LDA may not be useful if the initial number of speaker factors is small, for example, 50 or fewer. Whether to use LDA or not will be defined by the application requirements. Since the computational cost of extracting R speaker factors is $O(CDR+R^2)$, where D is the dimension of the feature vectors and C the number of components of the UBM, if the cost is critical, keeping R low and using WCCN will give very good performance. On the other hand, if computational cost is not critical, increasing R and using LDA for dimensionality reduction and then WCCN will give better results.

These approaches for intra-session variability compensation have been presented in [Vaquero *et al.*, 2011a].

5.3 Inter-session Variability Compensation and Modeling

In our approach for intra-session variability compensation, we have considered that two sequences of speaker factor vectors Y_m, Y_n, obtained from two different sessions m, n, belong to different classes, even if both sessions contain the same speaker s. It is known that inter-session variability is unavoidably captured when training the eigenvoice matrix V and as mentioned before, this variability may be helpful for speaker diarization.

But we known that this variability is present when training V and we can try to take advantage of it, either compensating it or modeling it properly. We can compensate for both inter and intra-session variability simply considering that when training intra-session variability compensation, every speaker is a different class, and that the speaker factor sequences Y_m and Y_n belong to the same class as far as sessions m and n contain the same speaker. LDA and WCCN transformations can be easily reformulated substituting n by s. This way, LDA minimizes the inter-session and intra-session variability and maximize the speaker variability, while WCCN forces the covariance of the speaker factors to be the identity over all training sessions of a single speaker, but not for every single session.

On the other hand, following the assumption that inter-session variability actually helps to separate speakers in a diarization task, we can capture inter-speaker and inter-session variability in V to take advantage of that effect. For this purpose, we train another V matrix, capturing the variability present among all available sessions. This variability is due to both inter-speaker and inter-session variability, and it is known as total variability. The factors extracted with this variability model are not speaker factors anymore, but total variability factors, also known as i-vectors [Dehak *et al.*, 2010] in the field of speaker recognition.

Configuration	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
intra-session WCCN $R = 50$	**1.38%**	**2113 (95.48%)**	**2155 (97.38%)**
inter/intra-session WCCN $R = 50$	1.50%	2100 (94.89%)	2141 (96.75%)
spk factors $R = 50$	**1.77%**	**2056 (92.91%)**	**2125 (96.02%)**
i-vectors $R = 50$	1.98%	2043 (92.32%)	2115 (95.57%)

Table 5.2: Accuracy of the proposed diarization system compensating and modeling inter-session variability

Table 5.2 shows the accuracy obtained by the proposed *heavy-weight* speaker diarization system, for different configurations. The first two entries of the table (intra-session WCCN and inter/intra-session WCCN) compare the results obtained when considering WCCN for intra-session variability compensation with those obtained considering the same technique for inter and intra-session variability compensation. it can be seen that there is degradation in accuracy when compensating for inter-session variability in addition to intra-session variability, so, as we expected, it does not seem interesting to compensate for inter-session variability in this task.

The last two entries compare the accuracy of the speaker diarization system when V is trained capturing inter-speaker variability (spk factors) and when V is trained capturing inter-speaker and inter-session variability (i-vectors). The accuracy degrades when modeling inter-session variability, probably because many eigenvectors in V are modeling mostly this variability instead of inter-speaker variability, reducing the separability among different speakers in the subspace generated by V.

Given these results, we can conclude that the best approach is to model inter-speaker variability with our V matrix and to compensate only for intra-session variability. However, the degradation introduced by compensating for or modeling inter-session variability is not dramatic.

5.4 Evaluation in Speaker Verification

The proposed techniques for intra-session variability compensation enables us to reduce the DER of the speaker factor diarization system from 1.77% to 1.31%. In this section we analyze how this improvement in terms of diarization accuracy is reflected in the speaker verification task defined in Chapter 3. The same experimental setup described in Section 3.2 is utilized and the verification task is evaluated in the three scenarios considered to analyze the speaker factor system in Section 4.3.2: the *mono-stereo*, *stereo-mono* and the *mono-mono* scenarios.

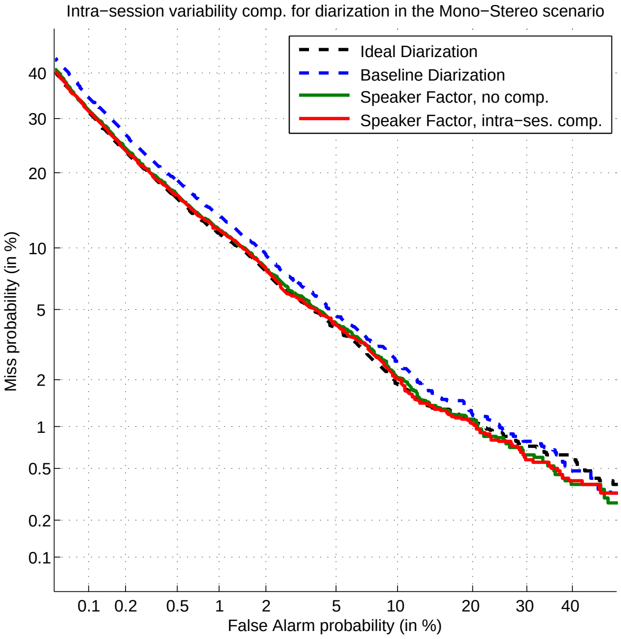


Figure 5.2: *DET curves considering the speaker factor diarization system with and without intra-session variability compensation in the mono-stereo scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

5.4.1 *Mono-Stereo* Scenario

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	4.92%	1021 (75.13%)	1163 (85.58%)
Speaker Factors no comp.	1.57%	1269 (93.38%)	1313 (96.62%)
Speaker Factors intra-ses. comp.	**1.11%**	**1307 (96.17%)**	**1335 (98.23%)**

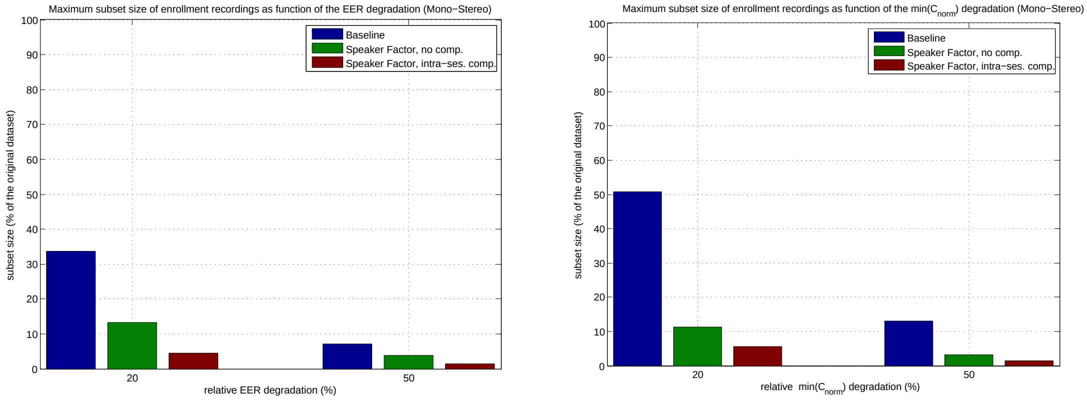
Table 5.3: Comparison of the accuracy obtained with the BIC AHC and the *heavy-weight* speaker factor diarization systems, with and without intra-session variability compensation, evaluated on the enrollment subset of the *summed-short2* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Table 5.3 compares the accuracy obtained by the baseline system, and the speaker factor diarization system with and without intra-session variability compensation, on the subset of the *summed-short2* condition of the NIST SRE 2008 considered for enrollment in the *mono-stereo* scenario. These results confirm the improvement introduced by the intra-session variability compensation. Note that the compensation for intra-session variability enables us to reduce the rate of recordings obtaining a $DER > 10\%$ from 3.38% to 1.77%, which is a reduction of more than 40%.

Figure 5.2 shows the DET curves obtained in the *mono-stereo* scenario considering the heavy configuration of the speaker factor system to diarize the enrollment dataset, with and without intra-session variability compensation. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. Note that the speaker factor system itself obtains little degradation with respect to an ideal diarization

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.40% (0.00%)	0.2042 (0.00%)
Baseline	4.76% (8.18%)	0.2295 (12.39%)
Speaker Factors no comp.	**4.53% (2.95%)**	**0.2095 (2.59%)**
Speaker Factors intra-ses. comp.	**4.49% (2.05%)**	**0.2074 (1.57%)**

Table 5.4: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system with and without intra-session variability compensation in the *mono-stereo* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of* $min(C_{norm})$

Figure 5.3: *Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, in the mono-stereo scenario.*

system. Introducing intra-session variability improves the results, but the improvement does not seem significant in the curves or the EER and $min(C_{norm})$ as observed in Table 5.4. However, it has been shown that the rate of recordings with $DER > 10\%$ is reduced in more than a 40%. Such a reduction would be clearly reflected in the speaker verification results if the initial rate of recordings with $DER > 10\%$ were much higher.

In fact, Figure 5.3 and Tables 5.5 and 5.6 show the percentage of recordings with highest DER of the enrollment dataset that can be accounted to keep the degradation in the EER (Figure 5.3(a)) and $min(C_{norm})$ (Figure 5.3(b)) with respect to the ideal diarization system over certain value. It can be observed that the use of intra-session variability compensation enables us to halve these percentages, when the maximum relative degradation allowed in terms of EER or $min(C_{norm})$ is 20% or 50%.

5.4.2 *Stereo-Mono* Scenario

Table 5.7 compares the accuracy obtained by the baseline system and the *heavy-weight* speaker factor diarization system, with and without intra-session variability compensation, on the subset of the *summed* condition of the NIST SRE 2008 considered for testing in the *stereo-mono* scenario. As observed before, intra-session variability compensation increases the accuracy of speaker diariation significantly. This technique reduces the percentage of

Diarization system	$\%_{Degrad(EER)} \geq 20\%$	$\%_{Degrad(EER)} \geq 50\%$
Baseline	33.61%	7.20%
Speaker Factors no comp.	13.37%	3.77%
Speaker Factors intra-ses. comp.	**4.46%**	**1.37%**

Table 5.5: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

Diarization system	$\%_{Degrad(min(C_{norm}))} \geq 20\%$	$\%_{Degrad(min(C_{norm}))} \geq 50\%$
Baseline	50.75%	13.03%
Speaker Factors no comp.	11.32%	3.09%
Speaker Factors intra-ses. comp.	**5.49%**	**1.37%**

Table 5.6: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

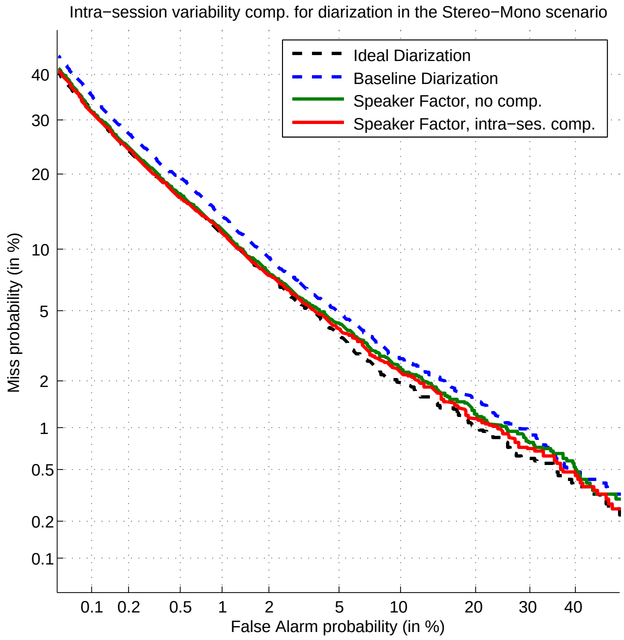


Figure 5.4: *DET curves considering the speaker factor diarization system with and without intra-session variability compensation in the stereo-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

recordings obtaining a $DER > 10\%$ from 3.99% to 2.45%, which is more than a 30% relative reduction.

However, this reduction is not enough to show significant improvement in the speaker verification task, as shown in Figure 5.4 and Table 5.8. Figure 5.4 shows the DET curves obtained in the *stereo-mono* scenario considering the speaker factor system with and without intra-session variability compensation to diarize the testing dataset. Table 5.8 shows the accuracy of the speaker verification system in terms of EER and $min(C_{norm})$ for these diarization systems. In this scenario we can extract similar conclusions to those

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	5.20%	1620 (73.54%)	1873 (85.02%)
Speaker Factors no comp.	1.78%	2046 (92.87%)	2115 (96.01%)
Speaker Factors intra-ses. comp.	**1.32%**	**2098 (95.23%)**	**2149 (97.55%)**

Table 5.7: Comparison of the accuracy obtained with the BIC AHC and the *heavy-weight* speaker factor diarization systems, with and without intra-session variability compensation, evaluated on the testing subset of the *summed* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.23% (0.00%)	0.2102 (0.00%)
Baseline	4.94% (16.78%)	0.2334 (11.04%)
Speaker Factors no comp.	4.51% (6.62%)	0.2125 (1.09%)
Speaker Factors intra-ses. comp.	**4.39% (3.78%)**	**0.2097 (-0.24%)**

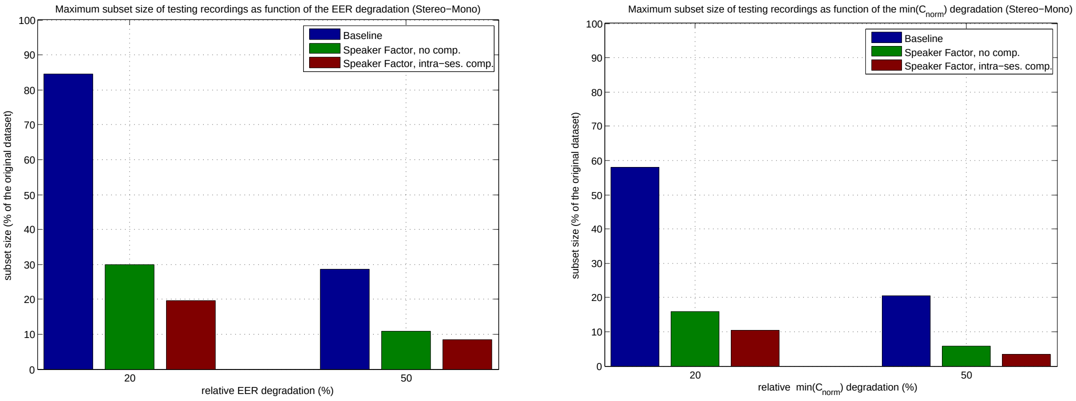
Table 5.8: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system with and without intra-session variability compensation in the *stereo-mono* scenario. The degradation with respect to the ideal diarization system is shown.

obtained in the *mono-stereo* scenario, because of the symmetry in the score of the PLDA speaker verification system. Again the improvement introduced by intra-session variability compensation is not significant, since the percentage of recordings that actually degrade the accuracy of the speaker verification task is very low without intra-session variability compensation. A further reduction is not reflected in these results. Note that the relative degradation in terms of $min(C_{norm})$ becomes negative for the speaker factor system whith intra-session variability compensation. This does not mean that we can outperform the ideal diarization system, but that the degradation is so insignificant that a much larger dataset will be needed to measure it.

Finally, Figure 5.5 and Tables 5.9 and 5.10 show the percentage of recordings with highest DER of the testing dataset that can be accounted to keep the degradation in the EER (Figure 5.5(a)) and $min(C_{norm})$ (Figure 5.5(b)) with respect to the ideal diarization over certain value. These statistics show that actually intra-session variability compensation enables us to reduce the size of the dataset that degrades the accuracy of the speaker verification system over certain value, thus increasing the number of recordings useful for speaker verification. This reduction would be reflected in the overall speaker verification results if the percentage of recordings with $DER > 10\%$ were much higher.

5.4.3 *Mono-Mono* Scenario

Finally, the *mono-mono* scenario is analyzed. This scenario makes use of the *summed-short2* enrollment and *summed* testing subsets. The accuracy of the speaker factor diarization system with and without intra-session variability compensation, evaluated on these datasets are shown in Tables 5.3 and 5.7. This scenario usually reflects with more clarity the differences in accuracy of the speaker diarization systems in the speaker verification task, since diarization errors are introduced in both enrollment and testing sides. However,



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of $min(C_{norm})$*

Figure 5.5: *Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the stereo-mono scenario.*

Diarization system	$\%_{Degrad(EER)} \geq 20\%$	$\%_{Degrad(EER)} \geq 50\%$
Baseline	84.43%	28.60%
Speaker Factors, no comp.	29.96%	10.89%
Speaker Factors, intra-ses. comp.	**19.52%**	**8.40%**

Table 5.9: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

given the high accuracy obtained in these datasets by the speaker factor diarization system without intra-session variability compensation, we do not expect the compensation for intra-session variability to introduce significant improvement in the overall results of the speaker verification task.

Figure 5.6 shows the DET curves obtained in the *mono-mono* scenario considering the speaker factor system with and without intra-session variability compensation to diarize both the enrollment and testing datasets. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. Table 5.11 shows the accuracy of the speaker verification system in terms of EER and $min(C_{norm})$ for the four diarization systems, in the *mono-mono* scenario. In this case, the DET curves are more separated than in the previous scenarios, but again the use of intra-session variability does not show significant improvement on the overall results, since the margin to improve is insignificant.

In the previous scenarios, the percentage of recordings with highest DER of a dataset that can be accounted to keep the degradation in the EER and $min(C_{norm})$ with respect to the ideal diarization over certain value has been analyzed. In this case, since both enrollment and testing recordings are processed by the diarization systems, we study the percentage of trials in the speaker verification task that involve recordings with high DER in both enrollment and testing sides, as in Section 4.3.2.3.

Figure 5.7 and Tables 5.12 and 5.13 show the percentage of trials that can be accounted

Diarization system	$\%_{Degrad(min(C_{norm}))} \geq 20\%$	$\%_{Degrad(min(C_{norm}))} \geq 50\%$
Baseline	58.10%	20.43%
Speaker Factors, no comp.	21.33%	5.90%
Speaker Factors, intra-ses. comp.	**10.44%**	**3.40%**

Table 5.10: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

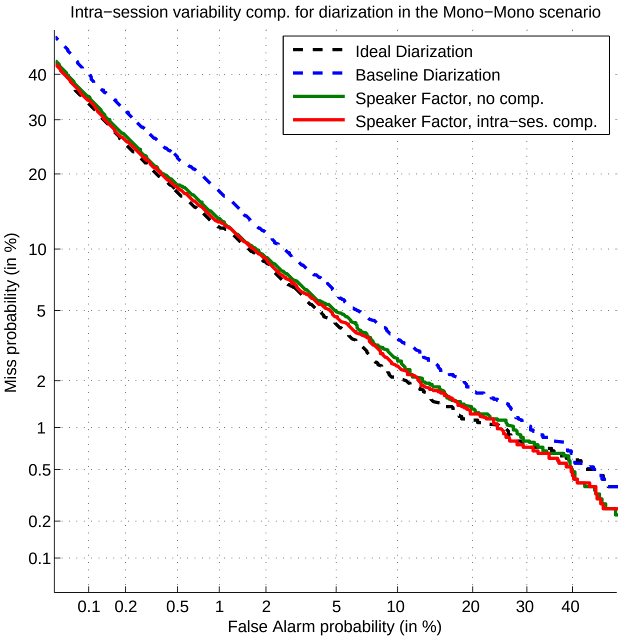


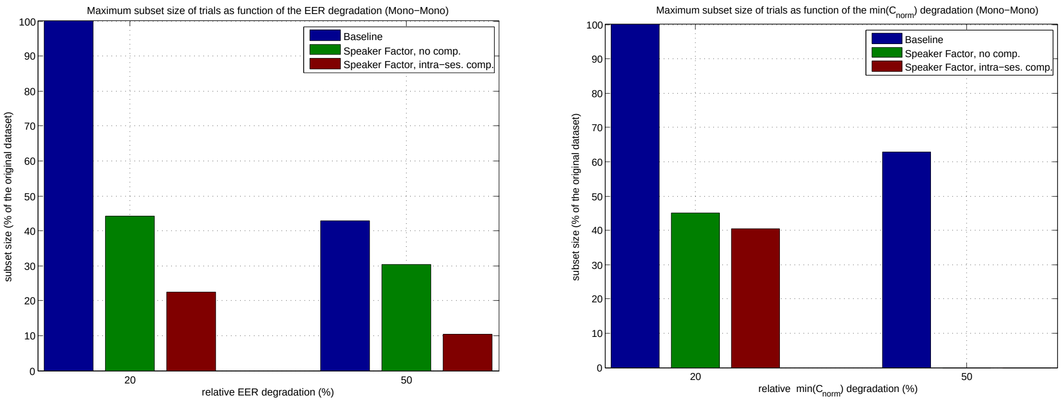
Figure 5.6: *DET curves considering the heavy-weight speaker factor diarization system with and without intra-session variability compensation in the mono-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

to keep the degradation in the EER (Figure 5.7(a)) and $min(C_{norm})$ (Figure 5.7(b)) with respect to the ideal diarization over certain value, considering those recordings with highest DER for enrollment and testing. In this scenario, it can be observed that the use of intra-session variability compensation enables us to account much fewer number of trials that obtain a considerable degradation in terms of EER. In terms of $min(C_{norm})$ the percentage of trials is only slightly reduced.

It has been shown that the improvement introduced by intra-session variability compensation in the diarization task produces a slight improvement in the speaker verification task, which is not clearly reflected in the overall results, since in all cases, the number of recordings that obtain a high DER without intra-session variability compensation is very low for the datasets considered. However, the reduction of the rate of recordings with high DER values will be reflected in the results of the speaker verification task when it faces datasets that have a high number of recordings with high DER.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.54% (0.00%)	0.2157 (0.00%)
Baseline	5.53% (21.81%)	0.2695 (24.94%)
Speaker Factors no comp.	4.99% (9.91%)	0.2289 (6.12%)
Speaker Factors intra-ses. comp.	**4.80% (5.73%)**	**0.2233 (3.52%)**

Table 5.11: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system with and without intra-session variability compensation in the *mono-mono* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum number of trials depending on the degradation in terms of EER*

(b) *Maximum number of trials depending on the degradation in terms of $min(C_{norm})$*

Figure 5.7: *Percentage of trials in the mono-mono scenario that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value, considering the enrollment and testing recordings with highest DER, in the mono-mono scenario.*

Diarization system	$\%_{Degrad(EER)} \geq 20\%$	$\%_{Degrad(EER)} \geq 50\%$
Baseline	100.00%	42.78%
Speaker Factors no comp.	44.23%	30.49%
Speaker Factors intra-ses. comp.	**22.39%**	**10.50%**

Table 5.12: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Diarization system	$\%_{Degrad(min(C_{norm}))} \geq 20\%$	$\%_{Degrad(min(C_{norm}))} \geq 50\%$
Baseline	100.00%	62.80%
Speaker Factors no comp.	47.02%	**0.00%**
Speaker Factors intra-ses. comp.	**40.53%**	**0.00%**

Table 5.13: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Part III

Speaker Clustering: Problems with Unknown Number of Speakers

6

Speaker Clustering

Speaker clustering has been traditionally associated to the task of speaker diarization, since it is an important stage of this task, as explained in Section 2.1. In the framework of speaker diarization, speaker clustering aims at grouping the set of acoustically homogeneous segments obtained as output of a segmentation system into a discrete set of priorly unknown classes, which correspond to different speakers. However, the concept of speaker clustering involves a more general task, since the audio segments to be clustered do not need to belong to the same recording.

In fact, there are speaker clustering applications that do not assume that the audio segments belong to the same recording, and thus, they are not related to speaker diarization. These applications are actually related to speaker characterization. For example, the clustering of different recordings according to the speaker present in them can be used to obtain adapted speaker models in ASR or more robust speaker models in a speaker verification task.

Traditional clustering metrics as BIC or KL2 can be used in the task of clustering different recordings, but their accuracy are severely affected by inter-session variability. Recently, two works, [van Leeuwen, 2010] and [Brummer and De Villiers, 2010], have proposed the use of complete speaker verification systems to solve speaker clustering tasks involving different recordings. In [van Leeuwen, 2010], the use of speaker verification techniques is shown to solve accurately the task of speaker clustering or *speaker linking* in large datasets (datasets with more than a thousand different recordings and some hundreds of speakers). In [Brummer and De Villiers, 2010], the speaker clustering task is referred as the speaker partitioning problem, and it is presented as a generalization of the speaker detection/verification problem. This last work proposes an optimal solution that becomes unfeasible as the size of the dataset increases.

6.1 The Speaker Clustering Task

As explained in Section 2.4.1, given $N = 2$ speech segments, the task of speaker detection/verification aims at selecting the correct hypothesis among the $K = 2$ possible hypotheses $\{H_1, H_2\}$. H_1 is the null hypothesis or target hypothesis, and it states that both segments belong to the same speaker, while H_2 is the non-target hypothesis, and it states that the segments belong to different speakers. The task of speaker clustering can be seen as a generalization of the task of speaker detection/verification, where a set Ω of $N \geq 2$

speech segments is available as input, and the objective is to cluster the speech segments into classes according to the speaker that uttered every segment.

6.1.1 The Speaker Partitioning Problem: optimal solution

We denote each speech segment n by χ_n. Then, the clustering problem reduces to finding the partition of the available set of speech segments $X = \{\chi_1, \chi_2 ..., \chi_N\}$ that clusters the speech segments according to the speaker they belong to. The desired partition is unique, and must be selected among all K possible hypothetical partitions $\mathcal{H} = \{H_1, H_2, ..., H_K\}$, from the coarsest partition H_1 that assumes that all segments belong to the same speaker, to the finest partition H_K that assumes that every segment contains a unique speaker. Every hypothetical partition H_k is composed of S_k non-overlapping clusters $\boldsymbol{\mathcal{C}_{H_k}} = (\mathcal{C}_k(1), \mathcal{C}_k(2), ..., \mathcal{C}_k(S_k))$, which together contains all speech segments of X. As commented in Section 2.4.1 and in [Brummer and De Villiers, 2010], The number of partitions K for a set of N members is given by $K = B_N$, where B_N denotes the N^{th} Bell number.

Let us define π_k as the prior probability for hypothesis H_k and $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$. We assume that a framework Θ that enables us to compute the posterior probability for any hypothetical partition of the given set X is available. Thus, the posterior probability for every partition H_k is defined as follows:

$$P(H_k|X, \boldsymbol{\pi}, \Theta) \tag{6.1}$$

Therefore, to solve the task of speaker clustering, we need to compute the posterior probability for the $K = B_N$ possible partitions and select the partition with highest posterior probability as the optimal speaker partition H_{opt} for the set X:

$$k_{opt} = \underset{k}{argmax}\, P(H_k|X, \boldsymbol{\pi}, \Theta) \tag{6.2}$$

$$H_{opt} = H_{k_{opt}} \tag{6.3}$$

This solution is optimal in the sense that all partitions are evaluated and the one with highest probability is selected. However, as explained in Section 2.4.1, this solution is not feasible in most real cases, since the number of hypothetical partitions B_N for a set increases dramatically as the number of elements N in the set increases.

Table 6.1 shows the number of partitions and non-empty subsets considered by this approach depending on the number of elements in the set. The number of non-empty subsets or possible clusters for a set of N elements, which can be computed as $2^N - 1$, is also shown. In the field of speaker recognition, the speaker partitioning problem is usually solved on datasets with sizes that go from some tens to some thousands of recordings. In the framework of speaker diarization, the number of audio segments considered as input of a speaker clustering system depends on the duration of the recording to diarize, but it can go from ten to a hundred. In the case of the speaker diarization system based on speaker factors, the number of speaker factor vectors obtained for a recording of 5 minute length is usually around twenty thousand (considering only those speaker factor vectors associated to speech frames). According to Table 6.1, the number of possible partition hypotheses to analyze in the problems that appear in the fields of speaker recognition and diarization is so huge to evaluate all of them.

Elements (N)	Partitions (K_{opt})	Non-empty subsets (C_{opt})
1	1	1
2	2	3
3	5	7
4	15	15
5	52	31
10	115975	1023
20	5.17×10^{13}	1048575
100	4.76×10^{115}	1.27×10^{30}
1000	2.99×10^{1927}	1.07×10^{301}

Table 6.1: Number of hypothetical partitions ($K_{opt} = B_N$) and number of non-empty subsets or clusters ($C_{opt} = 2^N - 1$) to evaluate in the optimal solution for several sizes (N) of the input set.

6.1.2 AHC suboptimal solution

In order to avoid the evaluation of all possible hypothetical partitions for a given set of speech segments, we consider a Bottom-up Hierarchical Clustering or AHC approach, which is a suboptimal solution described in 2.4.2. Bottom-up Hierarchical Clustering is a greedy solution for the partitioning problem that reduces significantly the number of partitions to evaluate by making locally optimal choices, so that the solution to the speaker clustering problem is feasible. In this approach, for simplicity, it is assumed that the prior probability for all partitions is equal $\pi = \pi_1 = \pi..., \pi_K$ so there is no need to take the prior into account to solve the clustering problem. Note that this assumption is equivalent to assume that there is no prior information on the speakers or the number of speakers involved. Thus, the prior considered is a non-informative prior.

The Algorithm 6.1 resumes the AHC process. The process starts from the finest partition, and it reduces the number of speakers iteratively until a stopping criterion is met. In every iteration, a distance matrix D is computed for all possible cluster pairs (i, j), and the closest clusters are merged. Meeting the stopping criterion usually means that the last pair of clusters merged do not seem to belong to the same speaker, so the previous partition is retrieved and returned as solution of the speaker clustering problem. Whenever a single cluster is obtained ($S = 1$) and the stopping criterion is not met, the coarsest partition is returned.

It is important to notice that this approach only selects one partition as possible solution on every iteration. This means that ignoring the stopping criterion, a total of N partitions are considered as possible solutions of the problem. Since every iteration the number of speakers s is decreased, these partitions will contain a decreasing number of clusters, from N to 1. Therefore, there is only one partition considered as solution for a given value of s, which is referred uniquely as $H_{S=s}$. Then, the task of the stopping criterion is to select among these N partitions and determine the actual number of speakers S_{AHC}.

But this approach explores more than N partitions for a given set of N speech segments. In fact, in every iteration, an initial partition $H_{S=s}$ and its corresponding set of s clusters $\boldsymbol{\mathcal{C}}_{\boldsymbol{H_{S=s}}}$ are available. The AHC approach explores all possible partitions considering $s - 1$ speakers obtained by merging two clusters of the initial set of s clusters $\boldsymbol{\mathcal{C}}_{\boldsymbol{H_{S=s}}}$. Thus, the number partitions considering $s - 1$ clusters explored in an iteration can be computed as

Algorithm 6.1: AHC process

```
input  : $X = \{\chi_1, ..., \chi_N\}$
output: $\boldsymbol{\mathcal{C}_{(H_{AHC})}}$ (clusters for the desired partition $H_{AHC} = H_{S=S_{AHC}}$)
begin
    $s = N$;
    for $n \leftarrow 1$ to $N$ do        // start from the finest partition $H_{S=N}$
        $\mathcal{C}_{(H_{S=s})}(n) = \{\chi_n\}$;
    end
    $\boldsymbol{\mathcal{C}_{H_{S=s}}} = \{\mathcal{C}_{H_{S=s}}(1), ..., \mathcal{C}_{H_{S=s}}(S)\}$;
    repeat                                       // Merge closest clusters
        if $s$==$1$ then               // Return coarsest partition if s=1
            $S_{AHC} = 1$;
            $\boldsymbol{\mathcal{C}_{H_{AHC}}} \longleftarrow \boldsymbol{\mathcal{C}_{H_{S=S_{AHC}}}}$;
            return $\boldsymbol{\mathcal{C}_{H_{AHC}}}$;
        end
        for $i \leftarrow 1$ to $s-1$ do                   // Compute distances
            for $j \leftarrow i+1$ to $s$ do
                $D(i,j) = Distance(\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j))$;
            end
        end
        find $(i_m, j_m)$ so that $D(i_m, j_m) \leq D(i,j)\,\forall(i,j)$;
        $\mathcal{C}_{H_{S=s-1}}(i_m) = \mathcal{C}_{H_{S=s}}(i_m) \cup \mathcal{C}_{H_{S=s}}(j_m)$;
        $\boldsymbol{\mathcal{C}_{H_{S=s-1}}} \longleftarrow (\boldsymbol{\mathcal{C}_{H_{S=s}}} \backslash \{\mathcal{C}_{H_{S=s}}(i_m), \mathcal{C}_{H_{S=s}}(j_m)\}) \cup \{\mathcal{C}_{H_{S=s-1}}(i_m)\}$;
        $s = s-1$;
    until $Stop(\boldsymbol{\mathcal{C}_{H_{S=s}}}, D)$==$true$  // until stopping criterion is met;
    $S_{AHC} = s+1$;
    $\boldsymbol{\mathcal{C}_{H_{AHC}}} \longleftarrow \boldsymbol{\mathcal{C}_{H_{S=S_{AHC}}}}$;
    return $\boldsymbol{\mathcal{C}_{H_{AHC}}}$;
end
```

$K_{H(S=s-1)} = \sum_{k=1}^{k=s-1} k = \frac{s(s-1)}{2}$, where $H(S = s-1)$ denotes the set of partitions explored in contrast to $H_{S=s-1}$ which denotes the partition selected among all partitions explored.

Therefore, the total number of partitions explored for a set of N speech segments can be computed as the sum of the partitions explored in all iterations:

$$K_{AHC} = \sum_{n=1}^{n=N-1} \frac{(N-n+1)(N-n)}{2} + 1 = \frac{N(N-1)(2N-1)}{12} - \frac{N(N-1)}{4} + 1, \quad (6.4)$$

where the initial (finest) partition has been included.

The total number of non-empty subsets or clusters considered during the AHC process can also be obtained easily. In the beginning, the finest partition is explored, considering a total of N non-empty subsets (the speech segments or initial). Then, in the first iteration all possible cluster pairs are considered for a total of $\sum_{n=1}^{n=N-1}(n) = \frac{N(N-1)}{2}$. Finally, during following iterations, the only new subsets considered are those involving the cluster merged in the previous iteration, since the subsets involving the other clusters have been accounted in previous iterations. Thus, the new number of subsets or clusters considered after the first iteration is given by $s-1=N-l$, where l is the iteration. Therefore, the number of non-empty subsets considered by the AHC suboptimal approach, for a set of N elements is:

$$\begin{aligned} C_{AHC} &= N + \frac{N(N-1)}{2} + \sum_{l=2}^{l=N-1}(N-l) \\ &= N + \frac{N(N-1)}{2} + \frac{(N-1)(N-2)}{2} \\ &= N(N-1)+1. \end{aligned} \tag{6.5}$$

Elements (N)	Partitions (K_{AHC})	Non-empty subsets (C_{AHC})
1	1	1
2	2	3
3	5	7
4	11	13
5	21	21
10	166	91
20	1331	381
100	166651	9901
1000	166666501	999001

Table 6.2: Number of hypothetical partitions ($K_{AHC} = \frac{N(N-1)(2N-1)}{12} + \frac{N(N-1)}{4} + 1$) and number of non-empty subsets ($C_{AHC} = N(N-1)+1$) explored by the AHC suboptimal solution for several sizes (N) of the input set.

Table 6.2 shows the number of hypothetical partitions and non-empty subsets explored by the AHC suboptimal solution for several sizes of the input set. Note that both the number of partitions and non-empty subsets have been reduced significantly, and the solution of the problem is feasible even for high values of N. Now the number of possible partitions for a set of N elements is $O(N^3)$, while the number of non-empty subsets to compute is $O(N^2)$.

It is important to notice that the number of partitions and non-empty subsets considered by the AHC process is in general lower than the numbers presented in Table 6.2. In Table 6.2, we are assuming that the AHC process is performed until the coarsest partition is reached. However, it is usual to meet the stopping criterion before this point, so many iterations are not performed, and thus many partitions and non-empty subsets are not considered.

Therefore, this approach is suboptimal in the sense that it does not consider all possible partitions, for two main reasons: First, the AHC process makes locally optimal choices, merging two clusters on every iteration. Every time two clusters are merged, they are never separated again, so several partitions will not be explored. Second, once the stopping criterion is met, the AHC algorithm stops and the remaining partitions are not explored.

6.1.3 Simplified AHC suboptimal solution

We have seen that using AHC to solve the speaker partitioning problem makes the solution of the problem feasible. However, the complexity of the clustering problem can be reduced further. The work presented in [van Leeuwen, 2010] shows that the clustering task can be solved accurately considering only the distance matrix D obtained in the first iteration. This is achieved using accurate distance metrics, as the scores provided by a speaker verification system.

Therefore, in the approach presented in [van Leeuwen, 2010], the distances are only computed for every possible pair of speech segments from the input set X, and in every iteration, no new distances are obtained. Instead of recomputing the distance metrics considering the new cluster obtained in every iteration, the initial distances obtained for the elements that belong to the new cluster are considered. Every time two elements from different clusters need to be merged, the clusters they belong to are merged instead.

Note that this approach is a simplification of the one described in Section 6.1.2. Given a set of N speech segments, this simplification explores the same partitions as the non-simplified AHC approach, but the actual number of non-empty subsets considered for computing the distances is smaller. In fact, the distances are only computed between all elements from the finest partition, thus, the non-empty subsets considered are those containing one or two speech segments. The final partition can provide clusters containing more than two elements, but the distances between these clusters are not computed properly, but approximated with the distances obtained for clusters containing a single element. Therefore, the actual number of non-empty subsets considered in this approach for distance computation is given by $C_{simpAHC} = N + \frac{N(N-1)}{2} = \frac{N(N+1)}{2}$.

Elements (N)	Partitions (K_{AHC})	Non-empty subsets ($C_{simpAHC}$)
1	1	1
2	2	3
3	5	6
4	11	10
5	21	15
10	166	55
20	1331	210
100	166651	5050
1000	166666501	500500

Table 6.3: Number of hypothetical partitions ($K_{AHC} = \frac{N(N-1)(2N-1)}{12} + \frac{N(N-1)}{4} + 1$) explored and number of non-empty subsets considered for distance computation ($C_{simpAHC} = \frac{N(N+1)}{2}$) in the simplified AHC suboptimal solution for several sizes (N) of the input set.

Table 6.3 shows the number of hypothetical partitions explored and non-empty subsets considered for distance computation by the simplified AHC suboptimal solution for several sizes of the input set. Note that the number of non-empty subsets have been reduced further, but it is still $O(N^2)$. Also note that the number of hypothetical partitions considered is the same as in the non-simplified AHC suboptimal solution. However the distances computed between clusters comprising more that one audio segment are not evaluated correctly, but using approximations, so the distance matrix for partitions involving a number of speakers $S < N - 1$ is not computed exactly.

Since the actual number of non-empty subsets considered for distance computations is halved compared to the non-simplified AHC suboptimal solution, the final computation cost will be reduced significantly, but not dramatically. Nevertheless, computational cost reduction is not the only advantage that the simplified method provides. Another interesting advantage is that this simplification enables us to explore constrained solutions with no additional computational cost. This simplified AHC algorithm only makes use of the initial distances, that do not depend on how the speech segments are agglomerated. For example, once a clustering problem has been solved using this simplified approach, the solution of subproblems of the original problem (for example subsets from the original dataset) can be obtained without any additional cost. On the other hand, the non-simplified AHC solution will need to recompute distance metrics for the new subproblem.

This advantage is very interesting for several reasons: in real speaker clustering problems related to speaker verification, it is usual to have informative meta-data that enables us to extract subsets out of the original dataset (for example telephone number, location,...). If several overlapped subsets can be built according to the available meta-data, the non-simplified AHC approach for speaker clustering will need to recompute the distances for several clusters, but the simplified approach can rearrange the speaker clusters without further computations. In addition it is known that the accuracy of a speaker clustering task depends on the size of the dataset N, obtaining lower accuracy as N increases [van Leeuwen, 2010]. The use of this simplification enables us to consider different subsets of the dataset in order to fragment the problem into easier problems without increasing the computational cost.

6.2 Stopping Criteria for Speaker Clustering

In the previous Section several solutions for speaker clustering have been described. The optimal solution is not feasible, so two modalities of a greedy bottom-up hierarchical approach have been proposed. These approaches, referred as AHC and simplified AHC, need a stopping criterion to determine the actual number of speakers present in the available set of audio segments. In this study we assume that the number of speakers is unknown and needs to be guessed. Note that if the number of speakers $s = S_{known}$ is known, since AHC approaches provide only a single partition for a given number of speakers $H_{S=S_{known}}$, the stopping criterion is reduced to reach the iteration that outputs the hypothetical partition with the desired number of speakers S_{known}.

Two stopping criteria are described in this Section: the use of a threshold on the distance metric to determine when two cluster should not be merged and the use of the Student´s t-test on the distributions of the distances computed for the available clusters [Nguyen *et al.*, 1998].

6.2.1 Threshold on the Distance metric

The most straightforward stopping criterion for the proposed AHC strategies for speaker partitioning is the use of a threshold for the value of the distance metric considered. This stopping criterion considers the two clusters that were merged during the last iteration $\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j)$ and checks whether the distance between them $D(\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j))$ is over or below ϵ. Whenever $D(\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j)) > \epsilon$, it means that the clusters merged during the last iteration $\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j)$ should not be merged.

In order to determine the value of ϵ, a development dataset is needed. Note that certain distance metrics as *well-calibrated* likelihood ratios (LRs) or the BIC can set $\epsilon = 0$, but they usually need a development dataset to perform calibration or to adjust the penalty term.

These processes are somehow setting an implicit threshold.

6.2.2 Student's t-test Stopping Criterion

The previous stopping criterion involves the use of a development dataset to set a threshold for the distance metric. However, it is known that for most distance metrics, the range of the values obtained may depend on the datasets. This is very clear for example in speaker recognition, where mismatched datasets obtain different behavior, and these differences in behavior may include variations in the optimal threshold value. This fact makes the use of a threshold a risky strategy unless the development dataset is known to be similar to the dataset that is going to be processed by the clustering algorithm.

To avoid the risk of mismatched development and testing datasets, a new stopping criterion for speaker clustering has been recently proposed in [Nguyen *et al.*, 1998]. This technique makes use of the Student's t-test statistic computed on the populations for the intra-cluster and inter-cluster distances to determine when to stop.

The Student's t-test provides a measure of the separation between two populations. To use it as stopping criterion, we define the population of intra-cluster D_{intra} and inter-cluster D_{inter} distances for a given partition $H_{S=s}$ obtained during the AHC process as:

$$D_{intra} = \bigcup_{\substack{i \in \mathcal{C}_{H_{S=s}}(k) \\ j \in \mathcal{C}_{H_{S=s}}(k) \\ \forall k}} D(\mathcal{C}_{H_{S=N}}(i), \mathcal{C}_{H_{S=N}}(j)) \tag{6.6}$$

$$D_{inter} = \bigcup_{\substack{i \in \mathcal{C}_{H_{S=s}}(k_i) \\ j \in \mathcal{C}_{H_{S=s}}(k_j) \\ \forall k_i, k_j, k_i \neq k_j}} D(\mathcal{C}_{H_{S=N}}(i), \mathcal{C}_{H_{S=N}}(j)) \tag{6.7}$$

where $\mathcal{C}_{H_{S=N}}(i) = \{\chi_i\}$, $\mathcal{C}_{H_{S=N}}(j) = \{\chi_j\}$ are clusters from the initial or finest partition and $\mathcal{C}_{H_{S=s}}(k)$ is the cluster k for the current partition $H_{S=s}$. Thus, D_{intra} is composed of the distances computed for all pairs of initial clusters $\mathcal{C}_{H_{S=N}}(i), \mathcal{C}_{H_{S=N}}(j)$ that belong to the same cluster $\mathcal{C}_{H_{S=s}}(k)$ given the current partition $H_{S=s}$, and D_{inter} is composed of the distances computed for all pairs of initial clusters $\mathcal{C}_{H_{S=N}}(i), \mathcal{C}_{H_{S=N}}(j)$ that belong to the different clusters $\mathcal{C}_{H_{S=s}}(k_i), \mathcal{C}_{H_{S=s}}(k_i)$, $k_i \neq k_j$ given the current partition $H_{S=s}$.

Therefore, for every obtained partition, we can obtain the populations D_{intra} and D_{inter}. Then, the separation between the distributions for both populations can be measured by means of the t-test statistic. But before obtaining the t-test statistic, some assumptions must be done for the distribution on the D_{intra} and D_{inter} populations. In those cases where both populations can be assumed to follow Gaussian distributions, the t-test statistic t_s for both distributions is computed as:

$$t_s = \frac{m_{intra} - m_{inter}}{\sqrt{\frac{\sigma_{intra}}{N_{intra}} + \frac{\sigma_{inter}}{N_{inter}}}}, \tag{6.8}$$

where m_{intra}, σ_{intra} and N_{intra} are the mean, standard deviation and size of D_{intra}, and the m_{inter}, σ_{inter} and N_{inter} are the mean, standard deviation and size of D_{inter}.

There are some distance metrics that are known to follow Gaussian distributions for D_{intra} and D_{inter}. Some examples are the LRs provided by some speaker recognition systems (for example PLDA) or the ΔBIC values. However, there are several distance metrics that do not follow Gaussian distribution, and the Gaussianity assumption is unreasonable for them. In [Nguyen *et al.*, 1998], another measure of the separation between the distributions for the populations D_{intra} and D_{inter} is proposed, which measures the overlap between both distributions without any assumption on the shape of the distribution they follow. This measure ρ is obtained following the next steps:

- $D_{all} = D_{intra} \cup D_{inter}$
- The distances in D_{all} are sorted in ascending order and then:

$$R_{intra} = \sum_{D_{all}(i) \in D_{intra}} rank(D_{all}(i)) \tag{6.9}$$

$$U_{intra} = R_{intra} - \frac{\|D_{intra}\|(\|D_{intra}\| + 1)}{2} \tag{6.10}$$

$$\rho = \left| \frac{U_{intra}}{\|D_{intra}\| \, \|D_{inter}\|} - 0.5 \right| \times 2, \tag{6.11}$$

where $rank(D_{all}(i))$ is the order of $D_{all}(i)$ in the sorted sequence of D_{all} and $\|.\|$ is the cardinal of the set. The measure ρ can take values between 0 and 1. A value of $\rho = 0$ represents complete overlap between D_{intra} and D_{inter}, while a value of $\rho = 1$ represent complete separation of the distributions.

In order to use these measures as stopping criteria, t_s or ρ is computed for the partition selected in every iteration of the AHC process. The partition obtaining maximum t_s or ρ value is the one selected as solution of the speaker clustering problem. The advantage of these stopping criteria is that they do not rely on any prior information and thus they do not need any development dataset to set a threshold.

6.3 PLDA for Speaker Clustering

Recently, two works presented in [van Leeuwen, 2010] and [Brummer and De Villiers, 2010] show that speaker verification techniques can be used for speaker clustering obtaining high accuracy. In this section, we analyze how the PLDA speaker verification system described in Section 3.1.2 can be used to solve the speaker partitioning problem, considering the solutions and stopping criteria presented in Sections 6.1 and 6.2.

6.3.1 PLDA Optimal Solution

The PLDA approach for speaker verification provides a framework that enables us to compute the likelihood for any hypothetical partition of the given set of speech segments. In this framework, every speech segment is represented using an i-vector, so from now on, the set of elements to cluster is a set of N i-vectors $\Phi = \{\phi_1, ..., \phi_N\}$. Every hypothetical partition H_k considered is composed of S_k non-overlapping clusters $\mathcal{C}_k(1), \mathcal{C}_k(2), ..., \mathcal{C}_k(S_k)$, which together contains all i-vectors of Φ. Assuming that the PLDA model is represented by Θ, and that the i-vectors belonging to different speakers are statistically independent (which is an intrinsic assumption of the PLDA model), the likelihood for the partition H_k is defined as the product of the likelihoods for all non-overlapping clusters of H_k:

$$\mathcal{L}(H_k|\Phi,\Theta) = \prod_{j=1}^{S_k} \mathcal{L}(\mathcal{C}_k(j)|\Theta) \tag{6.12}$$

The posterior probability of the partition H_k, given the prior for all partitions $\boldsymbol{\pi}$ is computed as:

$$P(H_k|\Phi,\boldsymbol{\pi},\Theta) = \frac{\pi_k \mathcal{L}(H_k|\Phi,\Theta)}{\sum_{i=1}^{K} \pi_i \mathcal{L}(H_i|\Phi,\Theta)} \tag{6.13}$$

For simplicity, it is assumed that there is no prior information available in the task of speaker clustering, so we consider $\pi = \pi_1, ..., \pi_k, ..., \pi_K$. Note that the term in the denominator is common for all partitions to evaluate. Therefore, to solve the speaker partitioning problem, under the assumption of equal prior probabilities for all partitions, we need to compute the likelihood for the $K = B_N$ possible partitions and select the partition with highest likelihood as the optimal speaker partition H_{opt} for the set Φ:

$$k_{opt} = \underset{k}{argmax}\, \mathcal{L}(H_k|\Phi,\Theta) \tag{6.14}$$

$$H_{opt} = H_{k_{opt}} \tag{6.15}$$

As an example let us assume that a set $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ composed of $N = 4$ i-vectors is available, and the speaker partition of Φ is desired. The total number of hypothetical partitions of this set is $K = B_4 = 15$. Appendix B comprises the 15 possible partitions $H_1, H_2, ..., H_{15}$ of the set Φ. To determine which one of these partitions is the optimal speaker partition, we first compute the likelihood for every partition (see Appendix B), and then the partition H_{opt} obtaining highest likelihood is selected, according to eq. (6.14) and eq. (6.15).

Since the likelihood for every partition can be obtained as the product of the likelihoods for all non-overlapping clusters of H_k, the number of likelihood computations $C_{opt}(\mathcal{L})$ in order to solve the clustering problem using this approach is equal to the number of non-empty subsets available in Φ. Thus $C_{opt}(\mathcal{L}) = 2^N - 1$. Table 6.1 shows the number of non-empty subsets and thus the number of likelihood computations $C_{opt}(\mathcal{L})$ required to solve the speaker clustering problem using this approach as function of N. Since it is usual to find speaker clustering problems considering hundreds or thousands of i-vectors as input, this approach is not feasible for most cases.

6.3.2 PLDA AHC suboptimal solution

The AHC solution for speaker clustering reduces the search space by making locally optimal choice based on a definition of distance between two clusters. The PLDA speaker verification system described in Section 3.1.2 can be easily integrated in this solution, since the LRs provided for speaker verification can be directly used as distance metric in the algorithm.

Therefore, given the set of i-vectors Φ, two clusters $\mathcal{C}_{H_{S=s}}(i), \mathcal{C}_{H_{S=s}}(j)$ obtained from a given partition $H_{S=s}$, and the PLDA model Θ, the distance metric for the PLDA AHC solution is defined as:

$$\begin{aligned}D(\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)) &= LR_{PLDA}(\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)) \\ &= \frac{\mathcal{L}(\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)|\Theta)}{\mathcal{L}(\mathcal{C}_{H_{S=s}}(i)|\Theta)\mathcal{L}(\mathcal{C}_{H_{S=s}}(j)|\Theta)}.\end{aligned} \tag{6.16}$$

It is important to notice that according to the definition of the LR provided by the PLDA speaker verification system, as it is usual in speaker verification, the higher the LR, the more likely are the i-vectors considered to belong to the same speaker. Thus, the LR is not actually a distance metric but a closeness metric. Then, in order to plug the PLDA LR as a distance metric in the Algorithm 6.1, the clusters selected to merge must be those obtaining maximum LR, or the LR can be inverted to be considered directly as distance metric.

One advantage of this distance metric is that the LR computed for a pair of clusters $\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)$ from a partition $H_{S=s}$ is actually the LR obtained when comparing the initial partition $H_{S=s}$ to the partition $H^{(i,j)}_{S=s-1}$ where the partition $H^{(i,j)}_{S=s-1}$ is obtained from $H_{S=s}$ by merging the clusters $\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)$. In fact, if $\mathcal{C}_{H^{(i,j)}_{S=s-1}}(i) = \mathcal{C}_{H_{S=s}}(i) \cup \mathcal{C}_{H_{S=s}}(j)$, it is shown that:

$$\boldsymbol{\mathcal{C}}_{\boldsymbol{H^{(i,j)}_{S=s-1}}} = \boldsymbol{\mathcal{C}}_{\boldsymbol{H_{S=s}}} \backslash \{\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)\} \cup \mathcal{C}_{H^{(i,j)}_{S=s-1}}(i) \tag{6.17}$$

$$\begin{aligned}LR(H^{(i,j)}_{S=s-1},H_{S=s}) &= \frac{\mathcal{L}(H^{(i,j)}_{S=s-1}|\Phi,\Theta)}{\mathcal{L}(H_{S=s}|\Phi,\Theta)} \\ &= \frac{\left(\prod_{k\neq i,j}\mathcal{L}(\mathcal{C}_{H_{S=s}}(k)|\Theta)\right)\mathcal{L}(\mathcal{C}_{H^{(i,j)}_{S=s-1}}(i)|\Theta)}{\left(\prod_{k\neq i,j}\mathcal{L}(\mathcal{C}_{H_{S=s}}(k)|\Theta)\right)\mathcal{L}(\mathcal{C}_{H_{S=s}}(i)|\Theta)\mathcal{L}(\mathcal{C}_{H_{S=s}}(j)|\Theta)} \\ &= \frac{\mathcal{L}(\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)|\Theta)}{\mathcal{L}(\mathcal{C}_{H_{S=s}}(i)|\Theta)\mathcal{L}(\mathcal{C}_{H_{S=s}}(j)|\Theta)},\end{aligned} \tag{6.18}$$

so the distance metric provided by the PLDA speaker verification system for a pair of clusters is in fact a LR that compares the current partition to the one obtained merging the two clusters considered, determining which partition better describes the set Φ. Note that the use of LRs also sets a straightforward stopping criterion for PLDA AHC. Since the LRs are comparing partitions, as far as the maximum LR value obtained for all available pairs of clusters $LR(\mathcal{C}_{H_{S=s}}(i_m),\mathcal{C}_{H_{S=s}}(j_m))$ in the current iteration is over 1, the AHC process should keep merging clusters. Actually, $LR(\mathcal{C}_{H_{S=s}}(i_m),\mathcal{C}_{H_{S=s}}(j_m)) \geq 1$ means that the likelihood for the partition $H_{S=s-1}$ obtained by merging the clusters $\mathcal{C}_{H_{S=s}}(i_m),\mathcal{C}_{H_{S=s}}(j_m)$ is greater than the likelihood obtained for the initial partition $H_{S=s}$. However, whenever $LR(\mathcal{C}_{H_{S=s}}(i),\mathcal{C}_{H_{S=s}}(j)) < 1$, the AHC process should stop and return the current partition $H_{S=s}$, since it means that the the likelihood for the new partition $H_{S=s-1}$ is below the likelihood for $H_{S=s}$. It is usual to consider log-likelihood ratio $LLR = log(LR)$ values instead of LR values, so the proposed stopping criterion reduces to setting a threshold $\epsilon = 0$.

As an example of the operation of the AHC (see Algorithm 6.1) considering PLDA LLRs as clustering metric, let us consider again the set $\Phi = \{\phi_1,\phi_2,\phi_3,\phi_4\}$ composed of

$N = 4$ i-vectors. The PLDA AHC approach starts from the finest partition $H_{S=4} = H_{15}$, $\boldsymbol{\mathcal{C}_{H_{S=4}}} = \{\mathcal{C}_{15}(1) = \{\phi_1\}, \mathcal{C}_{15}(2) = \{\phi_2\}, \mathcal{C}_{15}(3) = \{\phi_3\}, \mathcal{C}_{15}(4) = \{\phi_4\}\}$ (see Appendix B), and computes the LLR for all possible cluster pairs, which is equivalent to compare the initial partition to all possible partitions that assume that there are three speakers in Φ: $H_9, H_{10}, H_{11}, H_{12}, H_{13}, H_{14}$. We merge the pair of i-vectors obtaining maximum LLR, which is equivalent to select the partition obtaining maximum LLR. Let us assume that $LLR(H_{11}, H_{15}) > LLR(H_k, H_{15})$, $k = 9, 10, 12, 13, 14$ an thus the selected partition is H_{11}, whose LLR is given by:

$$\begin{aligned} LLR(H_{11}, H_{15}) &= log\left(\frac{\mathcal{L}(H_{11}|\Phi,\Theta)}{\mathcal{L}(H_{15}|\Phi,\Theta)}\right) \\ &= log\left(\frac{\mathcal{L}(\phi_1,\phi_4|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_3|\Theta)}{\mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_3|\Theta)\mathcal{L}(\phi_4|\Theta)}\right) \\ &= log\left(\frac{\mathcal{L}(\phi_1,\phi_4|\Theta)}{\mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_4|\Theta)}\right). \end{aligned} \tag{6.19}$$

Note that to compare the partitions H_{11} and H_{15} we only need to compute LLR for the pair ϕ_1, ϕ_4. Thus, the clusters to merge are $\mathcal{C}_{15}(1) = \{\phi_1\}, \mathcal{C}_{15}(4) = \{\phi_4\}$. Then the AHC process check whether the stopping criterion is fulfilled, in our case whether $LLR(H_{11}, H_{15}) < 0$. If so, it means that the clusters $\mathcal{C}_{15}(1), \mathcal{C}_{15}(4)$ should not be merged, and the selected partition $H_{AHC} = H_{S=4}$. Otherwise, the current partition is $H_{S=3} = H_{11}$ and the AHC process goes to the next iteration.

During the next iteration, the LLRs are computed for all possible pair of the clusters available in the current partition $H_{S=3} = H_{11}$. This means that only those partitions involving $S = 2$ speakers obtained by merging any pair of clusters from partition H_{11} are explored. In this case, these partitions are H_3, H_4, H_8 (see Appendix B). Note that the partitions H_2, H_5, H_6, H_7 are not explored and could never be selected as solution of the AHC approach. Among H_3, H_4, H_8, the partition obtaining maximum LLR when compared to $H_{S=3} = H_{11}$ is selected, let us assume that H_4 is the partition selected. Again the stopping criterion $LLR(H_4, H_{11}) < 0$ is checked, and if it is fulfilled, the partition selected is H_{11} and the process finishes. Otherwise, the current partition is $H_{S=2} = H_4$ and the AHC process continues iteratively until the coarsest partition $H_{S=1} = H_1$ is reached or the stopping criterion is met.

Reducing the search space by discarding partitions enables us to decrease the computational cost of the speaker clustering task. The computational cost of this approach can be measured in terms of likelihood computations $C_{AHC}(\mathcal{L})$ required to reach the desired partition H_{AHC}. The number of likelihood computations is actually the number of non-empty subsets explored by the algorithm before the algorithm is finished. Ignoring the stopping criterion, $C_{AHC}(\mathcal{L}) = N(N-1)+1$, but in general the stopping criterion will be met before reaching the coarsest partition so the number of likelihood computations will be lower. Table 6.2 shows the maximum number of non-empty subsets and thus the maximum number of likelihood that this approach will compute as function of the size of the input set N. It can be seen that this approach is feasible even for large values of N.

It is interesting to note that the PLDA AHC approach is suboptimal in the sense that the search space is reduced since it does not evaluate all possible partitions, but the partitions explored are evaluated exactly. Therefore, as far as the partition that obtains maximum likelihood is explored during the AHC process, the solution for the speaker clustering problem provided by this approach will be identical to that provided by the optimal solution.

6.3.3 Simplified PLDA AHC suboptimal solution

As explained in Section 6.1.3, the AHC approach can be simplified by making all clustering decisions based on the distance matrix D obtained in the first iteration. This is achieved using accurate distance metrics, as the scores obtained from a speaker verification system. In fact, this approach was proposed in [van Leeuwen, 2010], where it is shown that the score of a GMM-SVM-NAP speaker verification system [Campbell *et al.*, 2006] can be considered as distance metric to solve the speaker clustering problem accurately. Thus, the LLRs provided by a PLDA speaker verification system can be considered as distance metric for this clustering approach.

Given a set of N i-vectors, this simplification considers the same partitions as the non-simplified AHC approach, but it computes fewer likelihood values. The likelihood values are correctly computed for the finest partition and for all possible partitions with $S = N - 1$ speakers. Then the LLRs involving those partitions comprising $S < N - 1$ speakers are approximated by the LLRs obtained for the pairs of i-vectors belonging to the clusters to merge.

Coming back to our example of a set $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ composed of $N = 4$ i-vectors, considering this approach, only six likelihood ratios are computed: $LLR(H_k, H_{15})$, $k = 9, 10, ..., 14$, and thus a total of ten likelihood values need to be computed (the likelihood for every one of the four i-vectors and the joint likelihood for every one of the six possible pairs). Then, the two i-vectors obtaining maximum LLR are clustered together. According to our example, these i-vectors are ϕ_1, ϕ_4. The stopping criterion is checked considering the LR of the clustered i-vectors. Then and iteratively the next two i-vectors obtaining maximum LR are selected, let us assume that in our example, these i-vectors are ϕ_3 and ϕ_4. Since ϕ_4 belongs to a cluster that contains other i-vectors, ϕ_1 in this case, we merge both clusters obtaining a single cluster containing ϕ_1, ϕ_3, ϕ_4. Then the stopping criterion is checked again, considering the LR obtained for ϕ_3, ϕ_4. This process is repeated until the stopping criterion is met or the coarsest partition is reached.

Note that in this approach there is no need to compute new likelihoods as the clustering process evolves. This provides some advantages. The first and most obvious one is that the number of likelihood computations is reduced to the number of non-empty subsets that are considered exactly for LLR computations. Thus $C_{simpAHC}(\mathcal{L}) = N + \frac{N(N-1)}{2} = \frac{N(N+1)}{2}$. Table 6.3 shows the number of non-empty subsets exactly considered in this approach and thus the number of likelihood calculations depending on the size of the input set N.

It can be seen that the advantages in terms of computational cost provided by this simplification are significant but not dramatic. The number of likelihood values to compute is halved compared to the non-simplified AHC suboptimal solution. However, there is an additional and interesting computational cost reduction when considering the simplified approach. During the clustering process, several terms needed for likelihood calculations can be precomputed and considered for most of the calculations. One of these terms, the covariance of the posterior distribution on the PLDA speaker factors for every cluster, depend on the number of i-vectors in the clusters considered. The simplified approach always calculate likelihoods considering a single i-vector or a pair of i-vectors, so the two covariance matrices needed can be precomputed once at the beginning of the algorithm. However, the non-simplified PLDA AHC approach can consider a cluster as large as the complete set of i-vectors. Thus, precomputing this terms implies high memory consumption which is $O(N)$. In case it is not feasible to keep all the precomputed terms in memory, it can be solved increasing the computational cost for every trial. For example in our PLDA system, this memory consumption is $N \times 80Kb$, which means almost 1Gb for every 10000 i-vectors. If the dataset to cluster is huge, only part of the covariance matrices can be precomputed, and the others will be computed for every trial that demands them.

Moreover, the computational cost of a likelihood computation is not constant and depends on the number of i-vectors ($O(N_C)$) considered for the likelihood computation. Thus, the likelihood computations of the non-simplified PLDA AHC approach will be more costly as larger clusters are obtained, while the simplified approach do not calculate likelihoods for large clusters.

Finally, it is interesting to note that the simplified PLDA AHC approach is suboptimal in the sense that the search space is reduced but also in the sense that the LLRs involving partitions with a number of speakers $S < N-1$ are not computed exactly, but approximated by the LLRs considering the two i-vectors obtaining maximum LLR in the last iteration. Therefore, the partition selected by this approach may not be the same as the one provided by the optimal solution even if the partition obtaining maximum likelihood is in the search space.

6.3.4 Stopping Criteria for PLDA AHC

Both stopping criteria presented in Section 6.2 can be considered to solve the clustering problem using the PLDA AHC approach, but some considerations should be taken into account.

6.3.5 Threshold on the Likelihood Ratio

The most straightforward stopping criterion for the proposed PLDA AHC strategies for speaker partitioning is to set a threshold for the LLR obtained for the pair of clusters to merge in every iteration.

In Section 6.3.2, it has been suggested that the LLRs obtained in the PLDA AHC process can be directly considered as stopping criterion. In fact, if the LRs are correctly calibrated, their log-values will be positive as clusters containing the same speakers are merged, and will be negatives whenever two clusters containing different speakers are merged. Therefore, the PLDA AHC should stop merging clusters as soon as the LLR for the next clusters to merge is negative.

The experience in the field of speaker recognition tells that it is hard to obtain calibrated LRs, and that usually large development datasets are needed for this purpose. A solution to this problem is to set a threshold ϵ for the LLR value different from 0, simply obtained through a calibration process considering a development dataset. Thus, the clustering process is stopped whenever the two clusters to merge obtain a $LLR < \epsilon$.

Note that this stpping criteria can be considered for both the simplified and the non-simplified PLDA AHC approaches, and its computational cost is negligible.

6.3.6 Student's t-test Stopping Criterion

The maximization a the Student's t-test statistic for the population of intra-cluster distances and inter-cluster distances can be also considered as stopping criterion for the PLDA AHC strategy. In fact, the LLRs provided by a PLDA speaker verification system are known to follow Normal distributions for target and non-target trials. Thus the populations D_{intra} and D_{inter} will follow Normal distributions for the actual speaker partition of the given set Φ. Therefore, it is possible to consider the maximization of the t_s metric proposed in 6.2.2 as stopping criterion.

On the other hand, since the D_{intra} and D_{inter} populations are not expected to follow Gaussian distributions for all partitions, the maximization of the ρ metric proposed in 6.2.2 is also considered as stopping criterion. This stopping criterion does not rely on any assumption on the distribution for D_{intra} and D_{inter}, so it provides a more general approach to determine the actual number of speakers.

Note that the LLRs needed to obtain t_s and ρ are computed during the first iteration of the AHC process. Therefore, these stopping criteria can be considered for both the simplified and the non-simplified PLDA AHC approaches, and there is no need to compute additional likelihoods, so its computational cost is also negligible.

6.3.7 Variational Bayes

The use of PLDA enables us to consider another efficient approach for speaker clustering that do not reduce the search space and thus could be able to find the desired partition even in those cases where the AHC process is discarding it. This can be achieve by means of Variational Bayes (VB). The VB clustering using PLDA provides a framework that automatically detects the number of speakers and merge the i-vectors belonging to the same speaker. The first approach that considered VB for speaker clustering and diarization was presented in [Valente and Wellekens, 2004]. After that, a successful approach that combines recent advances in the field of speaker verification with VB has been presented in [Kenny, 2008] and validated in [Kenny *et al.*, 2010] in situations where the number of speakers is known.

The use of VB approaches enables us to model the partitioning problem with a parametric joint distribution on the the PLDA hidden variables and the mixing coefficients that indicates to which speaker belongs every i-vector. Since the joint distribution is not tractable, we can approximate it by factorizing the distribution and dealing separately with two distributions: the distribution on the i-vectors, which is given by the PLDA model (see Section 3.1.2), and the distribution on the number of speakers and the mixing coefficients i, where for each i-vector n, the corresponding mixing coefficients i_n are defined for every speaker s (i_{ns}) as the a priori probability that the speaker s is the one present in the i-vector n. The distribution on the mixing coefficients for a given i-vector is given by a Multinomial distribution with parameter θ.

Given the factorization, prior distributions are introduced in order to model each one of the parameters considered in each one of the distributions obtained from the factorization. In our case, the parameters of the distribution on the i-vectors are the hidden variables of the PLDA model, y, ε (see Section 3.1.2), and their priors are defined in the PLDA paradigm as Gaussian distributions. The only parameter of the distribution on the mixing coefficients is θ. It is usual on Variational Inference to consider conjugate priors since they lead to posterior distributions having the same functional form as the prior, thus simplifying the Bayesian analysis. Therefore, we choose a Dirichlet distribution $Dir(d_1, ..., d_S)$ as prior for θ, where the parameters $d_1, ..., d_S$ provide prior information related to the times that every speaker is expected to be observed. Usually, there is no prior information on the speakers present in the dataset indicating that one speaker is more likely than the others, so these parameters are set to d with $d = d_1 = ... = d_S$.

This factorization enables us to use the EM algorithm in order to iteratively obtain the posterior distribution on the PLDA hidden variables given the distribution on the mixing coefficients and then obtain the posterior distribution on the number of speakers and mixing coefficients considering the posterior distribution on the i-vectors previously obtained. The EM algorithm maximizes an auxiliary function that is a lower bound of the evidence for the given i-vectors. This process is similar to the one described in [Kenny, 2008], but in this case a PLDA model is considered instead of a JFA model, and i-vectors are considered as input instead of MFCC vectors. Note that the posterior on the mixing coefficients can set the coefficients for some of the speakers to zero, and thus, those speakers are removed from the algorithm. This way, we expect to determine the actual number of speakers [Bishop, 2006].

It is not our purpose to be exhaustive on VB theory, since the derivation of this approach for speaker clustering can be long a tedious. A derivation for a similar problem is presented in [Kenny, 2008], and a good reference for VB can be found in [Bishop, 2006].

The VB approach for speaker clustering is suboptimal in the sense that the joint distribution whose likelihood is maximized, is approximated by factorizing it, and a lower bound of the desired likelihood is maximized since it is not possible to compute the desired likelihood. Regarding the number of partitions considered by this algoritm, during the iterative EM process, this approach can lead to any one of all possible partitions, but it does not evaluate all of them. In fact, since the expected value of mixing coefficients during the EM algorithm is in general between 0 and 1 for every i-vector and hypothetical speaker, no clear information of the partitions evaluated is obtained. In fact, only a the end of the algorithm (when convergence is reached) a real partition, real in the sense that every i-vector is assigned to one and only one speaker, is obtained. Thus, only this final partition is evaluated.

6.4 Evaluation

In this Section the proposed approaches for speaker clustering are evaluated and compared. Firstly, the two PLDA AHC techniques (simplified and non-simplified) are analyzed and then the proposed stopping criteria are tested. In addition, the VB approach is evaluated. The optimal solution for the speaker partitioning partitioning problem is not analyzed since the experimental setup considered makes use of datasets too large to be processed by this approach. The use of large datasets is desired in order to obtain significant error rates, since it is known that the smaller the dataset considered the better the clustering task performs. In all cases the PLDA speaker recognition system described in 3.2 is considered.

6.4.1 Experimental setup

The 1781 recording sides from the *short2* condition of the NIST SRE 2008 database are considered to evaluate the accuracy of the proposed clustering techniques. There are a total of 1319 speakers in the set of 1781 recording sides. This 1781 recording sides come from a total of 1637 two-speaker telephone conversations recorded in stereo, but this fact is ignored in this analysis and the 1781 recording sides are considered as independent recordings (i.e. we do not use the conversation, information so the two recording sides from the same conversation can be clustered together). We do not consider all the sides of the 1637 stereo recordings for two reasons: Firstly, not all the sides have a speaker label. Secondly, the *short2* condition from the NIST SRE 2008 dataset is composed of the 1781 recording sides considered and it is a well-known dataset in the literature.

In those cases where a development dataset is needed, the training dataset for the *1conv4w-1conv4w* core condition of the NIST SRE 2006 database is considered. This dataset is composed of 811 sides of five-minute length telephone conversations, as the *short2* condition of the NIST SRE 2008 database. The set of 811 recording sides contain a total of 594 different speakers. The 811 sides are obtained from 741 two-speaker telephone conversations recorded in stereo, but as in the testing dataset, this information is not considered during the clustering process, and all recording sides are considered independent.

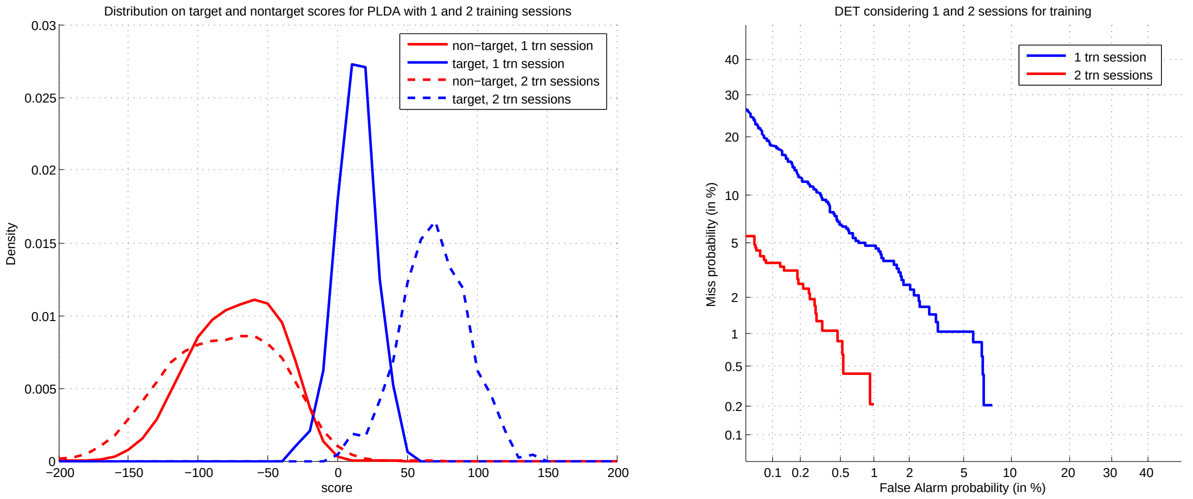
As evaluation measures we consider the Speaker Impurity (I_s) and the Cluster Impurity (I_c) for the partitions obtained. This measures are described in 2.5. In the analysis of the PLDA AHC approach, the results are compared in terms of Equal Impurity (EI), or the impurities obtained for the partition that obtains $I_s = I_c$. This partition has usually a number of clusters that is close to the number of speakers. Also, the impurities for the partition that obtains the actual number of speakers S_{act} are studied. When the stopping criterion is under analysis, the EI has not sense anymore, and the impurities are obtained for the final partition. Also, the number of clusters in the partition where the system decides to stop is compared to S_{act}.

6.4.2 PLDA for speaker clustering

Measure	simplified	non-simplified
EI	11.40%	14.77%
S_{EI}	1362	1318
$I_c(S_{act})$	13.92%	14.71%
$I_s(S_{act})$	10.78%	14.77%

Table 6.4: Accuracy of the PLDA AHC speaker clustering approach, considering the simplified and non-simplified methods, measured in terms of Equal Impurity (EI), number of speakers in the Equal Impurity point S_{EI}, and cluster and speaker impurities obtained for the actual number of speakers ($I_c(S_{act})$ and $I_s(S_{act})$).

Table 6.4 shows the accuracy of the PLDA AHC approach for speaker clustering in the task of agglomerating the recordings of the NIST SRE 2008 *short2* condition, considering the simplified and the non-simplified approaches. The results obtained for the optimal solution are not displayed, since this solution is not feasible for the size of the dataset considered (see Table 6.1). It can be seen that the simplified PLDA AHC approach obtains a significantly



(a) *Normalized target and non-target PLDA score distributions considering 1 and 2 sessions for training.*

(b) *DET curves obtained with the PLDA speaker verification system considering 1 or 2 sessions for training.*

Figure 6.1: *Normalized score distributions and DET curves obtained by the PLDA speaker verification system for a speaker verification task built on the NIST SRE 2008 short2 condition considered to analyze the clustering approaches. The cases where there are 1 and 2 sessions available for training are compared.*

lower EI than the non-simplified one. This is surprising since the non-simplified PLDA AHC should provide more robust scores as larger clusters are obtained. In fact, it is well-known that a speaker verification system obtains much better performance when there are more that one session available for training. This has been observed during the last decade in all NIST SREs involving conditions with multiple segments for training [NIST, 1998].

Training sessions	1 session	2 sessions
EER	2.09%	0.51%
$min(C_{norm})$	0.1152	0.0415

Table 6.5: EER and $min(C_{norm})$ obtained by the PLDA speaker verification system for a speaker verification task built on the NIST SRE 2008 short2 condition considered to analyze the clustering approaches. The cases where there are 1 and 2 sessions available for training are compared.

A deeper study on the scores obtained by the PLDA speaker recognition system considered for this clustering task can help us to understand why the simplified and faster approach gets better results. Figure 6.1 shows the normalized score distributions and the DET curves obtained by the PLDA speaker verification system considered in the clustering PLDA AHC approach. The results are obtained on all possible trials within the NIST SRE 2008 *short2* condition considered to analyze the accuracy of the clustering system. Two cases are analyzed: the case where there are only a single session available for training, and the case where there are two sessions available for training. In all cases only single sessions are considered in the testing sides. When two sessions are considered for training, the two sessions belong to the same speaker. Since the number of speakers that have two sessions in the NIST SRE 2008 *short2* condition is limited, the number of possible trials within the dataset is reduced. Note that the scores obtained considering a single session for training are those obtained during the first iteration of both PLDA AHC approaches (simplified and non-simplified). The scores obtained for two training sessions are the scores that the non-simplified PLDA AHC starts obtaining from the second iteration as the clusters containing a single session are merged building clusters containing two sessions.

The first conclusion that can be drawn is that the use of more than one session for training (or testing, since the PLDA speaker verification system is symmetric) improves the accuracy of speaker verification significantly. This can be observed in the DET curves in Figure 6.1(b) and in Table 6.5, which shows the EER and $min(C_{norm})$ depending on the number of sessions considered for training. This can also be observed in the score distributions in Figure 6.1(a): the overlap between the *target* and *non-target* score distributions is higher when considering a single session for training than when considering two sessions for training. According to these results, the non-simplified PLDA AHC should outperform the simplified one since as larger clusters are obtained, more accurate is the speaker verification system and more reliable are the scores obtained.

However, there is an undesired effect that can be observed in Figure 6.1(a). The score distributions considering one and two sessions for training are misaligned. Considering the score distributions obtained with one training session as reference, the score distribution for *target* considering two training sessions has moved to the right, and effect which is desirable, since it increases the separability of the scores for *target* and *non-target* trials. However, the score distribution for the *non-target* trials has not moved to the left significantly, and there is a tail in the distribution indicating that several *non-target* trials obtains high scores. The presence of this tail makes difficult to set a threshold suitable for a speaker verification task where trials having one and two training sessions coexist. This is the case of the non-simplified PLDA AHC approach. Once the clustering algorithm starts merging sessions, new clusters containing two or more sessions are obtained. The comparison of clusters containing a single session with clusters containing more that one session can obtain very high scores even if the clusters contain different speakers, as we have seen in Figure 6.1(a). These scores can be higher than the scores that the PLDA system would obtain comparing two clusters containing a single session even if both clusters belong to the same speaker. This means that, in many cases, large clusters containing different speakers will be merged before merging small clusters containing the same speaker, and thus, it is likely to obtain a impure cluster in the early stages of the clustering process. Once a impure cluster is available, the accuracy of the clustering task degrades iteration after iteration.

In fact, analyzing the obtained scores, it has been detected that the highest score for the *non-target* trials considering two sessions for training is higher than the highest score for the *target* trials considering one session for training. Therefore, the non-simplified PLDA AHC approach tends to increase the size of the larger clusters adding segments belonging to different speakers while the simplified approach is more robust against this effect, since the scores are always computed considering a single session for training and testing.

One solution to the misalignment in the score distributions that appears in the non-simplified PLDA AHC approach is to normalize and calibrate the system independently for every possible cluster size. However, this operation is tedious and may not be feasible when the size of the clustering problem is high.

Therefore, we consider the simplified PLDA AHC approach for the rest of this work, since it shows higher accuracy and is less expensive in terms of computational cost.

6.4.3 Stopping Criteria

The simplified PLDA AHC approach enables us to build a cluster tree that contains several clustering hypotheses or hypothetical partitions of the dataset, fast and easily. This tree, that will have the form as the one presented in Chapter 2, Figure 2.3, does not contain all possible hypothetical partitions, but those that are more likely for the bottom-up hierarchical clustering paradigm. These hypothetical partitions start from the finest partition and keep reducing the number of clusters in the partition, one by one, until the coarsest partition is reached. Thus, every partition has an associated number of speakers, and the task of selecting the desired partition of those simplified PLDA AHC approach is equivalent to the task of determining the number of speakers in the dataset.

For the purpose of determining the number of speakers in the dataset, several stopping criteria have been proposed. These criteria are evaluated and compared in the following subsections.

6.4.3.1 Threshold on the Likelihood Ratio

The first proposed stopping criterion makes use of a threshold for the distance metric considered, the PLDA score (LR) in this case. In Section 6.3.2, it is shown that in a PLDA AHC system, there is no need to compute the likelihood for every partition, but simply the LRs for every partition with respect to its initial partition, that is, the partition having one more speaker in the PLDA AHC tree. Therefore, the selection of the partition with maximum likelihood is equivalent to consider a stopping criterion that assumes that the clusters are merged until the LLR is below 0. In fact, if the LLR is below 0, it means that in the last iteration two clusters were merged and this is decreasing the likelihood of the partition. Although this is true for the non-simplified PLDA AHC approach, the same procedure can be applied to the simplified approach. However, it is not guaranteed that this stopping criterion selects the most likely partition out of those available in the PLDA AHC tree obtained by the simplified approach. In any case, as first approximation, we consider as stopping criterion that the LLR must be over 0.

Measure	NIST SRE 2008
S	535
I_c	65.69%
I_s	2.75%

Table 6.6: Accuracy of the simplified PLDA AHC speaker clustering approach, considering the threshold based stopping criterion, setting the threshold to 0. The accuracy is measured in terms of number of speakers determined by the stopping criterion (S), and cluster and speaker impurities obtained for the determined number of speakers (I_c and I_s).

Table 6.6 displays the accuracy of the simplified PLDA AHC speaker clustering approach, considering the threshold based stopping criterion, setting the threshold to $\epsilon = 0$. The obtained number of speakers S is far below the actual number of speakers S_{act}, which means that in many cases, several sessions belonging to different speakers are merged together. This is also reflected in the high value of the cluster impurity I_c.

Therefore, the threshold value $\epsilon = 0$ does not seem appropriate for this task. In order to find a suitable threshold value ϵ for the LLRs produced by the simplified PLDA AHC, the

Measure	NIST SRE 2006	NIST SRE 2008
EI	4.44%	11.40%
S_{EI}	600	1362

Table 6.7: Comparison of the accuracy of the simplified PLDA AHC speaker clustering approach between the NIST SRE 2006 and 2008 datasets considered for development and testing. The accuracy is measured in terms of EI and number of speakers determined in the EI point S_{EI} .

training dataset for the *1conv4w-1conv4w* core condition of the NIST SRE 2006 database is considered for development. Table 6.7 compares the accuracy of the simplified PLDA AHC speaker clustering approach between the NIST SRE 2006 and 2008 datasets considered for development and testing. Note that the EI obtained for the development dataset is significantly lower than that obtained for the testing dataset. The value ϵ is determined as the one that enables us to select the partition that contains a number of clusters identical to the actual number of speakers $\epsilon(S_{act})$.

Measure	NIST SRE 2006	NIST SRE 2008
$\epsilon(S_{act})$	11.60	10.02
$S(\epsilon_{2006}(S_{act}))$	594	1401
$I_c(\epsilon_{2006}(S_{act}))$	5.18%	8.87%
$I_s(\epsilon_{2006}(S_{act}))$	4.19%	11.85%

Table 6.8: Accuracy of the simplified PLDA AHC speaker clustering approach, considering the threshold based stopping criterion, using the NIST SRE 2006 data as development to set the optimal threshold to determine the actual number of speakers ($\epsilon(S_{act})$). The accuracy is measured in terms of number of speakers determined by the stopping criterion ($S(\epsilon_{2006}(S_{act}))$), and cluster and speaker impurities obtained for the determined number of speakers ($I_c(\epsilon_{2006}(S_{act}))$ and $I_s(\epsilon_{2006}(S_{act}))$).

The accuracy of the simplified PLDA AHC speaker clustering approach, considering the threshold based stopping criterion with ϵ trained on the development dataset extracted from NIST SRE 2006 is presented in Table 6.8. In the first entry, the values of $\epsilon(S_{act})$ for the development and testing datasets are displayed. It can be seen that they are not far from each other, given the wide range of values of the PLDA scores (see Figure 6.1(a) as an example of the range of the PLDA LLRs), so one can be used on the other dataset and vice-versa. Considering the threshold value obtained on the development dataset $\epsilon_{2006}(S_{act})$, the number of speakers obtained for the testing dataset is not far from S_{act}, 1401 against 1319. In this case, the stopping criterion is fulfilled before S_{act} is reached, and some clusters belonging to the same speaker are not merged together. This is reflected in the speaker impurity obtained, $I_s(\epsilon_{2006}(S_{act}))$, which is slightly higher than that obtained for the partition that corresponds to S_{act} (11.85% against 10.78%, see Table 6.4).

In general, since the development and testing conditions may present differences, there will be a slight misalignment between the estimated value of $\epsilon(S_{act})$ and the actual one. In this experiments this misalignment is tolerable, but in order to keep it tolerable in real conditions, it is necessary in general to obtain a development dataset which is not very different from the conditions that the clustering system will face.

6.4.3.2 Student's T-test

The Student's t-test is also considered as stopping criterion. Again a simplified PLDA AHC system is considered to build the clustering tree, but the partition selected this time will be the one that maximizes the t-test criterion described in Section 6.2.2. The t-test statistic is computed assuming that the score distributions are Gaussian, but the measure of separation between the distributions of the intra-cluster scores and inter-cluster scores ρ proposed in [Nguyen *et al.*, 1998] and described in Section 6.2.2 is also considered.

Measure	T-test Gaussian dist. (T_s)	T-test unknown dist. (ρ)
S	514	1779
I_c	67.38%	0.00%
I_s	2.64%	25.83%

Table 6.9: Accuracy of the simplified PLDA AHC speaker clustering approach, considering the t-test stopping criterion, assuming Gaussianity and unknown distribution for the PLDA scores. The accuracy is measured in terms of number of speakers determined by the stopping criterion (S), and cluster and speaker impurities obtained for the determined number of speakers (I_c and I_s).

Table 6.9 shows the accuracy of the simplified PLDA AHC speaker clustering approach, considering the t-test stopping criterion, assuming Gaussianity and unknown distribution for the PLDA scores. In both cases the accuracy is poor compared to that obtained considering the LLR calibration as stopping criterion. The problem that the t-test finds when Gaussian distributions are assumed is that the t-test statistic increases as more separated are the distributions considered, but it also takes into account the uncertainty of the distributions estimated on the given data. For a given set of score samples, as more balanced are the number of score samples considered to estimate every distribution, higher is the t-test statistic value. This makes sense since a distance metric between two distributions when one is poorly estimated (estimated with little data) may not be reliable, so it is penalized. However, in our clustering problem S_{act} is close to N, so for the desired partition, the number of score samples for the intra-cluster score distribution is much lower than the number of score samples for the inter-cluster score distribution. This leads to the detection of a number of speakers much lower than the actual one, and thus to a high I_c.

This t-test stopping criterion assuming Gaussian score distributions will probably have a better behavior in problems where the actual number of speakers is much lower than the initial number of clusters, situation that is usual in speaker diarization tasks. However, in the task of partitioning a given dataset by speaker, the number of speakers in the dataset can be as large as the number of sessions in the dataset, so this criterion does not seem to be suitable.

On the other hand, the t-test stopping criterion considering that the score distribution is unknown, leads to the opposite situation. The main problem is that this criterion selects the partition that maximizes the value of ρ, which is a measure of the separation or the absence of overlap between the intra-cluster and inter-cluster score distributions. Since the clustering process merge those clusters obtaining the highest score, during the first iterations, the intra-cluster and inter-cluster score distributions are not overlapped and $\rho = 1$. Considering the simplified PLDA AHC approach, once a cluster containing more than two sessions is created, there is a score in the intra-class distribution that has not been selected as maximum score, and it is likely to be below some inter-cluster scores. Thus, this stopping criterion hardly ever will select a partition that merges three or more sessions in a single cluster, unless the desired partition keeps the intra-cluster and inter-cluster score distributions separated.

Therefore, both t-test stopping criteria present problems that make them not suitable for the task of partitioning a dataset by speaker. In [Nguyen *et al.*, 1998], these problems are partially solved increasing the number of scores by simply splitting the available sessions and obtaining intra-cluster and inter-cluster score samples even for the finest partition. This way, the initial number of intra-cluster score samples is much higher but the proportion between intra and inter-cluster score samples will be similar. This may solve the problem observed when the score distributions are assumed to be Gaussian.

Splitting the available sessions may also solve the problem that the t-test stopping criterion presents under the assumption of no Gaussianity of the score distributions, since the initial partitions will hardly ever obtain a $\rho = 1$. The problem of this solution for the task of speaker partitioning in large datasets is that the number of PLDA score computations increases dramatically (note that the number of score computations is $O(N^2)$, and the number of i-vector extractions to be performed is $O(N)$). Therefore, this solution is not efficient unless very small problems are considered, which is not the case, in general, for the speaker partitioning problem in large datasets.

This solution might be suitable for speaker diarization, but still the computational cost will be high, and i-vector estimation would not be reliable as the length of the segments considered decreases, obtaining also less reliable scores to determine the optimal partition.

6.4.4 Variational Bayes

Another approach for speaker clustering under test is the PLDA VB method. This approach presents two main design issues. First, prior distributions must be defined for all the parameters present in the factorized distributions. These prior distributions are defined in Section 6.3.7, including those for the PLDA latent variables which are Gaussian with zero mean and identity as covariance matrix, and also for the θ parameter of the Multinomial distribution on the mixing coefficients, which follows a $Dir(d)$. The d parameter is a design parameter that is related to the number of i-vectors that are assumed to have been seen to belong to each one of the initial clusters in the past. Note that low values for this parameter encourages the clustering process to merge i-vectors, removing initial clusters, while high values encourages the clustering process to keep the initial clusters, not merging them.

The other design issue is to set the initial speaker clusters. This approach is not a hierarchical one where the merged clusters are not separated anymore, but as in a hierarchical clustering method, the final number of speakers cannot exceed the initial number of clusters. Thus, it seems reasonable to consider as many clusters as sessions are available to be clustered. In all the experiments presented in this section, the initial number of speakers or clusters considered is equal to the number of i-vectors in the dataset $S_{ini} = N = 1781$.

In addition, it is important to decide the initial values for the mixing coefficients, since as we will see later, considering the finest partition as initial partition does not lead to satisfactory results. Two strategies are adopted to initialize the mixing coefficients i:

- Finest: As first approach we consider that the initial partition is the finest partition, that is:

$$i_{ns} = 1 \iff n = s \tag{6.20}$$

$$i_{ns} = 0 \iff n \neq s \tag{6.21}$$

- Balanced: We also consider that all i-vectors are shared by all initial speakers, existing a unique i-vector for every speaker that is twice as probable to contain the speaker as the other i-vectors. This way we encourage the clustering process to merge the clusters since they are quite similar a priori.

$$i_{ns} = \frac{2}{N+1} \iff n = s \tag{6.22}$$

$$i_{ns} = \frac{1}{N+1} \iff n \neq s \tag{6.23}$$

Init. Strategy	Finest		Balanced	
d value	0.001	1000	0.001	1000
S	1781	1781	84	97
I_c	0.00%	0.00%	92.20%	91.07%
I_s	25.94%	25.94%	11.68%	12.63%

Table 6.10: Accuracy of the VB approach for speaker clustering for two initialization strategies for the mixing coefficients, setting the Dirichlet parameter to $d = 0.001$ and to $d = 1000$. The accuracy is measured in terms of number of speakers determined by the stopping criterion (S), and cluster and speaker impurities obtained for the determined number of speakers (I_c and I_s).

Table 6.10 shows the accuracy of the VB approach for speaker clustering obtained for both initialization strategies proposed and for two values of the Dirichlet free parameter d: a high value that discourages to merge clusters ($d = 1000$) and a low value that encourages to merge clusters ($d = 0.001$). Both initialization strategies and both d values obtain very poor accuracy, compared to a simplified PLDA AHC system setting a threshold for the LLR as stopping criterion. If we consider the finest partition as initialization, the VB system never merge two i-vectors, independently from the value of d. The output partition is always the finest partition.

On the other hand, giving non-zero values to all the mixing coefficients encourages the VB approach to merge the clusters, obtaining a very low number of speakers S, much lower than the actual number of speakers ($S_{act} = 1319$). The value of d has little impact on the clustering accuracy. In addition, I_s is very high, comparable to that obtained considering the simplfied PLDA AHC for the values of S close to S_{act}. This high value of I_s is telling us that the system is not merging the clusters correctly. This is probably due to the same effect that degrades the accuracy of the non-simplified PLDA AHC: the LRs obtained by the PLDA for clusters containing different number of i-vectors may not be in the same range.

From these results, we can conclude that the PLDA VB approach for clustering is highly dependent on the initialization strategy considered, so it does not provide a robust methodology to perform the clustering task. It seems that the method tends to get stuck on local maxima, and thus different initialization will provide completely different results. Further work must be carried out in order to take advantage of the PLDA and VB frameworks for speaker clustering, for example, deterministic annealing [Rose, 2008] can be used in order to provide robustness against local maxima. However, note that the number of possible solutions that this approach may find when the initial number of clusters is N is as huge as N^N which is higher than $B(N)$, since for a given partition, all possible permutations of the clusters in the partition over all the initial speakers are possible. This might be an explanation of the vulnerability of this approach to local maxima.

As conclusion of this analysis, we can affirm that the best clustering strategy to solve the speaker partitioning problem in large datasets, among all the strategies presented, is considering the simplified PLDA AHC for iteratively cluster merging, and obtaining a score threshold on a development dataset as stopping criterion.

6.5 Speaker Diarization with Unknown Number of Speakers

In the previous chapters, an innovative speaker diarization system has been presented and analyzed on the task of speaker diarization in telephone conversations, where the number of speakers is priorly known and equal to two. However, in general diarization problems, the number of speakers is unknown and must be determined by the speaker diarization system. In this section, the speaker diarization system proposed in Chapter 4 with intra-session variability compensation (see Chapter 5) is combined with speaker clustering strategies in order to face the general diarization problem. As approaches for speaker clustering, both traditional techniques considered for diarization and novel strategies are considered. The diarization system is evaluated with the assumption of not knowing the number of speakers in order to determine the best clustering techniques for this task, and finally compared to a traditional diarization system.

6.5.1 Diarization system description and configuration

Figure 6.2 shows the block diagram of the proposed approach for speaker diarization including intra-session variability compensation and speaker clustering to face problems where the number of speakers is unknown. The different stages and modules involved in the diarization process are described next.

- **Front End:** The *heavy-weight* configuration is considered since it has been shown to be more accurate (see Chapter 4). This configuration considers 19 MFCC plus delta as features. Since we are considering an stage for intra-session variability compensation that includes LDA for dimensionality reduction, a total of $R = 100$ speaker factors are extrated for every frame.

- **Variability Compensation:** To compensate for intra-session variability in the obtained stream of speaker factors, the LDA 100 → 50 + WCCN strategy presented in Chapter 5 is considered. In this strategy, LDA and WCCN transformations are trained using some development data composed of recordings containing a single speaker (NIST SRE 2004, 2005 and 2006), and considering that the speaker factor

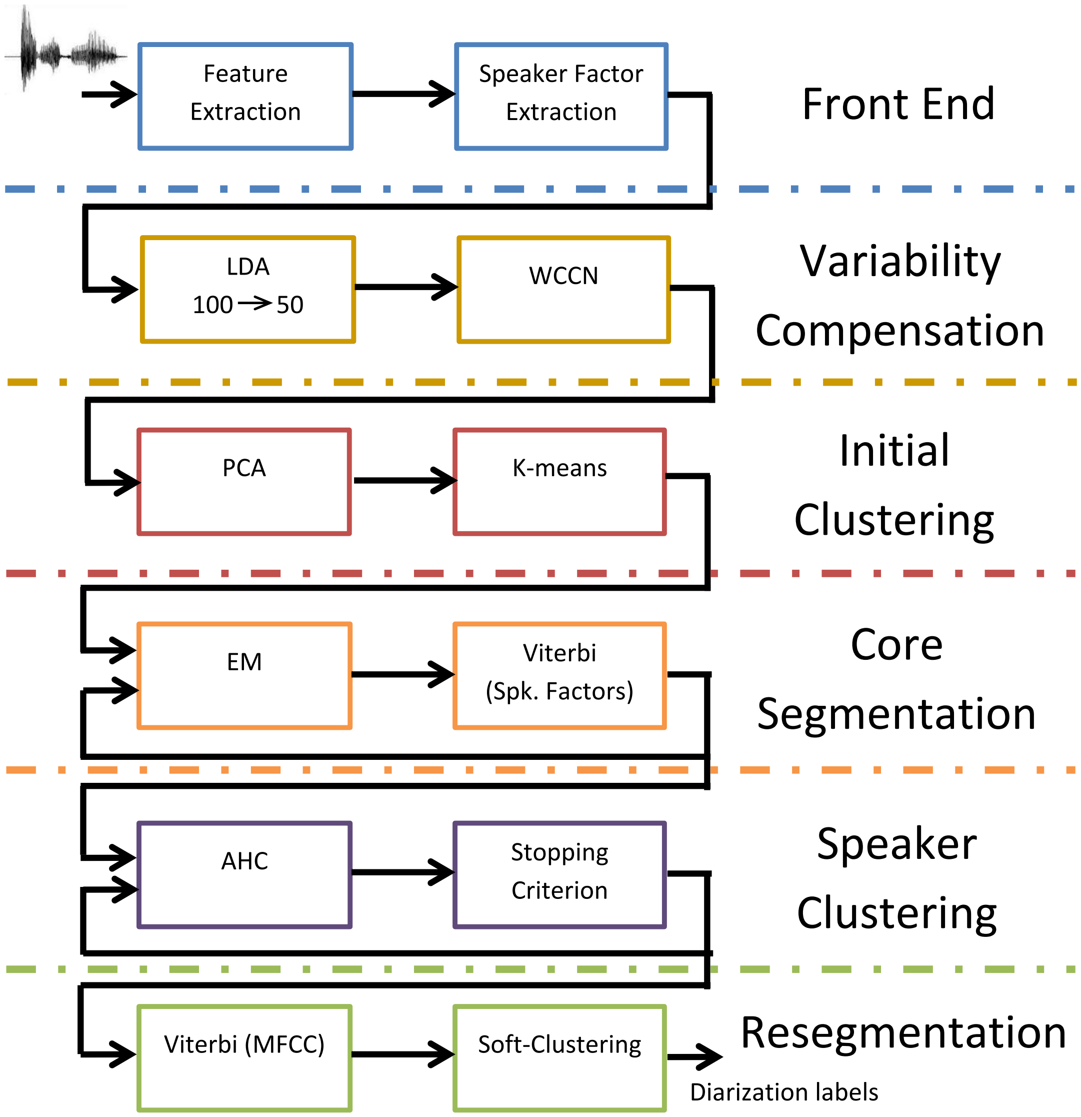


Figure 6.2: *Block Diagram of the proposed diarization system for problems with unknown number of speakers*

vectors extracted from every recording belong to a unique class. LDA is trained to reduce the dimensionality of the speaker factor vectors to 50, and then WCCN is applied.

- **Initial Clustering:** The initial clustering strategy is similar to the one described in Section 4.1.2, but, in this case, since the number of speakers in the recording under analysis is priorly unknown, a number of clusters N higher than the expected number of speakers present in the recording is considered. We set $N = 10$, assuming that the recordings under test will never contain more that 10 different speakers. However, for long recordings where this number could be small, it can be increased, or the recordings can be processed in large chunks (for example 5 minute length) where it is not expected to find such a high number of speakers.

  First, PCA is considered to find the $N-1$ directions of maximum variability among the input set of speaker factor vectors. The input speaker factor vectors are projected onto these directions and K-means are applied to obtain a first clustering hypothesis with N classes. Then, from this clustering hypothesis, the initial centroids for a second K-means process are obtained, and the complete speaker factor vectors, non-transformed with PCA, are grouped into N clusters during the second K-means algorithm. Note that this process is exactly the same as the one described in Section 4.1.2, but in that case $N = 2$ and only $N-1 = 1$ direction is considered for the first K-means algorithm.

- **Core Segmentation:** The core segmentation stage is identical to the one described in Section 4.1.3, but, in this case, $N = 10$ clusters are considered as input, so the stream of speaker factor vectors is segmented and classified into 10 classes instead of 2.

- **Speaker Clustering:** This stage is introduced to cluster the segments corresponding to the N different classes obtained as output of the core segmentation stage according to the speaker present in every cluster. It is composed of two modules. the first module (AHC in Figure 6.2) computes a distance metric between all possible pairs that can be obtained from the available clusters, and merge the closest pair of clusters. The second module checks whether the stopping criterion is fulfilled and thus determine the final number of speakers. The distance metrics and stopping criteria considered for speaker clustering are described later.

  The output of this stage is a set of S clusters with $1 \leq S \geq N$ that corresponds to different hypothetical speakers. Thus S is an estimate of S_{act}, the actual number of speakers in the recording.

- **Resegmentation:** The last stage is the resegmentation stage which is identical to that described in Section 4.1.4, but, in this case, a total of S speakers are considered, instead of setting $S = N = 2$ as in the original system.

6.5.1.1 Strategies for speaker clustering

For the speaker clustering stage, a total of fifteen different strategies are studied, obtained as combination of five distance metrics with three stopping criteria. The distance metrics considered are explained next:

- **ΔBIC:** As baseline distance metric the ΔBIC is considered. This metric is described in Chapter 2, and is the most popular one in traditional diarization systems. The ΔBIC is computed considering 12 MFCC without delta features, and modeling every cluster with a full covariance Gaussian. Every time two clusters are merged, the BIC value is recomputed, as it is usual in traditional BIC AHC speaker diarization systems. Note that this clustering strategy is exactly the same as the one considered in the BIC AHC diarization system described in Section 3.4.1. However in this case, the clustering stage does not stop when $S = 2$, but when the stopping criterion is fulfilled.

- **Simplified PLDA AHC:** The score of the PLDA speaker verification system described in Section 3.2 is considered as distance metric. In this case the PLDA scores are not recomputed every time two clusters are merged, but the initial scores are considered instead. Thus, this clustering strategy is the one referred as simplified PLDA AHC described in Section 6.3.3 and validated in Section 6.4.2.

- **Simplified PLDA AHC, intra-session compensation:** A PLDA model can be built to compensate for intra-session variability instead of compensating for inter-session variability. in this case, the PLDA system will not be a speaker recognition system, but a system capable of identifying segments belonging to the same speaker within a session, removing the variability that the speaker presents within that session, as explained in Section 5.2. In this case, for each one of the N clusters obtained as output of the core-segmentation stage, a speaker factor vector of dimension $R = 100$ is obtained, using the speaker factor extraction module in the front end stage of the speaker diarization system. This vectors are fed into the PLDA system, and the same procedure as in the simplified PLDA AHC strategy is followed.

- **Cosine distance, intra-session compensation:** Obtaining a speaker factor vector for every class enables us to consider other clustering strategies. A fast and straightforward approach is to use the cosine distance among the speaker factor vectors as distance metric. Intra-session variability is compensated using the same LDA 100 → 50 + WCCN strategy considered for speaker diarization, and the cosine distance is computed as in an i-vector speaker recognition system [Dehak *et al.*, 2010].

- **Euclidean distance, intra-session compensation:** Another distance metric that is fast and straightforward to compute once the speaker factor vectors for every segment are extracted and intra-session variability is compensated, is simply to compute the distance metric as the euclidean distance among the compensated speaker factor vectors. This metric is motivated by the high accuracy obtained by the K-means clustering strategy in the initial clustering stage.

These distance metrics are combined with three stopping criteria:

- **Ideal stopping criterion:** To analyze the best results that can be obtained for a given distance metric, the ideal stopping criterion is considered. This criterion is assumed to always select the partition that hypothesizes the actual number of speakers $S = S_{act}$. This stopping criterion enables us to compare distance metrics for clustering assuming that the number of speakers is known.

- **Threshold calibrated on a development dataset:** The stopping criterion that best accuracy has given in the task of speaker partitioning of large datasets is also a good candidate for the task of speaker clustering in diarization systems. To obtain the threshold a development dataset composed of recordings containing a known number of speakers is considered, and the threshold is set in order to detect the correct number of speakers for the maximum number of recordings. Note that this is the stopping criterion in most of the traditional diarization systems, for example in BIC AHC based diarization systems, it is usual to set the threshold to $\epsilon = 0$ and to adjust the value of the penalty parameter *lambda*, which is an implicit way of setting a threshold different from $\epsilon = 0$.

- **T-test under Gaussianity assumption:** The t-test stopping criterion under Gaussianity assumption is not suitable for the partitioning of large datasets when the actual number of speakers S_{act} is close to the initial number of clusters N. However, it may be suitable for problems where S_{act} is low compared to N. This situation is usual in speaker diarization. Thus, the t-test is also considered as stopping criterion for the proposed diarization system.

Note that to evaluate the proposed speaker clustering approaches for the speaker partitioning problem in large datasets, in Section 6.4, a dataset containing a S_{act} close to N has been considered. In diarization problems, S_{act} is usually significantly lower than N. This two extreme situations have been selected in order to see whether there is a speaker clustering solution valid for all possible situations.

6.5.2 Experimental setup

As datasets to evaluate the accuracy of the speaker diarization system previously described, we consider the NIST SRE 2008 *summed* dataset described in Section 3.2.1, as in previous analysis of diarization systems. All telephone conversations in this dataset contains two speakers, but in this analysis the conversations are processed as if the number of speakers were unknown. In addition, the *callhome* dataset from the NIST SRE 2000 is considered. This dataset was included in the NIST SRE 2000 to evaluate speaker diarization systems over telephone conversations. It is composed of 500 conversations with durations that vary from 40 seconds to 10 minutes. The number of speakers present in every conversation oscillates between 2 and 7. In those cases where a development dataset is needed, as for the threshold based stopping criterion, the NIST SRE 2008 *summed* dataset will be considered for development, and the technique under test will be validated on the *callhome* dataset. The accuracy of the proposed diarization system is measured in terms of DER, as usual.

The degradation that the proposed speaker diarization system introduces in a speaker verification task is also evaluated under the assumption of unknown number of speakers. For this purpose, the speaker verification task defined in Chapter 3 is considered. The same experimental setup described in Section 3.2 is utilized and the verification task is evaluated in the three scenarios considered to analyze the diarization system in previous Chapters: the *mono-stereo*, *stereo-mono* and the *mono-mono* scenarios. In this case, it is assumed that any mono conversation contains an unknown number of speakers that must be determined by the speaker diarization system.

6.5.3 Evaluation of speaker diarization with unknown number of speakers

	Stopping Criteria		
Measure	Ideal	Threshold	T-test, Gaussian
BIC AHC	1.88%	3.25%	23.71%
PLDA AHC	7.18%	12.77%	20.36%
PLDA AHC intra-ses.	7.61%	15.47%	13.45%
Cosine dist.	9.10%	17.06%	17.17%
Euclidean dist.	10.90%	18.04%	22.05%

Table 6.11: DER for the NIST SRE 2008 summed dataset, for several clustering strategies and stopping criteria.

Table 6.11 shows the accuracy of the proposed diarization system, measured in terms of DER, on the NIST SRE 2008 *summed* dataset, for several clustering strategies and stopping criteria. Analyzing the column corresponding to the ideal stopping criterion, it can be seen that the clustering strategy that obtain the best results is the BIC AHC. The results obtained for the strategies based on PLDA are similar, and these results are better than those obtained considering the cosine distance as metric. The later metric shows higher accuracy than the clustering strategy based on euclidean distance.

The assumption that the number of speakers is unknown degrades severely the results. In such a situation, the best strategy is to combine the BIC AHC with a threshold based stopping criterion, which is the traditional solution for speaker diarization. The t-test stopping criterion does not seem to work. In any case, the diarization system described in Section 4 with intra-session variability compensation (see Chapter 5) obtains a DER of 1.31% for the same configuration studied here, so the degradation is quite significant.

From all the results presented that assume an unknown number of speakers, the only one that is competitive combines BIC AHC with a threshold based stopping criterion, as traditional diarization systems, and obtains a DER of 3.25%. Note that the threshold is determined considering the same dataset as development dataset, thus, this is an optimistic situation. However, it is known that the threshold in BIC AHC systems is robust as far as the development and testing datasets are not significantly different.

The strategies proposed are also validated in the NIST SRE 2000 *callhome* dataset, which presents more difficulty than the NIST SRE 2008 *summed* dataset since its recordings may contain more than two speakers. In addition, it is neccesay to validate the threshold based stopping criterion in a dataset different from the one considered for development.

	Stopping Criteria		
Measure	Ideal	Threshold	T-test, Gaussian
BIC AHC	13.76%	19.89%	36.59%
PLDA AHC	18.46%	33.86%	35.36%
PLDA AHC, intra-ses	16.69%	26.34%	28.25%
Cosine dist.	18.63%	29.92%	31.73%
Euclidean dist.	18.78%	29.86%	33.04%

Table 6.12: DER for the NIST SRE 2000 callhome dataset, for several clustering strategies and stopping criteria.

Table 6.12 shows the the accuracy of the proposed diarization system, measured in terms of DER, on the NIST SRE 2000 *callhome* dataset, for several clustering strategies and stopping criteria. Again, it can be seen that the BIC AHC is the best clustering strategy, outperforming the rest of the proposed strategies when considering the ideal and the threshold based stopping criteria. The simplified PLDA AHC with intra-session variability compensation obtains also interesting results, but it is still far from the BIC AHC strategy, except for the t-test stopping criterion. The t-test criterion seems to work better for strategies based on speaker verification techniques than for the BIC AHC, but its accuracy is still not enough to outperform the BIC AHC with threshold stopping criterion.

Therefore, given these results, it seems that the best approach is to combine BIC AHC with a threshold stopping criteria to face problems with unknown number of speakers. This is actually the solution in traditional speaker diarization systems. The recent advances in the field of speaker recognition can be applied for the task of speaker clustering but they do not seem to outperform the traditional BIC. This is probably because the BIC AHC strategy uses models trained on data directly extracted from the recording to process. Speaker recognition based techniques are accurate in speaker verification and speaker clustering in large datasets since they provide robust methods to deal with undesired variability. However, in the task of speaker clustering for speaker diarization, within a single session, there is little undesired variability, and thus the simple speaker models considered by BIC obtain very good performance. Note that we do not expect to find sources related to inter-session variability within a single recording when performing diarization, and that intra-session variability is critical for very short segments (it helps for the initial clustering stage, that considers 1 second segments), but its impact is reduced as larger segments are considered, as in this speaker clustering task.

In addition, speaker recognition techniques make use of complex models that need to estimate several parameters, and their performance is degraded significantly as the size of the segment is reduced. Therefore, for the proposed task, the simplest and non-compensated speaker models that the BIC AHC strategy considers seems to perform better. Further research is needed in order to determine a size of the speaker recognition model (for example PLDA) and also to determine how to deal with undesired variability that might be still present among large segments from the same speaker.

From the previous results, it can also be concluded that prior knowledge on the number of speakers can be used in the initial clustering stage rather than in the speaker clustering stage (see Figure 6.2): the former solution provides a DER of 1.31%, while the later increases the DER to 1.88%. However, even if the number of speakers is known, we expect the speaker factor system to reduce its accuracy when there are more than 2 speakers present in the recording.

System	DER for *summed*	DER for *callhome*
Baseline BIC AHC	5.21%	19.76%
Spk factors, no spk clustering	1.31%	14.60%
Spk factors, spk clustering, ideal stop	1.88%	13.76%
Spk factors, spk clustering, threshold stop.	3.25%	19.89%

Table 6.13: DER for the NIST SRE 2008 summed and NIST SRE 2000 callhome datasets assuming that the number of speakers present in each recording is priorly known. Three systems are compared: the baseline BIC AHC, the speaker factor diarization system without and with speaker clustering. In addition, the results considering the threshold based stopping criterion for the last system are shown

Table 6.13 compares three different speaker diarization systems on the two datasets under analysis. The first system compared is the baseline BIC AHC system described in Section 3.4.1. The second one is the speaker factor based diarization system presented in Chapter 4, considering the *heavy-weight* configuration with intra-session variability compensation (LDA 100 → 50 + WCCN) as described in Chapter 5. The last one is the speaker factor based diarization system with speaker clustering described in this Chapter in Section 6.5.1, considering BIC as distance metric for clustering. The results are obtained assuming that the number of speakers on each testing recording is priorly known. Also, for comparison, the results obtained considering the last system when the number of speakers is priorly unknown are presented, considering the threshold based stopping criterion. The threshold is obtained on the NIST SRE 2008 *summed* dataset. Results are presented on NIST SRE 2008 *summed* and NIST SRE 2000 *callhome* datasets.

It can be seen that in all cases, when the number of speakers is assumed to be priorly known, the speaker factor based systems outperform the BIC AHC based one. For the NIST SRE 2008 *summed* dataset, the best results are achieved considering the speaker factor diarization system without speaker clustering. Thus, it seems that for two speaker problems, the best strategy is to search for both speakers during the initial clustering stage, avoiding the use of any speaker clustering approach later. However, for the NIST SRE 2000 *callhome* dataset, where the number of speakers is greater than two in several cases, the best strategy seems to be starting with a high number of speakers in the initial clustering stage, and then using a speaker clustering stage to merge the initial clusters. This is interesting, since it shows that the scheme proposed in Chapter 4 is valid for two speakers, but for an increasing number of speakers, traditional clustering strategies yield better accuracy. Finally, note that the assumption of not knowing the number of speakers degrades the accuracy but the results are still competitive.

6.5.4 Impact on speaker verification

In this Chapter a speaker diarization system capable to face situations where the number of speakers is priorly unknown has been presented. This system is similar to the one proposed in Chapter 4, but it starts with a high number of initial clusters and includes a speaker clustering stage that estimates the actual number of speakers and merges the initial clusters to obtain the estimated number of speakers. It has been shown that this approach shows significant degradation compared to the system proposed in Chapter 4, which assumes that the number of speakers is known and equal to two. However, in situations where the number of speakers is unknown, the last system cannot be used, and such a degradation must be assumed.

In this section, the impact of the mentioned degradation in the speaker verification task defined in Chapter 3 is analyzed. The same experimental setup described in Section 3.2 is utilized and the verification task is evaluated in the three scenarios considered to analyze the speaker factor system in Section 4.3.2: the *mono-stereo*, *stereo-mono* and the *mono-mono* scenarios. Four diarization systems are compared: the ideal diarization system, the BIC AHC baseline described in Section 3.4.1, the *heavy-weight* speaker factor diarization system described in Chapter 4 with the LDA 100 → 50 + WCCN intra-session variability compensation strategy as described in Chapter 5, and the system proposed in this Chapter, in Section 6.5.1, considering a BIC AHC clustering strategy with a threshold based stopping criterion. Note that the first three systems assume that the number of speakers is known and equal to two, while the last one assumes that is unknown and has to be estimated.

6.5.5 *Mono-Stereo* Scenario

Table 6.14 compares the accuracy obtained by the BIC AHC baseline and the speaker factor diarization system assuming that the number of speakers S is known $S = 2$, to the accuracy obtained by the proposed speaker factor diarization system with speaker clustering assuming that S is unknown. The accuracy is evaluated on the subset of the *summed-short2* condition of the NIST SRE 2008 considered for enrollment in the *mono-stereo* scenario. It can be seen that the speaker factor diarization system obtains higher accuracy when S is priorly known. But, even assuming that S is unknown, the accuracy of the speaker factor diarization system is higher than that obtained by the BIC AHC baseline system given that $S = 2$.

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline $S = 2$	4.92%	1021 (75.13%)	1163 (85.58%)
Speaker Factors, no spk clust. $S = 2$	**1.11%**	**1307 (96.17%)**	**1335 (98.23%)**
Speaker Factors, spk clust. S unknown	2.95%	1194 (87.86%)	1241 (91.32%)

Table 6.14: Comparison of the accuracy obtained with the BIC AHC baseline system, the speaker factor diarization system assuming that the number of speakers S is known ($S = 2$) and the speaker factor diarization system with speaker clustering assuming that S is unknown. The systems are evaluated on the enrollment subset of the *summed-short2* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

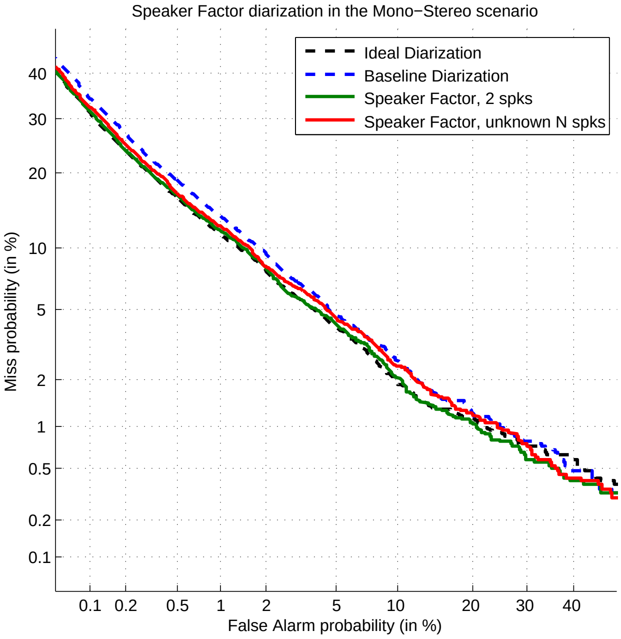
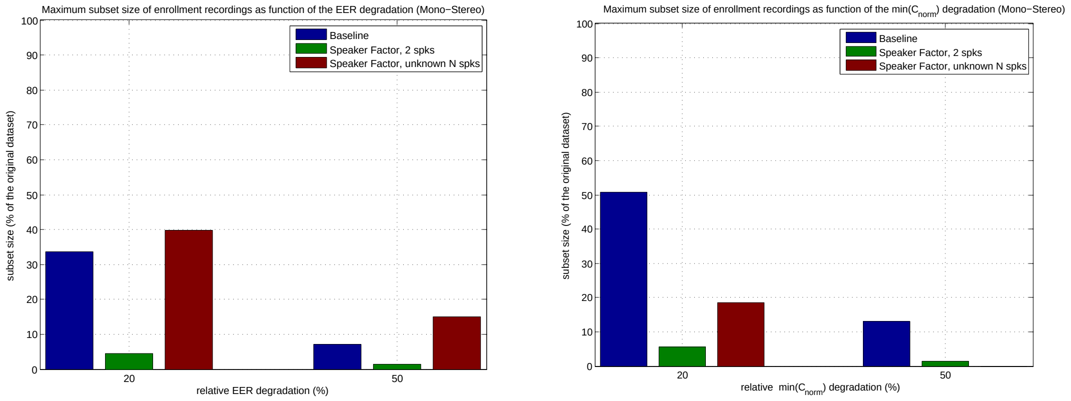


Figure 6.3: *DET curves considering the speaker factor diarization system assuming that S is known (S = 2) and unknown, in the mono-stereo scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

Figure 6.3 shows the DET curves obtained in the *mono-stereo* scenario considering the speaker factor system to diarize the enrollment dataset, assuming that S is known and unknown. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. It can be observed that the degradation introduced by the speaker diarization system with the assumption of not knowing the number of speakers in terms of DER is also reflected in the DET curves. The speaker factor system obtains higher accuracy in the speaker verification task if the number of speakers is priorly known. Nevertheless, the degradation is not significant and it is a small cost given the fact that the system is capable to face problems with unknown number of speakers. Note also that the accuracy of the speaker verification task when considering the speaker factor system, independently of whether S is priorly known or not, is higher that obtained considering the BIC AHC baseline diarization system, that assumes that S is known. This is also reflected in Table 6.15.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.40% (0.00%)	0.2042 (0.00%)
Baseline	4.76% (8.18%)	0.2295 (12.39%)
Speaker Factor, $S = 2$	**4.49% (2.05%)**	**0.2074 (1.57%)**
Speaker Factor, S unknown	4.72% (7.11%)	0.2125 (4.08%)

Table 6.15: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system assuming that S is known ($S = 2$) and unknown, in the *mono-stereo* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of* $min(C_{norm})$

Figure 6.4: *Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, in the mono-stereo scenario.*

It is interesting to notice that the degradation introduced in speaker verification, by assuming that S is unknown for the speaker factor diarization system, is higher in the low false rejection region (bottom-right area in Figure 6.3) than in the low false alarm region (top-left area in Figure 6.3). It can also be observed in Table 6.15, where the degradation in terms of EER is relatively higher than in terms of minimum C_{norm}, when the system is compared to the same speaker diarization system which assumes that $S = 2$.

Figure 6.4 and Tables 6.16 and 6.17 show the percentage of recordings with highest DER of the enrollment dataset that can be accounted to keep the degradation in the EER (6.4(a)) and $min(C_{norm})$ (6.4(b)) with respect to the ideal diarization system over certain value. In these graphs it is clear the effect of higher degradation in the low false rejection region, than in the low false alarm region. The subset that obtains certain degradation in terms of EER is larger when the number of speakers is unknown even when compared to the subset considering the baseline BIC AHC diarization system. However, this is not true if the degradation is measured in terms of $min(C_{norm})$.

Surprisingly, we could not find a subset obtaining a degradation in terms of $min(C_{norm})$ greater than 50%, when considering that S is unknown. The explanation we find to this effect is that the recordings obtaining higher DER values are those that overestimate the number of speakers. However, overestimating the number of speakers is not necessarily harmful for speaker verification in this scenario, as far as there is a pure cluster containing

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	33.61%	7.20%
Speaker Factor, $S = 2$	**4.46%**	**1.37%**
Speaker Factor, S unknown	39.37%	15.09%

Table 6.16: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	50.75%	13.03%
Speaker Factor, $S = 2$	**5.49%**	1.37%
Speaker Factor, S unknown	18.52%	**0.00%**

Table 6.17: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

enough data to enroll the desired speaker. Therefore, there might be such a subset that obtains a degradation in the $min(C_{norm})$ greater than 50%, and we expect this subset to be larger than the one obtained for the speaker factor system when the number of speakers is known. However, this subset will not be composed of those recordings obtaining the highest DER.

6.5.6 *Stereo-Mono* Scenario

Diarization System	DER	$N_{DER<5\%}$ (%)	$N_{DER<10\%}$ (%)
BIC AHC baseline	5.20%	1620 (73.54%)	1873 (85.02%)
Speaker Factor, $S = 2$	**1.32%**	**2098 (95.23%)**	**2149 (97.55%)**
Speaker Factor, S unknown	3.23%	1908 (86.61%)	2008 (91.15%)

Table 6.18: Comparison of the accuracy obtained with the BIC AHC baseline system, the speaker factor diarization system assuming that the number of speakers S is known ($S = 2$) and the speaker factor diarization system with speaker clustering assuming that S is unknown. The systems are evaluated on the testing subset of the *summed* condition of the NIST SRE 2008. The accuracy is measured in terms of DER and percentage of recordings with $DER < 5\%$ and $DER < 10\%$.

Table 6.18 compares the accuracy obtained by the BIC AHC baseline and the speaker factor diarization systems assuming that the number of speakers S is known $S = 2$, to the accuracy obtained by the proposed speaker factor diarization system with speaker clustering assuming that S is unknown. The accuracy is evaluated on the subset of the *summed* condition of the NIST SRE 2008 considered for testing in the *stereo-mono* scenario. As observed before, the speaker factor diarization system obtains higher accuracy when S is priorly known. Under the assumption of unknown S, the accuracy of the speaker factor diarization system is higher than that obtained by the BIC AHC baseline when the last system assumes that $S = 2$.

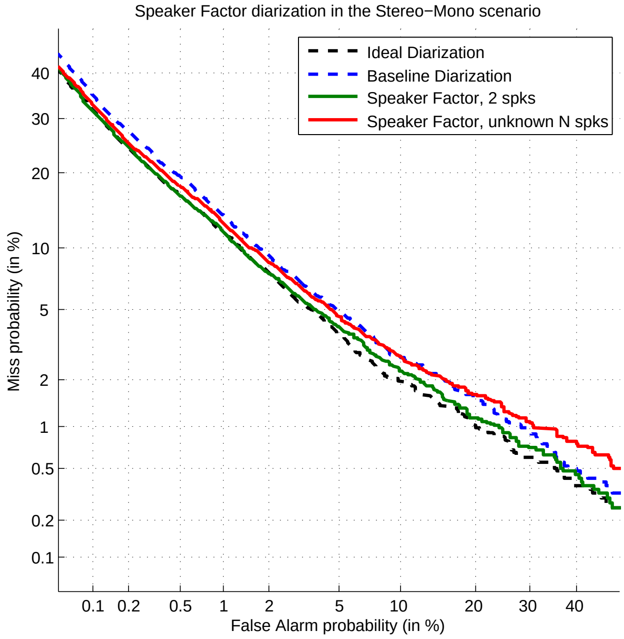


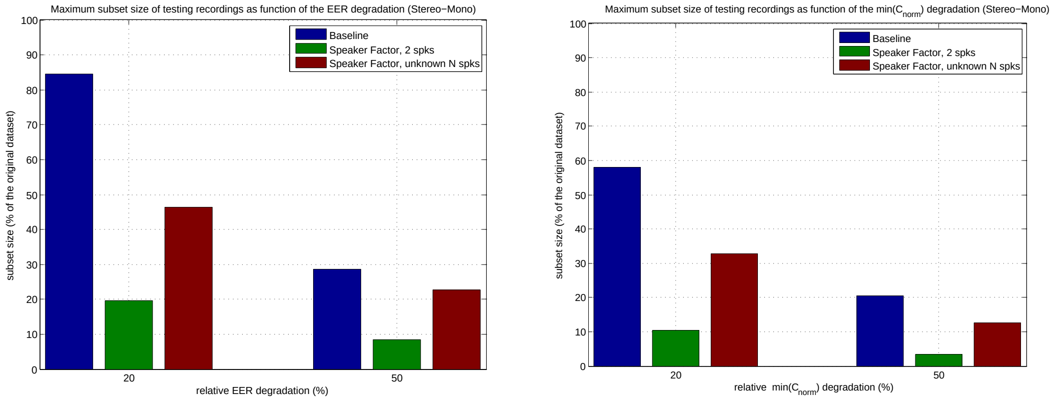
Figure 6.5: *DET curves considering the speaker factor diarization system assuming that S is known ($S = 2$) and unknown, in the stereo-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.23% (0.00%)	0.2102 (0.00%)
Baseline	4.94% (16.78%)	0.2334 (11.04%)
Speaker Factor, $S = 2$	**4.39% (3.78%)**	**0.2097 (-0.24%)**
Speaker Factor, S unknown	4.73% (11.81%)	0.2225 (5.86%)

Table 6.19: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system assuming that S is known ($S = 2$) and unknown, in the *stereo-mono* scenario. The degradation with respect to the ideal diarization system is shown.

Figure 6.5 shows the DET curves obtained in the *stereo-mono* scenario considering the speaker factor system to diarize the testing dataset, assuming that S is known and unknown. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. As in the previous scenario, it can be seen that the degradation introduced by the speaker diarization system with the assumption of not knowing the number of speakers in terms of DER is also reflected in the DET curves. Again, the speaker factor system obtains higher accuracy in the speaker verification task if the number of speakers is priorly known, but even if S is unknown, the accuracy is higher than that obtained for the BIC AHC baseline system.

In this scenario, the effect previously observed is even more significant: when we consider that the number of speakers is unknown for the speaker factor system, the degradation is more severe in the low false rejection region than in the low false false alarm region, as can be observed in Figure 6.5. Again the relative degradation with respect to the speaker factor system considering $S = 2$ in terms of EER is higher than in terms of $min(C_{norm})$, as shown in table 6.19.



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of $min(C_{norm})$*

Figure 6.6: *Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the stereo-mono scenario.*

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	84.43%	28.60%
Speaker Factor, $S = 2$	**19.52%**	**8.40%**
Speaker Factor, S unknown	46.30%	22.70%

Table 6.20: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

Figure 6.6 and Tables 6.20 and 6.21 show the percentage of recordings with highest DER of the testing dataset that can be accounted to keep the degradation in the EER (6.6(a)) and $min(C_{norm})$ (6.6(b)) with respect to the ideal diarization over certain value. As expected, not having prior information on the number of speakers degrades the accuracy obtained in the speaker verification task. In this case, overestimating the number of speakers is more harmful than in the previous scenario, since in the *mono-stereo* scenario, only the hypothetical speaker that best matches the desired speaker is considered. On the other hand, in this scenario, all hypothetical speakers are considered, and since the hypothetical speaker selected is the one that obtains the maximum score, the higher the number of speakers, the higher the obtained scores for impostor trials.

6.5.7 *Mono-Mono* Scenario

Finally, the *mono-mono* scenario is analyzed. This scenario makes use of the *summed-short2* enrollment and *summed* testing subsets. The accuracy of the speaker factor diarization system forcing $S = 2$, and with speaker clustering assuming that S is unknown has been previously evaluated on these datasets, and the results are shown in Tables 6.14 and 6.18. As it has been observed in previous Chapters, this scenario usually reflects with more clarity the differences in accuracy of the speaker diarization systems in the speaker verification task, since diarization errors are introduced in both enrollment and testing sides.

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	58.10%	20.43%
Speaker Factor, $S = 2$	**10.44%**	**3.40%**
Speaker Factor, S unknown	32.68%	12.71%

Table 6.21: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

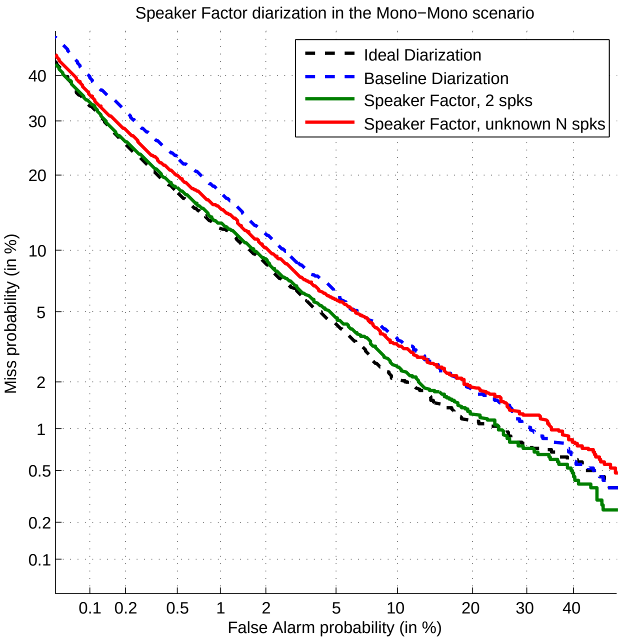


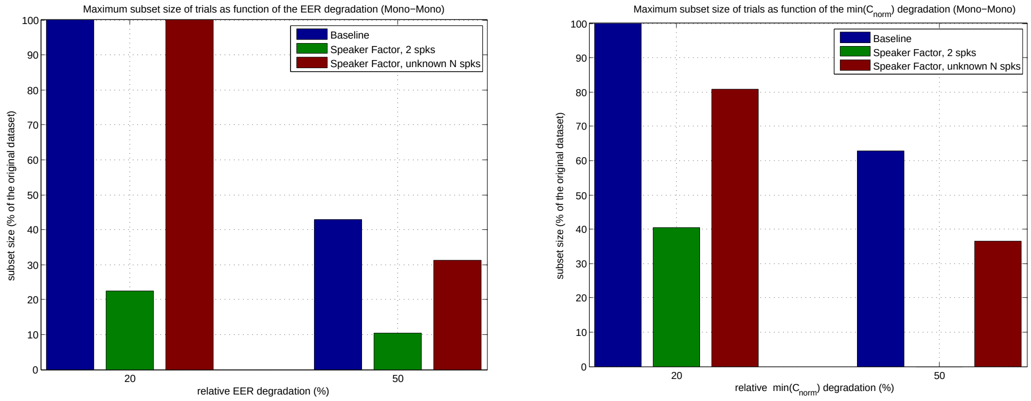
Figure 6.7: *DET curves considering the speaker factor diarization system assuming that S is known ($S = 2$) and unknown, in the mono-mono scenario. The DET curves obtained considering the baseline and ideal diarization systems are shown for comparison.*

Figure 6.7 shows the DET curves obtained in the *mono-mono* scenario considering the speaker factor system with and without intra-session variability compensation to diarize both the enrollment and testing datasets. For comparison, the DET curves considering the baseline BIC AHC and an ideal diarization systems are also shown. Table 6.22 shows the accuracy of the speaker verification system in terms of EER and $min(C_{norm})$ for the four diarization systems, in the *mono-mono* scenario. In this case, the DET curves are more separated than in the previous scenarios, but similar conclusions can be drawn. The degradation in terms of DER introduced when the number of speakers in priorly unknown is also reflected in the DET curves. Again, the degradation is significantly higher in the low false rejection region.

In the previous scenarios, the percentage of recordings with highest DER of a dataset that can be accounted to keep the degradation in the EER and $min(C_{norm})$ with respect to the ideal diarization over certain value has been analyzed. In this case, since both enrollment and testing recordings are processed by the diarization systems, we study the percentage of trials in the speaker verification task that involve recordings with high DER in both enrollment and testing sides, as in previous Chapters.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal	4.54% (0.00%)	0.2157 (0.00%)
Baseline	5.53% (21.81%)	0.2695 (24.94%)
Speaker Factor, $S = 2$	**4.80% (5.73%)**	**0.2233 (3.52%)**
Speaker Factor, S unknown	5.45% (20.11%)	0.2415 (11.99%)

Table 6.22: EER and minimum C_{norm} considering the ideal, the baseline and the speaker factor diarization system assuming that S is known ($S = 2$) and unknown, in the *mono-mono* scenario. The degradation with respect to the ideal diarization system is shown.



(a) *Maximum number of trials depending on the degradation in terms of EER*

(b) *Maximum number of trials depending on the degradation in terms of* $min(C_{norm})$

Figure 6.8: *Percentage of trials in the mono-mono scenario that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, considering the enrollment and testing recordings with highest DER, in the mono-mono scenario.*

Figure 6.8 and Tables 6.23 and 6.24 show the percentage of trials that can be accounted to keep the degradation in the EER (6.8(a)) and $min(C_{norm})$ (6.8(b)) with respect to the ideal diarization over certain value, considering those recordings with highest DER for enrollment and testing. In this scenario, it can be observed that not knowing the number of speakers increases significantly the percentage of trials that can be accounted in order to obtain a considerable degradation in terms of EER and $min(C_{norm})$. However, the results are still better than those obtained for the BIC AHC baseline diarization system assuming that $S = 2$.

Diarization system	$\%_{Degradation(EER)} \geq 20\%$	$\%_{Degradation(EER)} \geq 50\%$
Baseline	100.00%	42.78%
Speaker Factor, $S = 2$	**22.39%**	**10.50%**
Speaker Factor, S unknown	100.00%	31.19%

Table 6.23: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Diarization system	$\%_{Degradation(min(C_{norm}))} \geq 20\%$	$\%_{Degradation(min(C_{norm}))} \geq 50\%$
Baseline	100.00%	62.80%
Speaker Factor, $S = 2$	**40.53%**	**0.00%**
Speaker Factor, S unknown	80.86%	36.44%

Table 6.24: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Part IV

Quality Assessment for Speaker Diarization: Approaches and Applications

7

Quality Assessment for Speaker Diarization

It is known that Speaker Characterization techniques need a considerable amount of data to operate correctly. Considering a speaker verification system as example, and going back to Chapter 3, Figure 3.1, it can be seen that a database is needed during the development stage in order to train the background models needed for speaker recognition. In addition, during the enrollment stage, a dataset containing the target speakers to enroll is needed to train speaker models.

Most of Speaker Characterization techniques work reasonably well as far as the operating conditions during the evaluation stage are close to those present in the datasets considered for development and enrollment. Thus, it is usual to select or even collect the development and enrollment datasets to match the expected conditions.

It is also usual to find situations where, even when it is possible to collect data in the desired conditions, more than one speaker is present in the available recording, especially in the environment of telephone conversations. In these cases it is mandatory to use a speaker diarization system to separate the speakers present in the recording. Obviously, the most accurate the speaker diarization system the more useful the data will be, and lower degradation will be obtained during the operation of the Speaker Characterization system.

On the other hand, even for very accurate speaker diarization systems, it is easy to find recordings that obtain high diarization errors in large datasets. Considering again the example of a speaker verification system, taking into account these recordings either for development or speaker enrollment may degrade the performance of the system. Therefore, it would be interesting to assess the quality of the speaker diarization hypotheses for every recording in order to detect those recordings that obtain high diarization error and may degrade the performance of the speaker characterization system. Note that if it is possible to segregate these recordings from the available dataset accurately, they could be processed manually and utilized for speaker characterization introducing no degradation due to diarization errors. Moreover, depending on the size of the dataset and the number of recordings available per speaker, it may be case where not all recordings are needed, and those recordings obtaining high diarization errors could be discarded.

Thus, given a dataset intended to be used for a Speaker Characterization application, our goal is to extract a subset as representative (composed of as many recordings and speakers of the dataset as possible) and reliable (with low diarization error) as possible. For this purpose, two complementary objectives can be pursued. Firstly the speaker diarization accuracy can be increased, in order to increase the number of recordings with low diarization errors in the dataset. Secondly, an automatic method to detect those recordings with low diarization error can be developed, in order to select the useful recordings that will not degrade the accuracy of our Speaker Characterization application. This work focuses on two-speaker telephone conversations, which is a typical environment in speaker recognition applications.

In previous Chapters, we have focused on developing a speaker diarization system for speaker characterization applications as accurate as possible, to mitigate the degradation introduced because of diarization errors. As a result, a speaker factor based diarization system with intra-session variability compensation has been proposed and it has shown to obtain high overall accuracy on two-speaker telephone conversations. Moreover, the improvement obtained in the diarization accuracy is also reflected in a speaker verification task that makes use of the diarization hypotheses for speaker enrollment and testing in order to segregate speakers from mono recordings of two-speaker telephone conversations.

In this Chapter, we focus on the detection of those recordings that obtain low diarization error, in order to retrieve a useful subset of the available dataset. For this purpose, we first introduce the concept of usefulness of a dataset, a measure that takes into account the representativeness and reliability of the available dataset in order to determine whether the dataset is suitable for a speaker characterization application. Then, a set of confidence measures is analyzed in order to assess the quality of the speaker diarization hypotheses for every recording, and a strategy to fuse the confidence measures aiming at the detection of correctly diarized recordings (with low diarization error) is proposed. Finally the confidence measures and the proposed strategy are validated considering only the subset composed of the recordings detected as correctly diarized, and discarding the remaining recordings. The quality of the detected subset will be analyzed in terms of diarization accuracy but also in a speaker verification task where the subset is needed for enrollment, testing or both stages.

Although in this study we discard the recordings that are automatically classified as not correctly diarized, as mentioned before, these recordings could be processed manually in an semi-supervised process. Since the cost of manual diarization is usually high compared to the cost of an automatic system, the objective of obtaining a subset as representative of the available dataset as possible would not vary.

7.1 Usefulness of a Dataset for Speaker Characterization

Given a dataset Ω composed of two-speaker telephone conversations, our objective is to extract a subset $\Omega' \subset \Omega$ as reliable and representative as possible for a speaker characterization application. The concept of reliability and representativeness for the defined task are explained next. In addition, the concept of Dataset Usefulness, a figure of merit that involves both reliability and representativeness is introduced.

7.1.1 Reliability

The reliability of a recording is related to the accuracy of its diarization hypothesis, so to measure the reliability, we previously need a measure of the accuracy of a diarization hypothesis, that is a measure of the accuracy of the speaker diarization system. As accuracy measure for speaker diarization, we use the DER, as in previous Chapters.

Nevertheless, the reliability of a recording does not only depend on the accuracy of its diarization hypothesis, but also on the application that will use the recording. Depending on the application, it may be useful to consider diarization hypotheses with little error. In general, we consider that an application dependent threshold th is defined, so that every diarization hypothesis obtaining a DER below the threshold will probably be reliable and useful for the application. These threshold th is actually a "soft threshold" since a recording obtaining a DER over or below the threshold is not guaranteed to be useless or useful for the application. Nevertheless, independently of the threshold, the lower the DER the better the diarization hypothesis for our application. For example, assuming that $th = 10\%$, a diarization hypothesis with a DER of 0% will be more reliable than one with a DER of 9%, even though both will probably be useful for our application. Similarly, a diarization hypothesis with a DER of 11% is more reliable than one with a DER of 40%, even though both are probably useless for the given application.

Therefore, we can measure the reliability of a recording n for a given diarization hypothesis as:

$$Rl(n) = \frac{th - DER(n)}{th}, \tag{7.1}$$

where $DER(n)$ is the DER obtained for the recording n. The term in the numerator is the distance between the DER obtained for n and the application dependent threshold, and will be higher as better is the diarization hypothesis. Note that $Rl(n) > 0$ if $DER(n) < th$, so correct diarization hypotheses obtain a positive reliability while incorrect diarization hypotheses obtain negative reliability. The denominator is just a normalization term, so that the maximum value of the reliability is 1. A recording n will be completely reliable $Rl(n) = 1$ only if $DER(n) = 0\%$.

This way, the reliability of a subset Ω' is defined as the mean of the reliabilities of the recordings in Ω':

$$Rl(\Omega') = \frac{\sum\limits_{n \in \Omega'} Rl(n)}{N_{\Omega'}}, \tag{7.2}$$

so the reliability is higher as the recordings in Ω' are better diarized. Note that $Rl(\Omega')$ can be negative if many recordings in Ω' are incorrectly diarized. Actually, the negative values in the reliability mean that an unreliable recording can be "destructive" in the sense that the accuracy of the application that makes use of the subset can be severely degraded when considering the mentioned recording.

7.1.2 Representativeness

Nevertheless, increasing the reliability of a subset Ω' does not guarantee an increase in the accuracy of the application. Note that the reliability of a subset Ω' containing only a single recording which is perfectly diarized ($DER = 0\%$) will be 100%, but, for example, this is not the best subset to train a speaker recognition model based on eigenvoices. We also need Ω' to be as representative of Ω as possible. In this work, representativeness has to do with the fraction of the dataset that is extracted, being Ω' fully representative of Ω only if

$\Omega' = \Omega$. So we define the representativeness of $\Omega' \subseteq \Omega$ as:

$$Rp(\Omega') = \frac{N_{\Omega'}}{N_{\Omega}}, \tag{7.3}$$

where $N_{\Omega'}$ is the number of the recordings in the subset Ω' and N_{Ω} is the number of the recordings in the dataset Ω. Note that this definition considers that all recordings are equally important for our application, but the definition of representativeness can be modified to weight every recording, or to account the rate of the speakers considered rather than the rate of recordings.

7.1.3 Dataset Usefulness

As mentioned before, we need Ω' to be as reliable and representative of Ω as possible, so we define a figure of merit, the Dataset Usefulness for a given subset Ω', $DU(\Omega')$, that involves both the representativeness and reliability of Ω':

$$DU(\Omega') = Rp(\Omega') \times Rl(\Omega') = \frac{\sum\limits_{n \in \Omega'} Rl(n)}{N_{\Omega}} \tag{7.4}$$

The figure of merit $DU(\Omega')$ increases as more representative and reliable, that is, as more useful is the subset Ω' for the application given the diarization hypotheses obtained. In this work we study techniques to obtain a subset Ω' that maximizes $DU(\Omega')$.

There are two ways to increase the $DU(\Omega')$. Firstly, $DU(\Omega')$ can be increased obtaining diarization hypotheses with lower DER, that is, increasing the performance of the speaker diarization system. This way we increase the reliability of the whole dataset Ω and also the number of recordings that are correctly diarized in Ω. Note that $DU(\Omega')$ always increases when adding correctly diarized recordings ($Rl(n) > 0$) to Ω', so the more correctly diarized recordings the better $DU(\Omega')$. In the previous Chapters several techniques have been studied in order to increase the accuracy of the diarization system, thus reducing the overall DER on the dataset Ω. Thus, this techniques will increase the $DU(\Omega')$ when considering the whole dataset ($\Omega' = \Omega$).

Secondly, $DU(\Omega')$ can be increased selecting Ω' so that $\forall n \in \Omega', DER(n) < th$, which implies detecting those recordings correctly diarized, and discarding those incorrectly diarized. Since the reliability is negative for unreliable recordings, such recordings decrease $DU(\Omega')$, and removing them from Ω' will increase $DU(\Omega')$. To detect those recordings correctly diarized we study several confidence measures that assess the quality of the diarization output for every recording.

In this study, results are presented in terms of the overall DER, the percentage of recordings that are correctly diarized in the dataset Ω, and the usefulness of the subset Ω' measured using the figure of merit defined. We consider as speaker characterization application the same speaker verification task studied in previous Chapters. In Section 3.4.2, it has been shown that the degradation of the speaker verification task, for all scenarios considered, is severe whenever $DER > 10\%$. Therefore, $th = 10\%$ is considered for the whole study, but depending on the application, any other value could be considered.

7.2 Confidence Measures for Speaker Diarization

In this section we analyze a set of nine confidence measures to assess the quality of speaker diarization hypotheses obtained for recordings containing two speakers. These confidence measures can be classified into two groups. The first group includes speaker separability indicators: given an input recording and its hypothetical two-speaker diarization hypothesis, two speaker models can be built. Any measure indicating how close or separated are both models is a speaker separability indicator and can be considered as confidence measure to validate the given diarization hypothesis. The second group includes well conditioned data indicators: given an input recording and its hypothetical two-speaker diarization, it is possible to check whether the assumptions and prior models considered by the diarization system are suitable for the available recording. Any measure that determines whether the data is well conditioned for the correct operation of the diarization system is a indicator of well conditioned data. The following sections describe the proposed confidence measures.

7.2.1 Speaker Separability Indicators

The set of measures classified as speaker separability indicators aims at, given two speaker models built from a diarization hypothesis, determining how close or separated are both speakers. Assuming that an incorrect diarization will merge both speakers in every hypothetical speaker model, we can expect these speaker models to be close to each other if the diarization is wrong. In general, the opposite is not true: assuming that the diarization hypothesis is correct, we can not expect the two speaker models to be far from each other in all cases. A pair of speaker models may be closer to each other than other pair depending on the characteristics of the speakers. However, we can expect recordings containing two similar speakers to be harder to diarize, so a measure of how close both speakers are is more related to the difficulty of the task than to the quality of the diarization hypothesis. In such a situation, the measures described below can still be considered as confidence measures, providing a measure of the reliability of the hypothetical speaker diarization.

The proposed speaker separability indicators are:

- **Bayesian Information Criterion (BIC):** BIC has been considered in this work for speaker segmentation and clustering in the BIC AHC baseline diarization system (see Section 3.4.1), and as clustering metric for an AHC strategy. In Section 6.5, BIC has shown to be the most accurate clustering metric and also to provide a robust stopping criterion to determine the actual number of speakers when it is priorly unknown. In this case, the number of speakers is priorly known, so we do not need a stopping criterion. However, ΔBIC can be used as a measure of the accuracy of a given diarization hypothesis, since it is a distance between both clusters obtained.

  In this approach, given two sequences of acoustic feature vectors obtained by the diarization system, we compute the BIC for two hypotheses: Each sequence belongs to a different speaker or both sequences belong to the same speaker, as it is usual in AHC for diarization (see Section 6.1.2). The confidence measure is the ΔBIC, the difference between both BIC values. To avoid adjusting BIC penalty parameters, we force the models for both hypotheses to have the same complexity. That is, we model every speaker in the first hypothesis with a GMM of N Gaussians, and the global model in the second hypothesis with a GMM of $2N$ Gaussians. In our experiments we set N to 32 Gaussians. As feature vectors we use 12 MFCC including C0.

$$C_{BIC} = \Delta\ BIC = log(\frac{\mathcal{L}(\chi_1|\Theta_1)\mathcal{L}(\chi_2|\Theta_2)}{\mathcal{L}(\chi_{1,2}|\Theta_{1,2})}, \tag{7.5}$$

where $\mathcal{L}$ denotes log-likelihood, χ_s the acoustic feature vectors obtained for speaker s, and Θ_s the GMM model obtained from every vector sequence χ_s. We can expect $\Delta\ BIC$ to be high for correct diarization hypotheses and low for difficult recordings and thus possibly wrong diarization hypotheses. This measure was presented in [Vaquero *et al.*, 2010a] as confidence measure for speaker segmentation showing good performance.

- **Kullback-Leibler (KL) divergence:** another way to measure the accuracy of a given diarization is to compute the symmetrized KL divergence or KL2 (see Section 2.3.1) between the Gaussian speaker models obtained in the speaker factor space. In this approach we use the hypothetical diarization labels to obtain two sequences of speaker factors Y_1 and Y_2, and Gaussian models are trained for each sequence. We can expect higher KL divergences between both Gaussian models when the diarization is correct (i.e. the models are pure and separated). This measure was presented in [Vaquero *et al.*, 2010a] as confidence measure for speaker diarization showing good performance.

$$C_{KL} = KL(\mathcal{N}(\mu_{Y_1}, \Sigma_{Y_1}) \parallel \mathcal{N}(\mu_{Y_2}, \Sigma_{Y_2})) + KL(\mathcal{N}(\mu_{Y_2}, \Sigma_{Y_2}) \parallel \mathcal{N}(\mu_{Y_1}, \Sigma_{Y_1})), \tag{7.6}$$

Note that other distance metrics traditionally considered for segmentation or clustering, including those presented in Section 2.3.1, could be used as confidence measure for speaker diarization on two-speaker conversations. Since BIC and KL2 are among the most popular distance metrics for speaker diarization we will not consider other metrics.

- **Probabilistic LDA (PLDA) speaker verification score:** Another way to obtain a confidence measure for a diarization hypothesis is to see the problem as a speaker verification task: given two segments, obtained from the diarization output, determine whether or not they were uttered by the same speaker. Usually, speaker verification systems provide a score, which is a measure of how close both segments are and it is related to the likelihood for the two segments to belong to the same speaker. Such a score can be used directly as confidence measure.

  Therefore, if we consider the PLDA speaker verification system described in 3.2, the LLR it provides can be considered as confidence measure, as it has been previously considered as distance metric for clustering in Section 6.5.1. The confidence measure is obtained comparing both hypothetical speakers given by the diarization output using the PLDA model. This confidence measure will be high if both speakers extracted from the diarization hypothesis are similar, for example, if they are built from mixed samples extracted from the actual speakers, and will be low otherwise.

- **Cosine distance, intra-session compensation:** Since any clustering metric and thus any score provided by a speaker verification system could be utilized as confidence measure for this task, we can consider any of the clustering metrics proposed in Section 6.5.1 in addition to the ΔBIC and the PLDA speaker verification score. The remaining clustering strategies consider speaker factor vectors instead of i-vectors to model every speaker, and the distance metrics are obtained using a PLDA model, computing the cosine distance or computing the euclidean distance between both speaker factor vectors. The PLDA model is trained to compensate for intra-session variability compensation while to compute the cosine and euclidean distances, the speaker factor vectors are transformed using a LDA 100 → 50 + WCCN strategy to compensate for intra-session variability as in Chapter 5 (see Section 6.5.1).

Since the three approaches make use of very similar speaker modeling and essentially provide the same information, only one of them is considered. In this work we consider the cosine distance computed over the speaker factor vectors with intra-session variability compensation since it provides acceptable clustering accuracy (see Section 6.5.3) and the confidence measure is obtained over speaker models that are identical to those considered in the speaker diarization system. Thus, the score is directly related to the difficulty of the recording that the speaker diarization system has processed. The euclidean distance considers the same speaker models but its clustering accuracy is lower. Finally, the PLDA with intra-session variability compensation provides higher accuracy than the cosine distance, but the intra-session variability compensation strategy is different from that considered for diarization. However, there is no reason to believe than the cosine distance should be better confidence measure than the others, and the PLDA or the euclidean distance could be also considered.

Therefore, for every recording and its diarization hypothesis, the point estimates for two speaker factor vectors are obtained, m_{y_1} and m_{y_2}, one for each speaker, and then we normalize them applying LDA 100 → 50 + WCCN for intra-session variability compensation as in Chapter 5. To obtain the final score and thus our confidence measure, we compute the cosine distance between both normalized speaker vectors as in [Dehak *et al.*, 2010]:

$$C_{cosine} = cos(angle(m_{y_1}, m_{y_2})) = \frac{m'_{y_1} m_{y_2}}{||m_{y_1}||\,||m_{y_2}||}, \tag{7.7}$$

We can expect C_{cosine} to be low for correct diarization hypotheses and to be high for possibly wrong diarization hypotheses.

7.2.2 Well Conditioned Data Indicators

The indicators of well conditioned data aims at, given a recording and a diarization hypothesis for the recording, determining whether of not the data obtained from the recording matches the prior models that the diarization system uses and fulfills the assumptions made for the correct operation of the system. Again, if the data does not match the prior models or does not fulfill the assumptions made by the diarization system, the diarization hypothesis may be incorrect, but we cannot ensure that. However, if the data is very well matched and all assumptions fulfilled, it is very likely that the diarization hypothesis obtained is correct.

The proposed indicators of well conditioned data are:

- **Likelihood on the UBM (UBML):** Assuming that we have obtained the sequence of acoustic feature vectors χ for a given recording, we can obtain the likelihood for the sequence on the UBM Θ_{UBM} of the diarization system. Actually, the speaker factor diarization system evaluates the whole sequence to compute the speaker factors, so computing the UBM likelihood is costless. This likelihood, normalized by the number of frames of the feature vector sequence, indicates whether the observed sequence is well modeled by the UBM. We can expect a correct diarization hypothesis for a recording whose likelihood on the UBM is high, and a possibly incorrect diarization hypothesis for a recording whose UBM likelihood is low. This metric was proposed as a quality measure for speaker verification in [Harriero *et al.*, 2009].

$$C_{UBML} = log\mathcal{L}(\chi|\Theta_{UBM}), \tag{7.8}$$

- **Core Segmentation convergence:** The proposed approach for two-speaker diarization follows a iterative procedure during the Core Segmentation stage to determine the two Gaussian speaker models that best fit the sequence of speaker factor vectors extracted from the recording. This procedure converges when there is no change in the output labels of the Viterbi segmentation for two consecutive iterations, as explained in Section 4.1.3. We can expect a fast convergence when the system can easily find the correct diarization hypothesis since the speakers are easily separable, and a slow convergence otherwise. This measure was presented in [Vaquero *et al.*, 2010a] as confidence measure for speaker diarization showing acceptable performance.

$$C_{it} = \#iterations, \tag{7.9}$$

- **Skewness of the speaker factors:** According to the factor analysis paradigm, the speaker factor vectors belonging to a single speaker within a session should follow a Gaussian distribution. As far as a given diarization hypothesis involves two clusters of speaker factor vectors that follow Gaussian distributions we can expect the diarization hypothesis to be correct. Otherwise, the diarization hypothesis may be wrong. With this assumption, any Gaussianity measure can be used as confidence measure. For example, the skewness of the Gaussian speaker models should be zero, so we compute the skewness of the speaker factor vectors for every speaker over the principal axis, obtaining a vector of skewness values. The absolute value of these values are then accumulated, obtaining a scalar for every speaker. To obtain a single confidence measure, we compute the geometric mean of both scalar values. This confidence measure will be high when the Gaussian speaker models are highly asymmetric and thus the diarization could be incorrect. The confidence measure value will be close to zero if the speaker models are actually Gaussian and thus we can expect the diarization to be correct.

  The skewness based confidence measure is computed as follows:

$$C_{skew} = \sqrt{\sum_{r=1}^{r=R} abs(skew(Y_1(r))) \times \sum_{r=1}^{r=R} abs(skew(Y_2(r)))} \tag{7.10}$$

$$skew(Y_s(r)) = \frac{\frac{1}{T_s}\sum_{t=1}^{t=T_s}\left(m_{y_s}(r,t) - \mu_{Y_s}(r)\right)^3}{\left(\frac{1}{T_s}\sum_{t=1}^{t=T_s}\left(m_{y_s}(r,t) - \mu_{Y_s}(r)\right)^2\right)^{\frac{3}{2}}}, \tag{7.11}$$

where R is the dimension of the speaker factor vectors considered, $Y_s, s = \{1, 2\}$ are the sequences of speaker factor vectors for every speaker, T_s is the number of speaker factor vectors in the sequence Y_s, $m_{y_s}(r,t)$ is the r component of the t-th speaker factor vector in the sequence Y_s, and $\mu_{Y_s}(r)$ denotes the mean of the r component computed over all speaker factor vectors in the sequence Y_s.

- **Kurtosis of the speaker factors:** Along with the asymmetry of the speaker factor distributions we can compute the kurtosis to test the Gaussianity of the speaker models in the speaker factor space. The kurtosis of these models should be zero, so any deviation from this value is an indication of a possibly incorrect diarization hypothesis. We compute a kurtosis based confidence measure following the same procedure used for C_{skew}:

$$C_{kur} = \sqrt{\sum_{r=1}^{r=R} kur(Y_1(r)) \times \sum_{r=1}^{r=R} kur(Y_2(r))} \tag{7.12}$$

$$kur(Y_s(r)) = \frac{\frac{1}{T_s}\sum_{t=1}^{t=T_s}\left(m_{y_s}(r,t) - \mu_{Y_s}(r)\right)^4}{\left(\frac{1}{T_s}\sum_{t=1}^{t=T_s}\left(m_{y_s}(r,t) - \mu_{Y_s}(r)\right)^2\right)^{2}} \tag{7.13}$$

- **Normalized Eigenvalue Dispersion of the speaker factors**: In Chapters 4 and 5 it has been shown that for the correct operation of the speaker factor based diarization system, it is important to fulfill the assumption that the speaker factors from a single speaker follow a Gaussian distribution with covariance matrix close to the identity. So an indicator of how close our speaker models are to fulfill this assumption is a good candidate for confidence measure. We propose a normalized eigenvalue spread defined as:

$$C_{eig} = log\left(\frac{\frac{max(\lambda_{1,2})}{median(\lambda_{1,2})}}{\frac{max(\lambda_1)}{median(\lambda_1)}\frac{max(\lambda_2)}{median(\lambda_2)}}\right) \tag{7.14}$$

where λ_s are the eigenvalues obtained by Singular Value Decomposition (SVD) of the sequence Y_s. In all eigenvalue spread the median of the eigenvalues have been used in the denominator rather than the minimum, since the minimum may be noisier and less robust across different speakers. The term in the numerator is the eigenvalue spread considering the speaker factors from both speakers (joint hypothesis), and should increase as the speakers are more separable, while the term in the denominator is the product of the eigenvalue spread for every speaker and should be close to one if the mentioned assumption is fulfilled. In Section 4.1.2, Figure 4.3(b) it is shown a representation of the eigenvalues obtained for the two speakers in a single conversation separately and for the whole conversation. It can be seen that the eigenvalue spread for the whole conversation is much higher than eigenvalue spread for every speaker if the assumption is close to be fulfilled and the speakers are easily separable.

7.3 Quality Assessment for Speaker Diarization

Obtaining a reliable confidence measure can be useful to predict the accuracy of the diarization system for a diarization hypothesis, so that a given application can decide how to deal with the current recording. This usually means that the confidence measure is compared to a threshold or set of thresholds in order to classify the recording into different classes that will be processed differently. Then, given a dataset Ω, a partition of Ω is created according to the predicted accuracy of the diarization system so that the application can deal properly with every class in the partition. Therefore, the partition will be application dependent. For example, a semi-supervised diarization system can be built, so that the user only needs to check the diarization hypotheses for a small subset of the whole dataset. This subset will be composed of those recordings that the diarization system labels as *unreliable*.

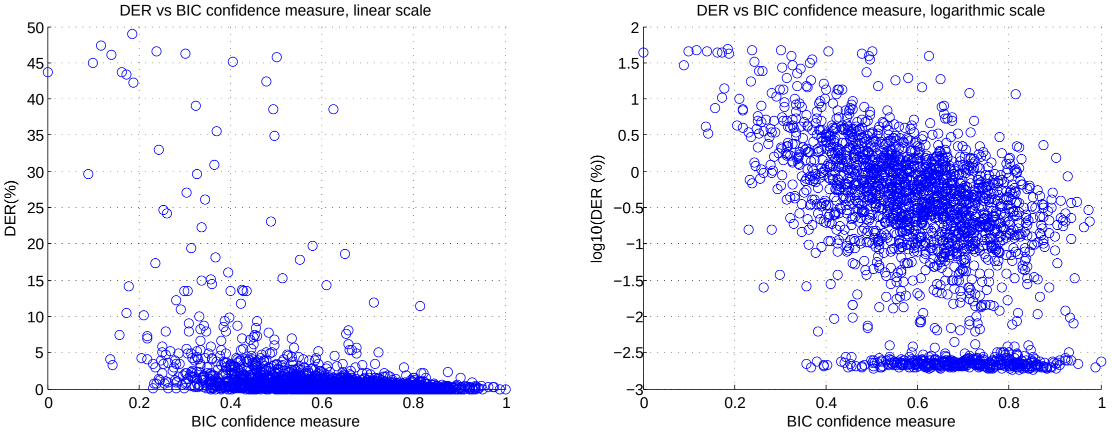
In this section, a model for partitioning a given dataset according to the quality of the diarization hypotheses is studied. As measure to represent the actual quality of a given diarization hypothesis the DER is considered, but other measures could be used as well, depending on the application.

7.3.1 Inferring Diarization Quality from Confidence Measures

In order to deal with every recording on a given dataset properly, we need to infer the accuracy of its corresponding diarization hypothesis. Thus, it would be desirable to find a function that enables us to represent approximately the accuracy of the diarization hypothesis in terms of DER given a confidence measure or a set of confidence measures by means of a regression model.

However, the nature of the confidence measures and the DER makes difficult to find a regression model that enables us to describe the DER properly, for two main reasons. First, most of the proposed confidence measures describe the difficulty of the diarization task rather than the quality of the diarization output as explained previously. Thus, low values of the confidence measures do not correspond to high DER values necessarily. Second, the DER does not completely describe the quality of a diarization hypothesis, since the concept of quality depends on the application. We are trying to model this dependency by fixing a threshold in the DER, but for some applications it may not be enough. For example the DER does not take into account the speaker balance. Therefore, in conversations where the speaker turns and their durations are highly unbalanced, a speaker may be completely missed and the DER will be still low. This situation can be unacceptable for speaker characterization applications where the missed speaker is needed.

The second problem is inherent of the concept of quality and it can be solved by defining a quality measure according to the application requirements. In this study we assume that the DER is a good quality measure to predict the accuracy of the speaker characterization application that will make use of the diarization hypotheses obtained. However, the first



(a) *DER as function of the BIC confidence measure.*

(b) $log_{10}(DER)$ *as function of the BIC confidence measure.*

Figure 7.1: *DER and* $log_{10}(DER)$ *as function of the BIC confidence measure.*

problem makes impossible to find a bijective function to infer the DER given a confidence measure, and thus it is not possible to obtain a continuous bijective function f so that $DER = f(C)$ where C is a confidence or set of confidence measures.

To show this effect, the *heavy-weight* speaker factor diarization system with intra-session variability compensation is considered to process the *summed* dataset from the NIST SRE 2008 (see Section 3.2.1). Figure 7.1 represent the DER and $log_{10}(DER)$ against the C_{BIC} confidence measure. The BIC confidence measure has been normalized to be a value between 0 and 1. It can be seen that as expected, the higher the value of the confidence measure the lower the DER, but the relation is not linear. The relation between C_{BIC} and $log_{10}(DER)$ seems linear, but the cloud of samples in Figure 7.1(b) spreads significantly to approximate it accurately by a continuous bijective function. However, it also can be seem specially in 7.1(a) that it is possible to find regions in the range of the BIC confidence measure values where the DER is guaranteed to be below or over certain threshold. For example, The DER is always below 5% for values of C_{BIC} over 0.85.

At sight of the previous figures it does not seem possible to find a continuous bijective function to build a regression model in order to infer the DER accurately given a confidence measure. However, we do not need to infer the DER accurately. Our objective, as explained before, is to classify every recording depending on the diarization accuracy, so the speaker characterization can deal with the recording properly. Thus, we want to build a multinomial logistic regression, capable of classifying the recordings in the desired quality classes.

In this study we limit our problem to two classes, assuming that our application will use only the subset Ω_c of correctly diarized recordings and discard the rest of the dataset (Ω_i). Thus, we want to solve the detection problem where only those diarization hypotheses obtaining a accuracy over certain threshold are desired. As mentioned before, in this study we consider that the correctly diarized or *reliable* recordings (retrieving the concept introduced in Section 7.2) are those obtaining $DER < 10\%$, so a threshold in the DER value $th_{DER} = 10\%$ is set.

7.3.2 Selection of Confidence Measures

In order to build a logistic regression model to detect *reliable* recordings, the first step is to validate the proposed measures, and to select the most informative ones. For this purpose, we consider the NIST SRE 2008 *summed* dataset (see Section 3.2.1). As diarization system, the *heavy-weight* speaker factor system (see Chapter 4) with intra-session variability compensation (see Chapter 5) is considered. With this framework, the relationship between every confidence measure and the DER is analyzed.

In order to use a confidence measure to classify a diarization hypothesis depending on its DER there must exist correlation between the confidence measure and the DER. Thus, to validate the confidence measures, the correlation between every confidence measure and the DER for the complete dataset considered is computed as follows:

$$corr(C_i, DER) = \frac{\sum_n (C_i(n) - \mu_{C_i})(DER(n) - \mu_{DER})}{\sqrt{\sum_n (C_i(n) - \mu_{C_i})^2 \sum_n (DER(n) - \mu_{DER})^2}} \tag{7.15}$$

$$\mu_{C_i} = \frac{\sum_n (C_i(n))}{N} \tag{7.16}$$

$$\mu_{DER} = \frac{\sum_n (DER(n))}{N} \tag{7.17}$$

Where $C_i(n)$ and $DER(n)$ are the confidence measure and DER obtained for the recording n, and $N = 2213$ is the number of recordings in the dataset. The correlation values are in the range [-1,1]. A correlation of 0 will indicate that the confidence measure and the DER are not correlated, and a correlation of 1 or -1 will indicate that there exist a linear relation between them.

Confidence measure	$corr(C_i, DER)$
C_{BIC}	-0.3696
C_{KL}	-0.2205
C_{PLDA}	0.3466
C_{cosine}	0.3158
C_{UBML}	-0.0922
C_{it}	0.1912
C_{skew}	0.0128
C_{kur}	0.0229
C_{eig}	-0.3836

Table 7.1: *Correlation between the confidence measures and DER.*

Table 7.1 shows the correlation values of the proposed confidence measures with the DER. The absolute value of the correlation presented in the table shows how correlated is every confidence measure to the DER. We see four confidence measures that are acceptably correlated with DER: BIC, PLDA, cosine distance and normalized eigenvalue spread. KL and number of iterations are still interesting to infer DER values, but UBM likelihood seems to be little correlated with the DER. Finally, the Gaussianity measures, skewness and kurtosis, seems to be non-informative to infer the DER for the available diarization hypotheses.

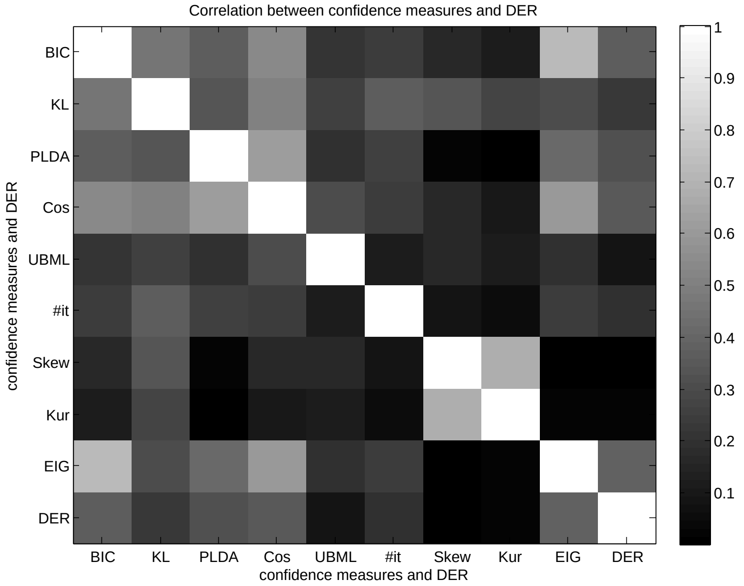


Figure 7.2: *Absolute value of the correlation between the proposed confidence measures and the DER for the NIST SRE 2008 summed dataset*

The sign of the correlation indicates the behavior of the confidence measures: for positive correlations, the confidence measure increases as the DER increases, and for negative correlations, the confidence measures decreases as the DER increases. If we want to name a metric confidence measure, such metric should increase as the performance increases, so we would expect negative correlations. Thus, we should change the sign of the metrics obtaining positive correlations with DER.

It is also interesting to determine the correlation between the different confidence measures, to see whether some of them are highly correlated and do not provide additional information.

Figure 7.2 shows the absolute value of the correlation between the proposed confidence measures and DER. We can see again that there are four confidence measures more correlated with the DER (BIC, PLDA, cosine distance and normalized eigenvalue spread), two of them less correlated (KL and number of iterations) and three of them almost uncorrelated (UBM likelihood, skewness and kurtosis). It is also interesting to see that some confidence measures are highly correlated between them: BIC with normalized eigenvalue spread, the last one with PLDA or PLDA with cosine distance, but also Skewness with Kurtosis.

From this correlation values we can discard some confidence measures that we expect not to be helpful when inferring DER values. The confidence measures to discard are Skewness, Kurtosis and UBML. The number of iterations and KL are candidates to be discarded since they do not show high correlation compared to BIC, cosine distance, PLDA and normalized eigenvalue spread.

On the other hand it is also interesting to analyze the behavior of the confidence measures in order to make use of them to infer the DER. In Figure 7.1(a), it can be seen that a high value of C_{BIC} ensure that a diarization hypothesis obtains a low DER, but low values of C_{BIC} do not guarantee that the DER is high. The correlation does not indicate whether a confidence measure can be useful to ensure that a recording is *reliable*, *unreliable* or both.

To analyze how every confidence measure can be useful to infer the DER, we first normalize the confidence measures to provide values between 0 and 1. The sign of those confidence measure that give positive correlations is changed, since we want the confidence measures to decrease as the DER increases. Then, for every confidence measure, the dataset is fragmented into five subsets depending on the values of the confidence measure. The subsets are obtained according to an uniform division of the value range of the confidence measures. Thus, for every confidence measure the first subset comprises all recordings obtaining confidence measure values from 0.0 to 0.2, the second from 0.2 to 0.4 and so on. Finally, for every confidence measure, the average DER, the 90% confidence interval of the DER and the number of recordings is computed for every subset. This analysis is only performed for the six confidence measures that has shown to be correlated with the DER.

Figure 7.3 shows the average DER and the 90% confidence interval of the DER depending on the value range of the confidence measure, for the six confidence measures selected. In addition, the number of recordings for every value range for the confidence measures is represented. It can be seen that all confidence measures follow a similar behavior. Low values of the confidence measures are related to higher DER values in average, but the confidence interval is huge, and the average DER and the confidence interval decrease as the confidence measures increase. Thus, as observed before, high values of the proposed confidence measures ensure that the diarization hypotheses are *reliable*, but there is no confidence measure capable of identifying *unreliable* diarization hypotheses: if the confidence measure is low, the confidence interval for the DER is so high to determine whether or not the diarization hypothesis is *unreliable*. Note that this behavior is expected: as explained in Section 7.2, the proposed confidence measures can also be seen as indicators of the difficulty of the diarization task. Therefore, if the task is easy, the DER is expected to be low, but when the task is difficult, it is hard to predict the DER.

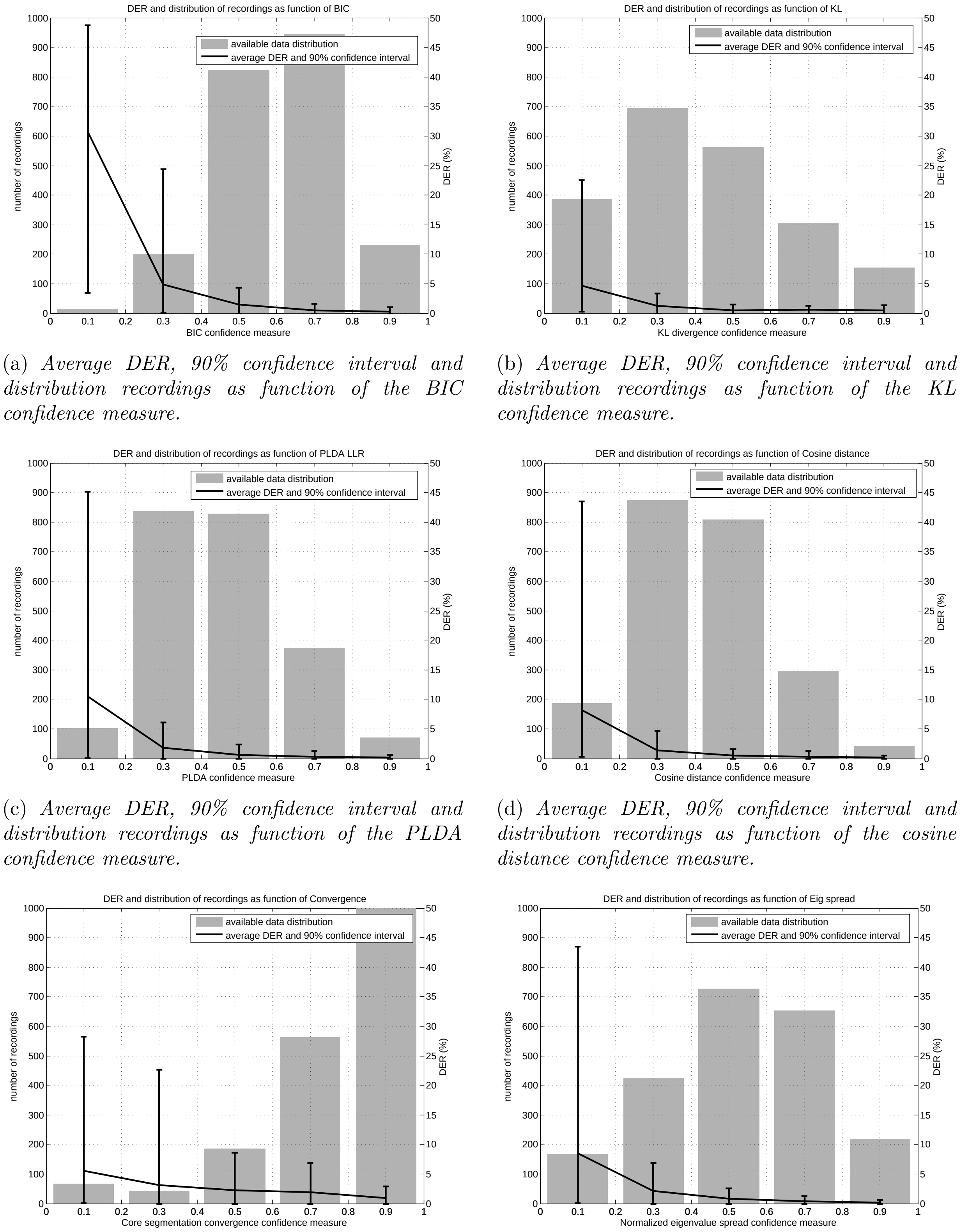
Analyzing every confidence measure, it can be observed that the BIC shows a very interesting behavior. It is the only confidence measure that ensures that for very low confidence values, the DER will hardly ever be very low. For normalized C_{BIC} values below 0.2, the 90% confidence interval does not include recordings obtaining a DER below 3%. On the other hand, the same confidence interval for all the other confidence measures includes recordings with DER as low as 1%.

The behavior for the PLDA, cosine distance and eigenvalue spread confidence measures is very similar, and they all, along with the BIC confidence measure, seem appropriate to detect most of the *reliable* diarization hypotheses in the dataset with high precision.

The KL confidence measure does not show capability of segregating *unreliable* recordings at all. Note that the number of recordings with normalized C_{KL} values below 0.2 is very high compared to other confidence measures, and these recordings include some obtaining a DER as low as 0%. In addition, the average DER and the 90% confidence interval do not decrease as the confidence increases for confidence values above 0.4, so variations of the normalized KL over 0.4 do not provide information.

The convergence confidence measure shows a behavior quite different from all the other confidence measures. The average DER and 90% confidence intervals decrease as the confidence measure increases, but the decrease seems linear with the confidence measures. However, it shows wide confidence intervals for all value ranges of the confidence measure, so the convergence confidence measure has low precision detecting both *reliable* and *unreliable* diarization hypotheses.

Therefore, we only consider four confidence measures in order to detect those recordings



(a) *Average DER, 90% confidence interval and distribution recordings as function of the BIC confidence measure.*

(b) *Average DER, 90% confidence interval and distribution recordings as function of the KL confidence measure.*

(c) *Average DER, 90% confidence interval and distribution recordings as function of the PLDA confidence measure.*

(d) *Average DER, 90% confidence interval and distribution recordings as function of the cosine distance confidence measure.*

(e) *Average DER, 90% confidence interval and distribution recordings as function of the convergence confidence measure.*

(f) *Average DER, 90% confidence interval and distribution recordings as function of the normalized eigenvalue spread confidence measure.*

Figure 7.3: *Average DER, 90% confidence interval and number of recordings obtained for every confidence measure, fragmenting the dynamic range of the normalized confidence measures into five uniform intervals.*

with *reliable* diarization hypotheses in the dataset: BIC, PLDA, cosine distance and normalized eigenvalue spread confidence measures. Once more, note that all the confidence measures considered are precise in the detection of recordings with *reliable* diarization hypotheses, but the precision in the detection of recordings with *unreliable* diarization hypotheses is very poor.

7.3.3 Detection of Recordings with *Reliable* Diarization Hypotheses

In order to detect those recordings correctly diarized, or recordings with *reliable* diarization hypotheses, we train a linear logistic regression model using the selected confidence measures. This model will give as output the certainty of a diarization hypothesis for a given recording to be correct. The main problem of this approach is that linear logistic regression considers that all recordings are equally important for the detection task during the training stage of the algorithm. That is, the logistic regression model does not take into account the concept of *reliability* of every recording. It only considers the fact the DER is over or below th_{DER}. To overcome this problem we can use a weighted linear logistic regression model, which is a logistic regression model where every recording has a weight or an importance in the training algorithm. The weights for the correctly (w_c) and incorrectly (w_i) diarized recordings are related to the reliability of every recording and are defined as follows:

$$w_c(n) = \frac{c_m Rl(n)}{Rp(\Omega_c)c_m Rl(\Omega_c) - Rp(\Omega_i)c_{fa}Rl(\Omega_i)} \tag{7.18}$$

$$w_i(n) = \frac{-c_{fa} Rl(n)}{Rp(\Omega_c)c_m Rl(\Omega_c) - Rp(\Omega_i)c_{fa}Rl(\Omega_i)} \tag{7.19}$$

where c_m and c_{fa} are the cost of a false negative or a miss and the cost of a false positive or a false alarm respectively, Rp() denotes representativeness as defined in eq. (7.3), and Rl() denotes reliability as defined in eq. (7.2). The minus sign in the denominators of both weights and in the numerator of $w_i(n)$ is used since Rl() is negative when it is evaluated on a recording (n) or a subset (Ω_i) that is *unreliable* (see Section 7.1.1). This way, every recording is weighted by the absolute value of the reliability of its diarization hypothesis, which is the distance between the DER obtained for that hypothesis and the application dependent threshold th. Using these weight definitions for training the logistic regression minimizes the following cost function:

$$cost(c_m, c_{fa}) = Rp(\Omega_c)c_m Rl(\Omega_m)p_m - Rp(\Omega_i)c_{fa}Rl(\Omega_{fa})p_{fa}, \tag{7.20}$$

where Ω_m and Ω_{fa} are the subsets composed by the missed and false alarms recordings respectively, and p_m and p_{fa} are the rate of miss and false alarm respectively. Note that this expression is analogous to that used in the NIST SRE [NIST, 2010c] to define the cost, where the prior probability of target and non-target trials are in this case $Rp(\Omega_c)$ and $Rp(\Omega_i)$ respectively, and the cost of every miss and false alarm depends on every diarization hypothesis and it is given by its reliability.

In the definition of our objective figure of merit $DU(\Omega')$ we have considered that $c_m = c_{fa} = 1$. Taking this into account, and realizing that $Rp(\Omega_c)Rl(\Omega_m)p_m = DU(\Omega_m)$ and $Rp(\Omega_i)Rl(\Omega_{fa})p_{fa} = DU(\Omega_{fa})$, since $Rp(\Omega_c)p_m = Rp(\Omega_m)$ and $Rp(\Omega_i)p_{fa} = Rp(\Omega_{fa})$

the cost can be defined as:

$$cost(c_m = 1, c_{fa} = 1) = DU(\Omega_m) - DU(\Omega_{fa}), \tag{7.21}$$

It can be seen that minimizing this cost is equivalent to detecting the subset Ω' that maximizes the Dataset Usefulness $DU(\Omega')$, since $DU(\Omega')$ can be expressed as:

$$\begin{aligned} DU(\Omega') &= DU(\Omega_c) - DU(\Omega_m) + DU(\Omega_{fa}) \qquad (7.22)\\ &= DU(\Omega_c) - cost(c_m = 1, c_{fa} = 1). \qquad (7.23)\end{aligned}$$

Note that eq. (7.21) will also be true when $c_m \neq c_{fa}$, $c_m \neq 1$, $c_{fa} \neq 1$ if we generalize the definition of Dataset Usefulness in eq. (7.4) to include c_m and c_{fa}. Then, we can define the Generalized Dataset Usefulness of a dataset Ω for a given subset Ω', given c_m and c_{fa}, as:

$$GDU(\Omega', c_m, c_{fa}) = Rp(\Omega') \times \left(Rl(\Omega_c) - Rl(\Omega_m) + \frac{c_{fa}}{c_m} Rl(\Omega_{fa}) \right), \tag{7.24}$$

so that $GDU(\Omega', c_m, c_{fa}) = GDU(\Omega_c, c_m, c_{fa}) - cost(c_m, c_{fa})$.

In this work we consider $c_m = c_{fa} = 1$. Thus, according to eq. (7.23), the proposed logistic regression model is trained to detect the subset Ω' that maximizes the $DU(\Omega')$.

7.3.4 Evaluation of Quality Assessment for Diarization

To analyze the accuracy of the weighted linear logistic regression for retrieving the correctly diarized recordings in a dataset, we consider again the NIST SRE 2008 *summed* channel condition as development dataset Ω, and we obtain the diarization hypotheses using the best performing configuration for the speaker factor diarization system: the *heavy-weight* configuration considering $R = 100$ speaker factors and LDA 100 → 50 + WCCN for intra-session variability compensation (see Chapter 5). It is assumed that the retrieval of those recordings obtaining a DER below $th = 10\%$ is desired.

The proposed configuration of the diarization system obtains an overall DER as low as 1.31% and enables us to retrieve up to 2158 correctly diarized recordings out of 2213 (see Chapter 5, Table 5.1). The development dataset Ω and the four selected confidence measures are considered to train the weighted logistic regression model.

Table 7.2 shows the results obtained for the task of detecting *reliable* recordings on the NIST SRE 2008 *summed* dataset. The overall DER, the standard deviation of the DER values, the representativeness and the proposed figure of merit *DU* are displayed. We can see that 96.75% (2141 out of 2213) of the recordings are correctly detected as *reliable* recordings (true positives, $\Omega_{true\,positives}$), and the overall DER and the standard deviation of the DER for these recordings are as low as 0.75% and 1.30% respectively. Only 0.63% (14 out of 2213) of the recordings are false alarms or false positives, which is a low number compared to the number of correctly diarized recordings detected. The most important result is the Dataset Usefulness for the selected subset $\Omega' = \{\Omega_{true\,positives} \cup \Omega_{false\,positives}\}$, $DU(\Omega')$. We can see that $DU(\Omega')$ is over 2.5% higher than $DU(\Omega)$. This means that the subset Ω' of detected recordings is more useful for the application than the whole dataset Ω, since the most of the recordings in Ω' are *reliable*. This can be noted also in the DER of the subset Ω', 0.86%, which is lower than that obtained for the whole dataset. Actually, the DER of the complementary subset $\Omega \backslash \{\Omega'\}$ is very high (19.96%), since the weighted logistic regression

Subset	DER	σ_{DER}	Representativeness	DU
Ω	1.31%	4.58%	100.00%	85.63%
$\Omega_{correct}$	0.77%	1.34%	97.51%	89.55%
$\Omega_{incorrect}$	25.36%	12.93%	2.49%	-3.92%
$\Omega_{true\,positives}$	0.75%	1.30%	96.75%	89.05%
$\Omega_{true\,negatives}$	26.78%	12.91%	1.85%	-3.18%
$\Omega_{false\,positives}$	20.84%	12.57%	0.63%	-0.73%
$\Omega_{false\,negatives}$	3.67%	2.73%	0.77%	0.50%
Ω'	**0.86%**	**2.34%**	**97.38%**	**88.31%**
$\Omega\backslash\{\Omega'\}$	19.96%	15.38%	2.62%	-2.69%

Table 7.2: Results for the detection task using weighted linear logistic regression on the NIST SRE 2008

model is discarding many *unreliable* recordings. Note that $DU(\Omega\backslash\{\Omega'\})$ is negative, which means that this subset $\Omega\backslash\{\Omega'\}$ is harmful for the application and it is better to discard it.

In order to validate the weighted logistic regression model, a different dataset Ω is considered. This new dataset is extracted from the NIST SRE 2010 *summed* dataset. In this case, Ω is composed of 7044 five minute two-speaker telephone conversations selected from the 7130 summed channel recordings available in the NIST SRE 2010. These 7044 recordings are selected since their ASR transcriptions are provided by NIST, so it is possible to determine the correct diarization labels using these ASR transcriptions (see Section 3.2.1).

Subset	DER	σ_{DER}	Representativeness	DU
Ω	2.45%	4.87%	100.00%	72.79%
$\Omega_{correct}$	1.76%	1.84%	95.83%	77.57%
$\Omega_{incorrect}$	21.72%	11.15%	4.17%	-4.78%
$\Omega_{true\,positives}$	1.75%	1.81%	95.34%	77.35%
$\Omega_{true\,negatives}$	29.37%	10.94%	1.46%	-2.82%
$\Omega_{false\,positives}$	17.04%	8.75%	2.71%	-1.96%
$\Omega_{false\,negatives}$	5.33%	3.02%	0.48%	0.21%
Ω'	**2.08%**	**3.41%**	**98.06%**	**75.39%**
$\Omega\backslash\{\Omega'\}$	23.48%	14.07%	1.94%	-2.60%

Table 7.3: Performance of the detection task using weighted linear logistic regression on the NIST SRE 2010

Table 7.3 shows the results obtained for the detection task on the NIST SRE 10 dataset. In this case, the detection task enables us to select a more reliable subset Ω' so that $DU(\Omega')$ is more than 2.5% higher than that obtained for the whole dataset Ω. This increase is similar to that obtained for the NIST SRE 2008 dataset. Actually, the conclusions that can be extracted from table 7.3 are, in general, those extracted for table 7.2. Therefore, the weighted logistic regression model and the proposed confidence measures enable us to retrieve a subset of *reliable* recordings from a dataset, increasing the usefulness of the dataset for a speaker characterization application.

Comparing the results in Tables 7.2 and 7.3, we can see that the accuracy of the diarization system is higher for NIST SRE 2008 than for NIST SRE 2010 (1.31% against

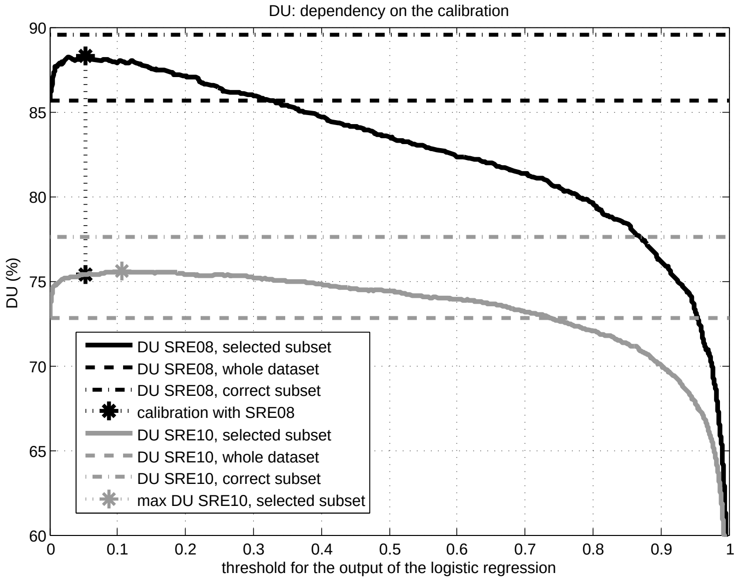


Figure 7.4: *Dependency of DU on the calibration*

2.45%), and so the percentage of correctly diarized recordings and the Dataset Usefulness is higher for NIST SRE 2008 (97.51% against 95.83% and 85.63% against 72.79% respectively). The diarization system shows a significantly lower accuracy on the the NIST SRE 2010 dataset. Exploring the causes of this reduced performance, we have detected that the percentage of overlapped speech over the net speech is much higher for NIST SRE 2010 than for NIST SRE 2008 dataset, 12.7% against 5.8%. This high percentage of overlapped speech is one of the main reasons for the lower accuracy shown on the NIST SRE 2010. Thus, although it is not tackled in this work, future work on overlapped speech detection must be carried out to avoid degradation in the proposed approach for speaker diarization.

Since the logistic regression model has been trained and calibrated on the NIST SRE 2008, and the proportion of incorrectly diarized recordings is notably different in both datasets (2.49% against 4.17%), the logistic regression model may not be optimal for NIST SRE 2010. The dependency of the $DU(\Omega')$ on the calibration for both datasets is analyzed in Figure 7.4. We can see that the optimum threshold for the output of the logistic regression is different for both datasets. On the other hand, $DU(\Omega')$ does not change significantly for variations on this threshold, so this approach is robust against calibration errors. Particularly, for the NIST SRE 2010 dataset, $DU(\Omega')$ is higher that $DU(\Omega)$ for most of the threshold range.

7.4 Evaluation on speaker verification

In this Chapter, a methodology to detect correctly diarized recordings in a dataset has been proposed. This approach makes use of a set of four confidence measures and a weighted logistic regression model in order to detect those recordings with *reliable* diarization hypotheses, in this study, those obtaining a DER is below 10%. The approach has shown to work properly, detecting most of the *reliable* diarization hypothesis, and introducing a relatively small number of false alarms.

The goal of this methodology is to obtain a subset suitable for speaker characterization, by means of detecting those recordings that may introduce degradation because of their incorrect diarization hypothesis. Once this recordings are detected, they could be discarded or processed by human supervision, depending on the size and the importance of the available and discarded data.

Therefore, to validate this methodology completely, the accuracy of a speaker characterization application must be analyzed for the selected and discarded subsets. For this purpose, the speaker verification task defined in Chapter 3 is considered as speaker characterization. The same experimental setup described in Section 3.2 is utilized and the verification task is evaluated in the three scenarios considered to analyze the speaker factor system in Section 4.3.2: the *mono-stereo*, *stereo-mono* and the *mono-mono* scenarios. For every scenario, the accuracy of the spekaer verification task considering the whole dataset is compared to that obtained considering the selected subset of *reliable* recordings and also to that obtained considering the discarded subset.

Since we consider the NIST SRE 2008 *summed* dataset for the evaluation of the speaker verification task, in this case, we use the NIST SRE 2010 *summed* dataset as development dataset to train the weighted logistic regression model. The NIST SRE 2010 *summed* dataset used for development is the one considered in Section 7.3.4 to validate task of detecting *reliable* recordings.

Subset	DER	σ_{DER}	Representativeness	*DU*
Ω	1.31%	4.58%	100.00%	85.63%
Ω'	**0.90%**	**2.62%**	**97.29%**	**87.72%**
$\Omega\backslash\{\Omega'\}$	18.15%	16.38%	2.71%	-2.09%

Table 7.4: Results for the detection task using weighted linear logistic regression on the NIST SRE 2008, using the NIST SRE 2010 as development dataset

Table 7.4 shows the results obtained for the task of detecting *reliable* recordings on the complete NIST SRE 2008 *summed* dataset when the NIST SRE 2010 *summed* dataset is considered development. The overall DER, the standard deviation of the DER values, the representativeness and the proposed figure of merit *DU* for the whole dataset and for the accepted subset Ω' and rejected $\Omega\backslash\{\Omega'\}$ subsets are displayed. Comparing this results to those presented in Table 7.2, it can be seen that using a different dataset for development introduces a slight degradation, but the detection task is still working very well, increasing the *DU* in more than 2% absolute.

7.4.1 *Mono-Stereo* Scenario

In this scenario we consider the *summed-short2* dataset from the NIST SRE 2008 for enrollment. We refer to this dataset as Ω_{enr}. The proposed approach for detecting *reliable* recordings is considered to determine the subset Ω'_{enr} of recordings that should be taken into account for enrollment in the speaker verification task.

Table 7.5 compares the quality of the diarization hypotheses obtained by the baseline and the speaker factor diarization system when considering the complete dataset and a *reliable* subset selected using the approach proposed in this Chapter. The results are obtained on the subset of the *summed-short2* condition of the NIST SRE 2008 considered for enrollment in the *mono-stereo* scenario. These results confirm that the proposed approach

Diarization System and subset	DER	σ_{DER}	Representativeness	*DU*
BIC AHC baseline, all data (Ω_{enr})	4.92%	9.08%	100.00%	46.79%
Spk Fact, all data (Ω_{enr})	1.11%	3.47%	100.00%	88.07%
Spk Fact, accept subset (Ω'_{enr})	**0.84%**	**1.90%**	**98.01%**	**89.15%**
Spk Fact, reject subset ($\Omega_{enr} \backslash \{\Omega'_{enr}\}$)	15.49%	15.16%	1.99%	-1.08%

Table 7.5: Comparison of the quality of the enrollment subset from the *summed-short2* condition considering the BIC AHC and the speaker factor diarization systems, for the complete dataset and selecting a subset of *reliable* recordings. The quality is presented in terms of overall DER, standard deviation of the DER values, representativeness and *DU*.

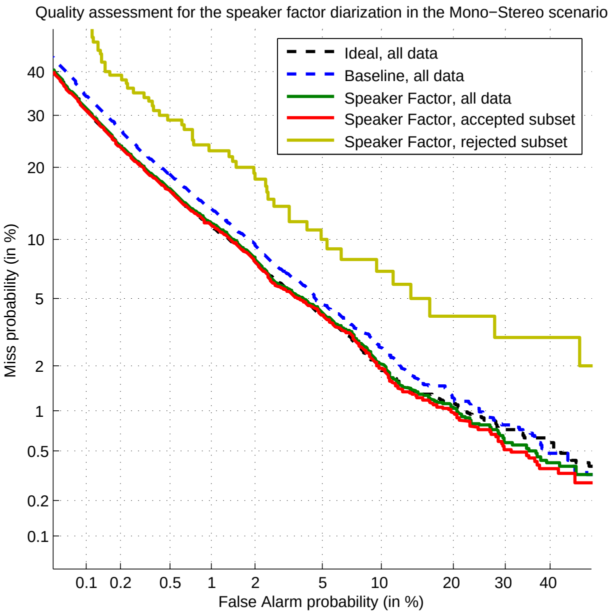


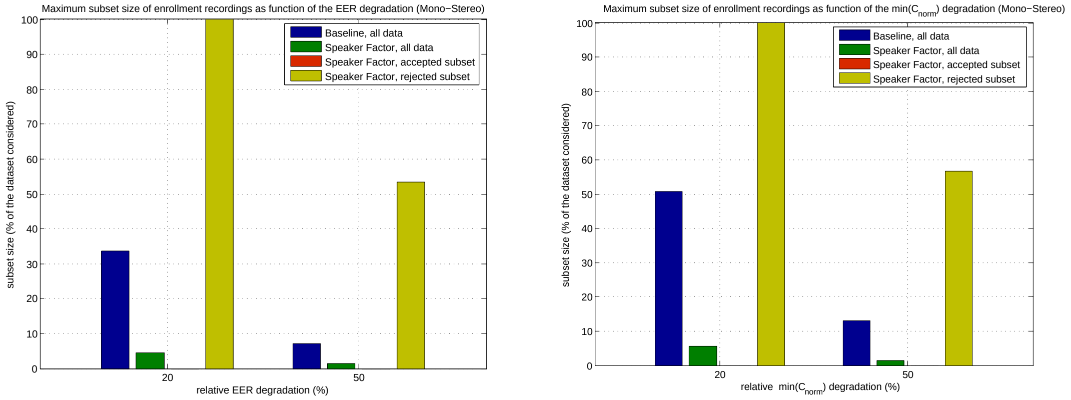
Figure 7.5: *DET curves considering the speaker factor diarization system for the complete dataset and selecting a subset of reliable recordings in the mono-stereo scenario. The DET curves obtained for the rejected subset and considering the baseline and ideal diarization systems are shown for comparison.*

for quality assessment is working as expected, retrieving a large subset (98.01%) of the original dataset, with very low DER (0.84%). This subset is expected to be more suitable for speaker verification than the complete dataset, since the *DU* has increased in more than 1% absolute. Note also the improvement in terms of *DU* introduced by the speaker factors system with respect to the BIC AHC baseline system. This improvement reflects the reduction of the DER and thus the increase in the number of recordings with $DER < 10\%$ achieved by the speaker factor diarization system.

Figure 7.5 shows the DET curves obtained in the *mono-stereo* scenario using the speaker factor system to diarize the enrollment dataset, considering the complete dataset and the accepted subset of *reliable* recordings. For comparison, the DET curves considering the rejected subset, and considering the whole dataset using the baseline BIC AHC and an ideal diarization systems are also shown. The difference between the DET curves obtained for the complete dataset and the selected subset is negligible. However, observing the DET curve obtained for the rejected subset, it can be seen that the selection of *reliable* recordings

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal, all data (Ω_{enr})	4.40% (0.00%)	0.2042 (0.00%)
Baseline, all data (Ω_{enr})	4.76% (8.18%)	0.2295 (12.39%)
Spk Fact, all data (Ω_{enr})	4.49% (2.05%)	0.2074 (1.57%)
Spk Fact, accept subset (Ω'_{enr})	**4.41% (0.22%)**	**0.2044 (0.09%)**
Spk Fact, reject subset ($\Omega_{enr}\backslash\{\Omega'_{enr}\}$)	7.51% (70.55%)	0.3138 (53.69%)

Table 7.6: EER and minimum C_{norm} for the ideal, the baseline and the speaker factor diarization system, considering for the later the complete dataset, and the accepted and rejected subsets, in the *mono-stereo* scenario. The degradation with respect to the results obtained with the ideal diarization system for the complete dataset is shown.



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of $min(C_{norm})$*

Figure 7.6: *Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER and $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the mono-stereo scenario.*

is working as expected. The rejected or discarded subset contains recordings that obtain low accuracy in the speaker verification task. Note that the rejected subset is very small since the speaker diarization is very accuracte in this dataset. Therefore, the DET curves for the complete dataset and the accepted subset are almost identical, since they both involve almost the same trials. However, for datasets where the accuracy of the diarization system is much lower, the size of the rejected subset will be bigger, and we can expected the DET curve for the accepted subset to show higher accuracy than that obtained for the complete dataset. In such a case, the improvement will be achieved at the expense of rejecting more recordings.

Similar conclusions can be extracted from Table 7.6, which shows the EER and $min(C_{norm})$ obtained in the same scenario for the diarization systems and subsets under analysis. The rejected subset shows a significant degradation with respect to the complete dataset, but since it is very small, discarding this subset does not improve significantly the results obtained for the accepted subset.

Finally, Figure 7.6 and Tables 7.7 and 7.8 show the percentage of recordings with highest DER of the subset considered in each case that can be accounted to keep the degradation in the EER (Figure 7.6(a)) and $min(C_{norm})$ (Figure 7.6(b)) with respect to the ideal diarization

Diarization system	$\%_{Deg(EER)} \geq 20\%$	$\%_{Deg(EER)} \geq 50\%$
Baseline, all data (Ω_{enr})	33.61%	7.20%
Spk Fact, all data (Ω_{enr})	4.46%	1.37%
Spk Fact, accept subset (Ω'_{enr})	**0.00%**	**0.00%**
Spk Fact, reject subset ($\Omega_{enr} \backslash \{\Omega'_{enr}\}$)	100.00%	53.33%

Table 7.7: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

Diarization system	$\%_{Deg(min(C_{norm}))} \geq 20\%$	$\%_{Deg(min(C_{norm}))} \geq 50\%$
Baseline, all data (Ω_{enr})	50.75%	13.03%
Spk Fact, all data (Ω_{enr})	5.49%	1.37%
Spk Fact, accept subset (Ω'_{enr})	**0.00%**	**0.00%**
Spk Fact, reject subset ($\Omega_{enr} \backslash \{\Omega'_{enr}\}$)	100.00%	56.67%

Table 7.8: Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-stereo* scenario.

system over certain value. It is interesting to observe that it is not possible to account even a small subset of recordings from the accepted subset Ω'_{enr} that obtains a degradation in terms of EER and $min(C_{norm})$ over 20%. This means that most of the degradation obtained considering all the dataset is introduced by the recordings discarded in the rejected subset.

7.4.2 *Stereo-Mono* Scenario

In the *stereo-mono* scenario we consider a subset from the NIST SRE 2008 *summed* condition as testing dataset. We refer to this dataset as Ω_{tst}. The proposed approach for detecting *reliable* recordings is considered to determine the subset Ω'_{tst} of recordings that should be taken into account for testing in the speaker verification task.

Diarization System and subset	DER	σ_{DER}	Representativeness	*DU*
BIC AHC baseline, all data (Ω_{tst})	5.20%	9.41%	100.00%	44.04%
Spk Fact, all data (Ω_{tst})	1.32%	4.63%	100.00%	85.52%
Spk Fact, accept subset (Ω'_{tst})	**0.90%**	**2.63%**	**97.28%**	**87.67%**
Spk Fact, reject subset ($\Omega_{tst} \backslash \{\Omega'_{tst}\}$)	18.15%	16.38%	2.72%	-2.15%

Table 7.9: Comparison of the quality of the testing subset from the *summed* condition considering the BIC AHC and the speaker factor diarization systems, for the complete dataset and selecting a subset of *reliable* recordings. The quality is presented in terms of overall DER, standard deviation of the DER values, representativeness and *DU*.

Table 7.9 compares the quality of the diarization hypotheses obtained by the baseline and the speaker factor diarization system when considering the complete dataset and a *reliable* subset selected using the approach proposed in this Chapter. The results are obtained on the subset of the *summed* condition of the NIST SRE 2008 considered for testing in the *stereo-mono* scenario. Again, the results confirm that the proposed approach for quality assessment

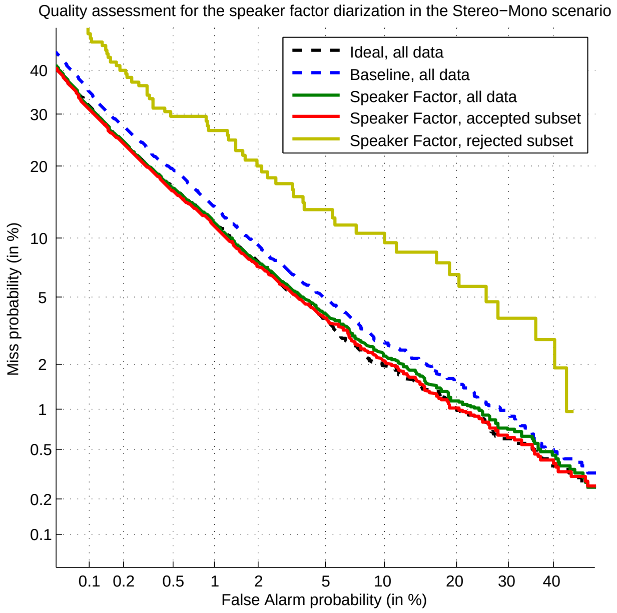


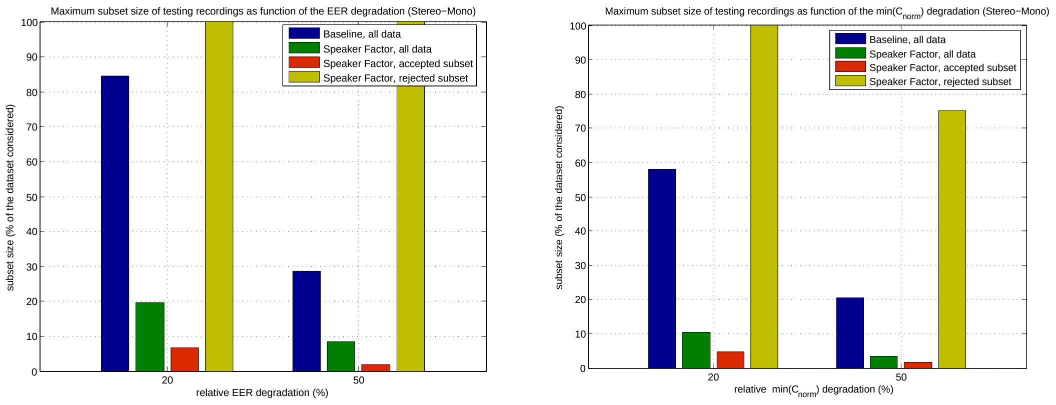
Figure 7.7: *DET curves considering the speaker factor diarization system for the complete dataset and selecting a subset of reliable recordings in the stereo-mono scenario. The DET curves obtained for the rejected subset and considering the baseline and ideal diarization systems are shown for comparison.*

is working as expected, retrieving a large subset (97.28%) of the original dataset, with very low DER (0.90%). This subset is expected to be more suitable for speaker verification than the complete dataset, since the DU has increased in more than 2% absolute. Note once more the improvement in terms of DU introduced by the speaker factors system with respect to the BIC AHC baseline system.

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal, all data (Ω_{tst})	4.23% (0.00%)	0.2102 (0.00%)
Baseline, all data (Ω_{tst})	4.94% (16.78%)	0.2334 (11.04%)
Spk Fact, all data (Ω_{tst})	4.39% (3.78%)	0.2097 (-0.24%)
Spk Fact, accept subset (Ω'_{tst})	**4.26% (0.62%)**	**0.2056 (-2.19%)**
Spk Fact, reject subset ($\Omega_{tst}\backslash\{\Omega'_{tst}\}$)	9.44% (122.99%)	0.3429 (63.16%)

Table 7.10: EER and minimum C_{norm} for the ideal, the baseline and the speaker factor diarization system, considering for the later the complete dataset, and the accepted and rejected subsets, in the *stereo-mono* scenario. The degradation with respect to the results obtained with the ideal diarization system for the complete dataset is shown.

Figure 7.7 shows the DET curves obtained in the *stereo-mono* scenario using the speaker factor system to diarize the testing dataset, considering the complete dataset and the accepted subset of *reliable* recordings. For comparison, the DET curves considering the rejected subset, and considering the whole testing dataset using the baseline BIC AHC and the ideal diarization systems are also shown. In addition the EER and $min(C_{norm})$ obtained for the diarization systems and subsets under analysis are shown in Table 7.10. In this scenario, the difference between the DET curves obtained for the complete dataset and



(a) *Maximum subset size depending on the degradation in terms of EER*

(b) *Maximum subset size depending on the degradation in terms of* $min(C_{norm})$

Figure 7.8: *Percentage of recordings of the enrollment dataset with highest DER that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, in the stereo-mono scenario.*

Diarization system	$\%_{Deg(EER)} \geq 20\%$	$\%_{Deg(EER)} \geq 50\%$
Baseline, all data (Ω_{tst})	84.43%	28.60%
Spk Fact, all data (Ω_{tst})	19.52%	8.40%
Spk Fact, accept subset (Ω'_{tst})	**6.77%**	**1.87%**
Spk Fact, reject subset ($\Omega_{tst} \backslash \{\Omega'_{tst}\}$)	100.00%	100.00%

Table 7.11: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

the selected subset is still negligible, although the difference is observable in the low false rejection region. Again, the DET curve for the rejected subset shows that the discarded recordings are degrading the overall accuracy, but the percentage of discarded recordings (2.72%) is small compared to the size of the dataset, so we do not get significant improvement removing these recordings. However, as mentioned before, we would obtain significant improvement if the speaker diarization system were not as accurate as it is for this dataset.

Finally, Figure 7.8 and Tables 7.11 and 7.12 show the percentage of recordings with highest DER of the subset considered in each case that can be accounted to keep the degradation in the EER (Figure 7.8(a)) and $min(C_{norm})$ (Figure 7.8(b)) with respect to the ideal diarization system over certain value. These statistics show that the accepted subset is quite reliable for speaker verification since the size of the subset obtaining certain degradation is very small. In fact, the percentage of recordings of the discarded subset that obtains a degradation in terms of EER and $min(C_{norm})$ over 20% and 50% is 100% or close to 100% in all cases.

7.4.3 *Mono-Mono* Scenario

The *mono-mono* scenario makes use of the *summed-short2* enrollment and *summed* testing datasets. The quality of these datasets depending on the diarization system and on the

Diarization system	$\%_{Deg(min(C_{norm}))} \geq 20\%$	$\%_{Deg(min(C_{norm}))} \geq 50\%$
Baseline, all data (Ω_{tst})	58.10%	20.43%
Spk Fact, all data (Ω_{tst})	10.44%	3.40%
Spk Fact, accept subset (Ω'_{tst})	**4.67%**	**1.63%**
Spk Fact, reject subset ($\Omega_{tst}\backslash\{\Omega'_{tst}\}$)	100.00%	75.00%

Table 7.12: Percentage of recordings of the testing dataset with highest DER that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *stereo-mono* scenario.

subsets considered is shown in Tables 7.5 and 7.9. This scenario may reflect with more clarity the impact of selecting a *reliable* subset for the task of speaker verification for two main reasons. First, in this scenario, we only accept those trials involving accepted recordings in both enrollment and testing sides. Therefore, the subset of trials considered will be smaller. Second, diarization errors are introduced in both enrollment and testing sides, so the degradation because of diarization errors will be higher.

Subset for Enrollment	Subset for Testing		
	all (Ω_{tst})	accepted (Ω'_{tst})	rejected ($\Omega_{tst}\backslash\{\Omega'_{tst}\}$)
all (Ω_{enr})	3211974 (100.00%)	3124494 (97.28%)	87480 (2.72%)
accepted (Ω'_{enr})	3145884 (97.94%)	3060204 (95.27%)	85680 (2.67%)
rejected ($\Omega_{enr}\backslash\{\Omega'_{enr}\}$)	66090 (2.06%)	64290 (2.00%)	1800 (0.06%)

Table 7.13: Number and rate of trials for all possible dataset and subset combinations from the *summed-short2* (Ω_{enr}) and *summed* (Ω_{tst}) datasets in the *mono-mono* scenario.

Table 7.13 shows the number and rate of trials for all possible dataset and subset combinations from the *summed-short2* (Ω_{enr}) and *summed* (Ω_{tst}) datasets in the *mono-mono* scenario. In those cases where all the data is considered, the 100% of the trials are evaluated. Whenever the accepted subsets are considered, only those trials involving accepted recordings in both enrollment and testing are accounted. These trials constitute 95.27% of the total number of trials. Whenever the rejected subsets are considered, all the remaining trials are accounted, that is, all the trials involving a rejected recording either on enrollment or testing sides. The percentage of rejected trials is thus 4.73% = 100.00% − 95.27% = 2.67% + 2.00% + 0.06%. Note that the number of accepted trials is smaller than in the previous scenarios, but it is still large to observe differences in the accuracy of the speaker verification system when comparing the complete set and the accepted subset of trials.

Figure 7.9 and Table 7.14 show the results of the speaker verification task obtained in the *mono-mono* scenario using the speaker factor system to diarize both enrollment and testing datasets, considering the complete set of trials and the accepted subset of *reliable* trials (trials with *reliable* recordings for both enrollment and testing). For comparison, the results considering the rejected subset of trials, and considering all the trials using the baseline BIC AHC and the ideal diarization systems are also shown. As expected, in this scenario it is possible to appreciate a slight improvement in the results obtained for the accepted subset when compared to the results for the complete set of trials. However, the improvement is still small since the size of the accepted subset is similar to the size of the complete set

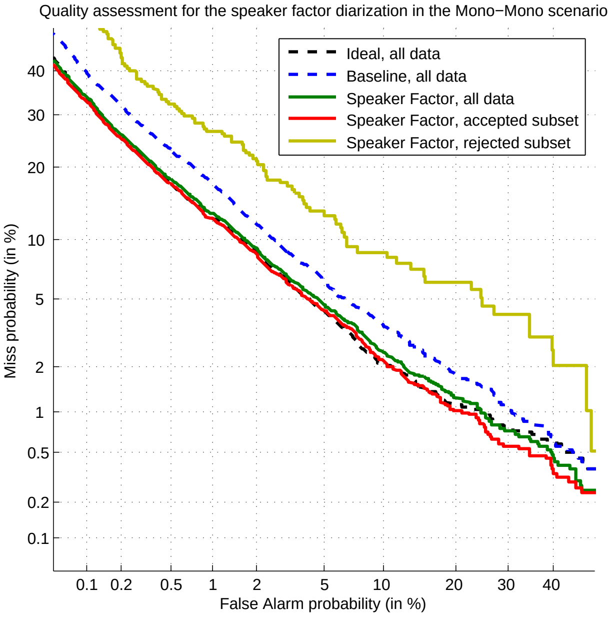


Figure 7.9: *DET curves considering the speaker factor diarization system for the complete dataset and selecting subsets of reliable recordings for enrollment and testing in the mono-mono scenario. The DET curves obtained for the rejected subsets and considering the baseline and ideal diarization systems are shown for comparison.*

of trials. Again, we need to analyze the rejected subset of trials to see that the detection system is discarding the *unreliable* ones. This can be observed in the DET curve as well as in the EER and $min(C_{norm})$ obtained for the rejected subset.

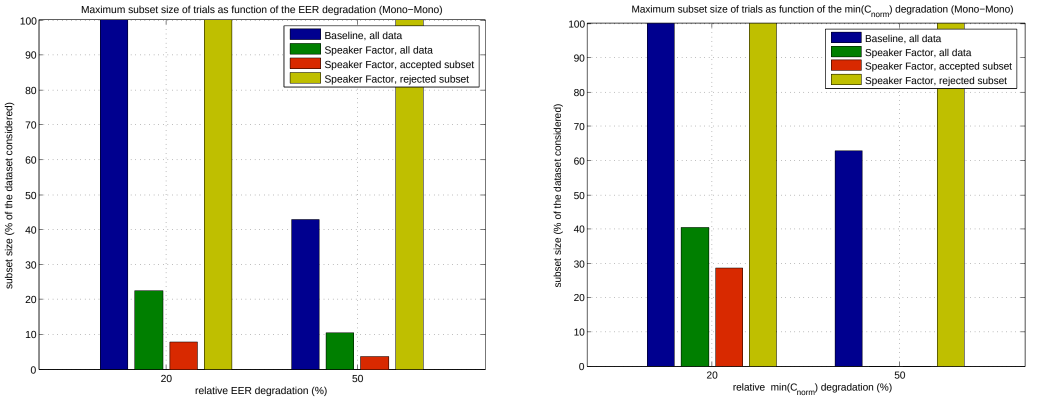
It is important to remark that the accepted subset of trials do not obtain significant improvement compared to the complete set of trials because the speaker factor diarization system is very accurate in this dataset, and thus the detection system accepts most of the recordings. For another dataset with higher DER, the number of *unreliable* recordings would be higher and would impact significantly on the accuracy of the speaker verification task when considering the complete set of trials. In such a situation, discarding these *unreliable* recordings and keeping only the accepted subset would yield and important improvement, at the expense of missing a large number of trials.

In the previous scenarios, the percentage of recordings with highest DER for a given dataset that can be accounted to keep the degradation in the EER and $min(C_{norm})$ with respect to the ideal diarization over certain value has been analyzed. In this case, since both enrollment and testing recordings are processed by the diarization systems, we study the percentage of trials in the speaker verification task that involve recordings with high DER in both enrollment and testing sides, as in Section 4.3.2.3. In those cases where a subset of trials is considered as initial set (accepted and rejected subsets of trials), the percentage is computed over the size of the initial set considered, instead of over the size of the complete set of trials.

Figure 7.10 and Tables 7.15 and 7.16 show the percentage of trials that can be accounted to keep the degradation in the EER (Figure 7.10(a)) and $min(C_{norm})$ (Figure 7.10(b)) with respect to the ideal diarization over certain value, considering those recordings with highest DER for enrollment and testing. Again these statistics show that the accepted subset is quite reliable for speaker verification. In addition, it can be observed that the discarded

Diarization system	EER (degradation)	$min(C_{norm})$ (degradation)
Ideal, all data	4.54% (0.00%)	0.2157 (0.00%)
Baseline, all data	5.53% (21.81%)	0.2695 (24.94%)
Spk Fact, all data	4.80% (5.73%)	0.2233 (3.52%)
Spk Fact, accept subset	**4.58% (0.93%)**	**0.2159 (0.09%)**
Spk Fact, reject subset	8.40% (85.00%)	0.3563 (65.20%)

Table 7.14: EER and minimum C_{norm} for the ideal, the baseline and the speaker factor diarization system, considering for the later the complete set and the accepted and rejected subsets of trials, in the *mono-mono* scenario. The degradation with respect to the results obtained with the ideal diarization system for the complete dataset is shown.



(a) *Maximum number of trials depending on the degradation in terms of EER*

(b) *Maximum number of trials depending on the degradation in terms of* $min(C_{norm})$

Figure 7.10: *trials in the mono-mono scenario that can be accounted to keep the degradation in terms of EER and* $min(C_{norm})$ *with respect to the ideal diarization system over certain value, considering the enrollment and testing recordings with highest DER, in the mono-mono scenario.*

subset contains most of the trials involving recordings that degrade the accuracy of the speaker verification task.

From these results, we can conclude that the selection of recordings with *reliable* diarization hypotheses using the confidence measures and the weighted logistic regression model proposed in this Chapter can help to improve the accuracy of speaker characterization applications. The improvement is obtained at the expense of discarding a subset of the dataset considered in the application, however, the discarded subset could be retrieved by means of human inspection. Human inspection is not feasible in most applications because of the size of the datasets considered, but in this study it has been shown that the speaker factor diarization system is accurate enough to obtain a high percentage of *reliable* recordings. Thus, given the small size of the rejected subsets, and depending on the application, human inspection might be feasible.

Diarization system	$\%_{Deg(EER)} \geq 20\%$	$\%_{Deg(EER)} \geq 50\%$
Baseline, all data	100.00%	42.78%
Spk Fact, all data	22.39%	10.50%
Spk Fact, accept subset	**7.89%**	**3.53%**
Spk Fact, reject subset	100.00%	100.00%

Table 7.15: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of EER with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

Diarization system	$\%_{Deg(min(C_{norm}))} \geq 20\%$	$\%_{Deg(min(C_{norm}))} \geq 50\%$
Baseline, all data	100.00%	62.80%
Spk Fact, all data	40.53%	0.00%
Spk Fact, accept subset	**28.55%**	**0.00%**
Spk Fact, reject subset	100.00%	100.00%

Table 7.16: Percentage of trials, involving those recordings with highest DER, that can be accounted to keep the degradation in terms of $min(C_{norm})$ with respect to the ideal diarization system over certain value, in the *mono-mono* scenario.

8

Use Cases

In this Chapter we present two use cases where the proposed technique for quality assessment for speaker diarization hypotheses can help to increase the accuracy of a final application. The first use case makes use of quality assessment to increase the accuracy of the speaker factor diarization system on two-speaker telephone conversations. The second one considers a speaker clustering task over two-speaker telephone conversations, where a *reliable* subset is selected with the purpose of increasing the accuracy of speaker clustering.

8.1 Quality Assessment for Hypothesis Generation and Selection

In this section we present an approach that makes use of the proposed diarization system presented in Chapter 4 considering intra-session variability compensation as described in Chapter 5, and the methodology for quality assessment presented in Chapter 7, in order to generate and select *reliable* diarization hypotheses for a given recording, increasing the diarization accuracy and thus increasing the DU for a given dataset (see Section 7.1). In addition, the *reliable* subset of correctly diarized recordings is detected to increase the DU further. The selected subset could be used for training speaker models in a speaker verification system from unlabeled data in an unsupervised or semi-supervised way, improving the accuracy of the system. This approach has been also presented in [Vaquero *et al.*, 2011a].

8.1.1 Diarization Hypothesis Generation and Selection

In Section 4.2.3, it is shown that a good initialization is critical for the correct operation of the speaker factor diarization system. Here, we propose a method to generate several diarization hypotheses for a conversation, using the confidence measures to select the best fragments to initialize the diarization system, and also to select the final diarization hypothesis.

In order to generate different diarization hypotheses for a given conversation, we split the conversation into halves iteratively until we obtain a set of non-overlapping slices of sufficient length. Figure 8.1 shows how a conversation is split into slices iteratively. The point to split every slice is selected to be in a silence segment, as close as possible to the middle of the slice.

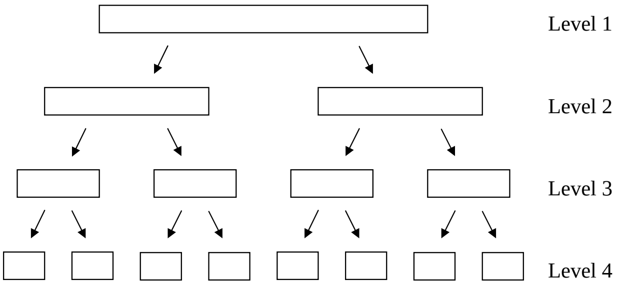


Figure 8.1: *Slice splitting diagram for diarization hypothesis generation.*

Thus, we obtain 2^{l-1} slices in every level l, and every slice is processed by the speaker factor diarization system independently. Therefore, for every conversation and level we obtain several slices with independent diarization hypotheses, and we can expect some of them to be correct even if the diarization hypothesis obtained for $l = 1$ (that is, for the whole conversation) is not correct. In fact, assuming that within a level l the diarization process for different slices is independent (which is reasonably since they are not overlapped), the number of slices correctly diarized in every level l follows a binomial distribution $K \sim B(2^{l-1}, p_{correct})$ where $p_{correct}$ is the probability for obtaining a correct diarization hypothesis for a given slice. Thus the probability of obtaining at least one slice correctly diarized is given by:

$$P(K \geq 1|2^{l-1}, p_{correct}) = \sum_{k=1}^{2^{l-1}} \binom{2^{l-1}}{k} p_{correct}^{k}(1 - p_{correct})^{(n-k)} = 1 - (1 - p_{correct})^{2^{l-1}} \quad (8.1)$$

If we assume that $p_{correct}$ does not depend on the length of the slice, the probability of obtaining at least one slice correctly diarized increases as the level l and so the number of slices increases. In general the accuracy of the proposed two-speaker diarization system will depend on the length of the slice to diarize, but if the conversation is long enough, $p_{correct}$ will not decrease significantly as l increases, for low values of l. Therefore, we can expect $P(K \geq 1|2^{l-1}, p_{correct})$ to increase as l increases. As an example, let us imagine a conversation which contains mostly clean speech, but at some point a severe noise appears in the recording. The diarization system may not work properly when processing the complete recording, but it may work when processing a slice that only contains clean speech.

Then, the idea is to use those slices correctly diarized at every level as a good initialization to process the whole conversation, obtaining a unique diarization hypothesis for every level. To do so, for every recording, we first select the best diarized slices at every level l. In order to select the best diarized slices at every level we make use of the confidence measures and the logistic regression model described in Chapter 7. The logistic regression model combines all the confidence measures obtaining a single fused confidence measure. At each level l, only the slice obtaining the maximum fused confidence and those obtaining a fused confidence a 20% below the highest one are kept, and the remaining slices are discarded.

Then, for each level l, we agglomerate the hypothetical speakers obtained by the diarization system in the selected slices using BIC based AHC. Every hypothetical speaker is modeled with a full covariance Gaussian, as in Section 6.5.1.1, and the hypothetical speakers are agglomerated until we obtain two clusters. The agglomeration is constrained to avoid merging two hypothetical speakers from the same slice. The two final clusters obtained are assumed to represent the two speakers involved in the conversation, and the whole recording is resegmented using 32-component GMM trained on those clusters as speaker models. A final soft-clustering resegmentation on the whole recording is done to refine the speaker boundaries. Note that this is equivalent to fed the obtained clusters directly to the Resegmentation stage of the speaker factor diarization system (see Figure 4.1 in Chapter 4).

This way, we generate a diarization hypothesis for every level l and then, the best one is selected as final diarization hypothesis for the recording. Again, in order to select the best diarization hypothesis for a given recording, the confidence measures and logistic regression model proposed in previous Chapter are considered. Thus, the selected diarization hypothesis is the one obtaining the maximum fused confidence measure.

8.1.2 Experimental Setup

To evaluate our approach for hypothesis generation and selection, we use NIST SRE 2008 and NIST SRE 2010 *summed* datasets (see Section 3.2.1). We consider that the NIST SRE 2008 *summed* dataset is available and labeled to train our logistic regression model and this approach will be validated on the NIST SRE 2010 *summed* channel dataset. Again we consider the *heavy-weight* configuration for the speaker factor diarization system, using 100 speaker factors and LDA 100 → 50 + WCCN for intra-session variability compensation. To select the best diarized slices and the best diarization hypothesis for every recording we use the four confidence measures and the logistic regression model proposed in Chapter 7. We consider four levels $l = 4$ for hypothesis generation and selection.

The results are presented in terms of DER, percentage of *reliable* recordings, which are those obtaining a DER below 10%, and DU for the complete dataset (see Section 7.1). In addition, we utilize the fused confidence measure to select the subset of *reliable* recordings of the dataset, presenting the overall DER, its standard deviation, the representativeness and the DU for the accepted and rejected subsets.

8.1.3 Experiments

Diarization	DER	$\%_{DER<10\%}$	$DU(\Omega)$
$l = 1$	1.31%	97.51%	85.63%
$l = 2$	1.23%	97.70%	86.43%
$l = 3$	1.45%	97.24%	83.97%
$l = 4$	1.73%	96.29%	81.70%
Max Conf hypotheses	**1.00%**	**98.69%**	**88.89%**
Min DER hypotheses	0.70%	99.14%	92.27%

Table 8.1: Accuracy of the diarization system with hypothesis generation and selection and *DU* for the NIST SRE 2008 dataset

Table 8.1 shows the results obtained for every level and for the proposed approach for hypothesis selection on the NIST SRE 2008 *summed* subset. As we can see, the accuracy in terms of DER, the number of *reliable* recordings, and thus $DU(\Omega)$ using the proposed approach for the selection of correct diarization hypotheses (Max Conf hypotheses entry in Table 8.1) is better than that obtained at every level. However, the results could be much better if we could always select the best diarization hypothesis among the four available hypotheses (Min DER hypotheses entry in Table 8.1). To increase the Dataset Usefulness further, we can try to detect the subset Ω' of *reliable* recordings, that is, those whose DER is below 10%.

Subset	DER	σ_{DER}	Representativeness	*DU*
Ω	1.00%	2.98%	100.00%	88.89%
$\Omega_{correct}$	0.79%	1.33%	98.69%	90.41%
$\Omega_{incorrect}$	21.47%	10.98%	1.31%	-1.52%
Ω'	**0.92%**	**2.38%**	**99.28%**	**89.35%**
$\Omega\backslash\{\Omega'\}$	13.39%	10.98%	0.72%	-0.46%

Table 8.2: Results for the detection task when considering the diarization system with hypothesis generation and selection on the NIST SRE 2008

Table 8.2 shows the results for the detection task on the selected diarization hypotheses for the NIST SRE 2008 *summed* dataset. This time the detection of *reliable* recordings is helpful, but not as significantly as shown in Chapter 7, in Table 7.2. The *DU* only increases 0.46% compared to the increase of 2.68% obtained when considering only $l = 1$ (Table 7.2). This is due to the fact that, after the hypothesis selection, most recordings are *reliable* and those that are not, are not far from the threshold. Note that selecting always the diarization hypotheses with maximum fused confidence measure for every recording increases the fused confidence measure value for the *unreliable* recordings.

However, comparing these results to those obtained at $l = 1$ with no detection of the *reliable* subset, the increase in terms of *DU* is quite significant, from 85.63% to 89.35%. Even considering as baseline the detection of *reliable* recordings at $l = 1$ there is a significant increase in the *DU* (from 88.31% to 89.35%).

Diarization	DER	$\%_{DER<10\%}$	$DU(\Omega)$
$l = 1$	2.45%	95.83%	72.79%
$l = 2$	2.43%	95.88%	73.07%
$l = 3$	2.46%	95.71%	72.64%
$l = 4$	2.73%	96.29%	69.93%
Max conf hypothesis	**2.12%**	96.96%	76.17%
Min DER hypothesis	1.73%	98.11%	80.56%

Table 8.3: Accuracy of the diarization system with hypothesis generation and selection and *DU* for the NIST SRE 2010 dataset

In order to validate the hypothesis generation and selection strategy, we test it on the NIST SRE 2010 dataset. Again we can see in Table 8.3 that this approach is useful to reduce the overall DER and thus to increase the reliability and usefulness of the dataset Ω, obtaining an increase of more than 3% in $DU(\Omega)$. Note that the results could be improved significantly if the confidence measures could always select the best diarization hypothesis for every recording.

Table 8.4 shows the results for the detection task on the selected diarization hypotheses for the NIST SRE 2010 *summed* dataset. It can be observed that after hypothesis generation and selection, we do not get great improvement by using the proposed approach to detect

Subset	DER	σ_{DER}	Representativeness	DU
Ω	2.12%	3.78%	100.00%	76.17%
$\Omega_{correct}$	1.74%	1.82%	96.96%	78.73%
$\Omega_{incorrect}$	18.24%	9.97%	3.04%	-2.56%
Ω'	**2.06%**	**3.49%**	**99.49%**	**76.53%**
$\Omega\backslash\{\Omega'\}$	15.84%	14.72%	0.51%	-0.46%

Table 8.4: Results for the detection task when considering the diarization system with hypothesis generation and selection on the NIST SRE 2008

the subset Ω' containing *reliable* recordings, as for the NIST SRE 2008. But again, the improvement in DU is significant compared to that obtained for $l = 1$. Using hypothesis generation and selection and detecting a *reliable* subset Ω', we obtain a $DU(\Omega')$ of 76.53% while for $l = 1$ we obtained a $DU(\Omega)$ of 72.79% for the whole dataset Ω, and 75.39% after detecting the *reliable* subset Ω'.

8.1.4 Comparison to other Diarization Approaches based on Speaker Recognition

The approach for speaker diarization presented in this thesis is a novel technique that makes use of recent advances in speaker recognition to improve speaker separability by means of inter-speaker variability modeling, intra-session variability compensation to reduce within speaker variability, and quality assessment to generate and select the best diarization hypothesis for a given recording. However, this is not the only approach in the literature that has been motivated by the recent advances in the field of speaker recognition. In the following sections, the main contributions in the field of speaker diarization that make use of speaker recognition techniques are briefly commented, and compared to the proposed approach.

8.1.4.1 Streaming Diarization using Speaker Factors

The first approach that made use of inter-speaker variability models and speaker factors to improve the accuracy of speaker diarization was presented in [Castaldo *et al.*, 2008]. This approach is similar to the one presented in this thesis, in the sense that it extracts a sequence of speaker factors that are treated as features by posterior processing stages.

The approach presented in [Castaldo *et al.*, 2008] does not use a Factor Analysis model to capture inter-speaker variability. Instead of that, PCA is performed over several speaker supervectors to build the eigenvoice matrix. The features considered also differ from those considered in this thesis. The features utilized in [Castaldo *et al.*, 2008] are processed using the same normalization techniques considered for speaker verification, but it is known that normalization techniques do not usually help for speaker diarization.

The stream of speaker factors is used to build a GMM with a number components equal to the expected number of speakers, and Viterbi segmentation is performed considering the Gaussian models extracted from the GMM. This system has been evaluated on the NIST SRE 2008 *summed* dataset, obtaining a DER of 5.8% that can be reduced to 4.6% using Viterbi Resegmentation in the MFCC space.

8.1.4.2 Variational Bayes and Speaker Factors

Recently, a set of approaches for speaker diarization based on inter-speaker variability modeling have been presented in [Reynolds *et al.*, 2009] and [Kenny *et al.*, 2010]. Among them, the most innovative and best performing system tries to model the joint distribution of the features extracted from the complete conversation considering the Factor Analysis model described in (4.1) and the probabilities of every segment considered to belong to one speaker or the other. Since the joint distribution of both variables is not tractable, a Variational Bayes (VB) approach is proposed [Valente and Wellekens, 2004], [Bishop, 2006]. This approach shows great potential since it enables the use of more information to estimate the posterior distribution of the speaker factors for every speaker based on the prior segment assignment (which is soft). This enables us to increase the number of speaker factors to values usually considered for speaker recognition. Thus, the segment reassignment is performed based on a robust estimation of the speaker factors. Once the segments are assigned to one speaker or the other, a Viterbi resegmentation pass is performed. Finally, the segmentation obtained with Viterbi is processed again by the VB algorithm to reassign the segments.

This system was firstly presented in [Reynolds *et al.*, 2009]. Achieving a DER of 3.8% on the NIST SRE 2008 *summed* dataset. Then, in [Kenny *et al.*, 2010], the system is refined, and the DER is reduced to 1.0% for the same dataset. Among the refinements, it is important to remark the use of features without any normalization or compensation (20 MFCC including C0), the use of a 1024-component UBM and 300 eigenvoices, and the consideration of all silence intervals as candidate speaker changes in the process.

8.1.4.3 Total Variability for Speaker Diarization

Finally, the most recent diarization system based on variability modeling has been proposed in [Shum *et al.*, 2011] at the time of the conclusion of this work. The system presented in [Shum *et al.*, 2011] is similar to the one presented in this thesis. However, instead of modeling inter-speaker variability, the total variability is modeled, including inter-speaker and inter-session variability, and i-vectors are extracted over short speech segments. The segments are defined by the silence intervals, and only these intervals are considered as speaker boundaries. As in this thesis, the Total Variability system makes use of PCA and K-means to cluster the i-vectors, but more than a single PCA dimension is considered. Then a Viterbi resegmentation is performed and the resulting segments are reassigned by a second pass, recomputing the i-vectors and the PCA+K-means clustering. This approach enables the authors to obtain a DER of 0.9% in the NIST SRE 2008 *summed* dataset.

8.1.4.4 Comparison and Discussion

Table 8.5 compares the accuracy of the mentioned speaker diarization systems to the one presented in this thesis, in terms of overall DER and the standard deviation of the DER values for all recordings for the NIST SRE 2010. The results for the diarization system proposed in this thesis are presented considering and not considering hypothesis generation and selection. When considered, the quality assessment strategy is trained on the NIST SRE 2010 *summed* dataset. The results for the remaining diarization systems are extracted from the literature, from [Kenny *et al.*, 2010] and [Shum *et al.*, 2011].

System	DER	σ_{DER}
Streaming spk fact	4.6%	8.8%
VB spk fact (2 passes)	1.0%	3.5%
i-vectors (2 passes)	0.9%	3.2%
Spk fact intra-ses. comp	1.3%	4.6%
Spk fact intra-ses. comp, q. assessment	1.0%	3.8%

Table 8.5: Comparison of state-of-the art speaker diarization systems for two-speaker telephone conversations.

Note that the proposed diarization system obtains competitive accuracy. Comparing the proposed approach to the VB based and the i-vector based it can be seen that the accuracy is quite similar. It is interesting to mention that there is not significant degradation when considering the NIST SRE 2010 *summed* dataset as development for the quality assessment methodology compared to the results obtained using NIST SRE 2008 directly as development data. So the quality assessment technique proposed seems to be robust across different datasets.

8.2 Speaker Clustering on Two-Speaker Conversation Datasets

It is usual in telephone environments to find large unlabeled datasets composed of two speaker conversations recorded on a single channel. In some situations, these datasets are intended for speaker modeling, however, they are not directly useful, unless two obvious problems are solved.

The first problem is the fact that two speakers are present on each recording. In Chapter 3 it is shown that the presence of two speakers in a recording degrades severely the performance of a speaker characterization system. As concluded previously, to avoid this effect, an accurate speaker diarization system is needed.

The second problem is the fact that a single speaker is usually present in several recordings within the dataset. Thus, it is mandatory to cluster those segments extracted from different recordings that belong to the same speaker for two main reasons. Firstly, we want to avoid obtaining different models for a single speaker. Secondly, using more than a single session to model every speaker provides robustness to the speaker models.

In this section, we address the speaker clustering problem on datasets composed of two-speaker telephone conversations. A system that makes use of the diarization and clustering techniques previously presented is proposed, and the impact of the diarization error on the accuracy of the clustering process is studied. Finally, quality assessment is considered in order to detect *reliable* recordings and thus to improve the accuracy of the clustering task. This work is also presented in [Vaquero *et al.*, 2011b].

8.2.1 System Description

The proposed approach for speaker clustering on two-speaker conversation datasets follows four steps. First, every conversation is processed by the speaker factor diarization system,

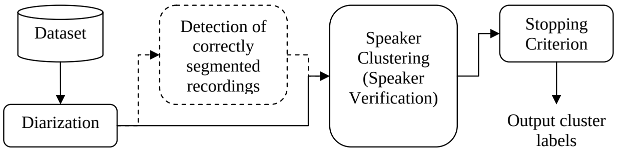

Dataset
Diarization
Detection of correctly segmented recordings
Speaker Clustering (Speaker Verification)
Stopping Criterion
Output cluster labels

Figure 8.2: *Block Diagram of the system for partitioning of two-speaker conversation datasets.*

using the best available configuration, in order to segregate the two speakers present. This configuration considers the *heavy-weight* speaker factor diarization system, using 100 speaker factors and LDA 100 → 50 + WCCN for intra-session variability compensation. Then, an optional detection of *reliable* conversations can be performed, in order to avoid feeding impure segments into the clustering process. Then, speaker clustering is performed over the remaining segments. For speaker clustering, the simplified PLDA AHC approach is considered (see Section 6.3.3, Chapter 6). Finally, a stopping criterion estimates the number of speakers in the dataset. These four steps are summarized in Figure 8.2.

8.2.2 Experimental Setup

The datasets considered for testing the proposed system are the NIST SRE 2008 and 2010 *summed* conditions. From the NIST SRE 2008 *summed* dataset, only 2894 out of the 4426 sides (2213 × 2) of the conversations that can be extracted belong to a speaker whose identity is provided by NIST. Thus, the clustering problem is evaluated only on the segments corresponding to these 2894 sides of the conversation. This dataset contains 1040 different speakers.

Since quality assessment will use the NIST SRE 2008 *summed* condition as development dataset, a subset of the NIST SRE10 *summed* channel dataset is considered to validate the quality assessment paradigm in the proposed speaker clustering task. This subset contains 2794 recordings, and only 3198 out of the 5588 sides of the conversations that can be extracted belong to a known speaker. The speaker clustering task is evaluated on these 3198 sides of the conversation. This dataset contains 461 different speakers.

The accuracy of the speaker diarization system is evaluated in terms of DER as usual. The accuracy of the speaker clustering system is measured in terms of cluster impurity (I_c) and speaker impurity (I_s) as defined in Section 2.6, Chapter 2. The cluster impurity increases as segments containing different speakers are clustered together, while the speaker impurity increases as the set of segments belonging to a single speaker are assigned to different clusters. In every experiment, the point of equal impurity (EI, $I_c = I_s$) is obtained as measure of accuracy of the clustering system. To analyze the impact of the detection of correctly diarized recordings, we will study the EI as well as the fraction of speakers that have been kept in the dataset after discarding incorrectly diarized recordings.

This study is not focused on the stopping criterion, but on the impact that the diarization errors may produce in the accuracy of the clustering task. Thus, it is assumed that the stopping criterion is capable of stopping at the point if EI in all experiments.

8.2.3 Impact of diarization errors on speaker clustering

To analyze the impact of the diarization error on the speaker clustering task, we evaluate the system described in 8.2.1 without quality assessment for speaker diarization. We consider several diarization systems obtaining different accuracies and we study the clustering accuracy depending on the diarization system considered. The diarization systems considered are listed below.

- The BIC AHC baseline, with a soft-clustering resegmentation pass [Reynolds *et al.*, 2009]
- The *light-weight* diarization system without resegmentation passes.
- The speaker factor diarization system with the best available configuration (intra-session variability compensation) described in Chapter 5.
- The reference diarization labels.

system	DER	EI	C
BIC AHC + soft-clustering	4.09%	18.87%	1314
Light-weight spk fact, no reseg	2.89%	18.04%	1304
Spk fact, best config.	**1.31%**	**16.97%**	**1289**
Reference labels	0.00%	15.38%	1122

Table 8.6: *Diarization accuracy (in terms of DER), speaker clustering accuracy (in terms of EI) and number of clusters (C) obtained for different diarization systems, on the NIST SRE 2008 summed dataset.*

Table 8.6 shows the diarization and speaker clustering accuracy for every speaker diarization approach evaluated on the NIST SRE 2008 *summed* dataset, in terms of DER and EI, as well as the number of clusters (C) obtained for the EI point. It can be seen that the accuracy of the speaker clustering task degrades significantly as the accuracy of the diarization system degrades, even though the approaches for speaker diarization we are presenting here obtain very low DER. Even the best configuration of the speaker diarization system, that obtains a DER as low as 1.31%, degrades the EI from 15.38% to 16.97%. Therefore, for this use case, the most accurate speaker diarization system available is needed. So from now on, only the speaker factor system with the most accurate configuration is considered (see Chapter 5).

The significant degradation obtained in the speaker clustering task given the low DER obtained can be explained by the fact that, although the overall DER is low, there are a few recordings with high DER. Thus, we are feeding audio segments containing two speakers into the clustering system, and these segments will mislead the iterative clustering process. Note that a single error during the AHC process can propagate and degrade the accuracy iteration after iteration.

Therefore, it would be interesting to detect those recordings with DER high enough to degrade the accuracy of the clustering task, so they can be discarded or processed manually. To do so, we need to set a threshold in the DER so that the recordings with DER below that threshold will not degrade the clustering performance. This threshold may be different

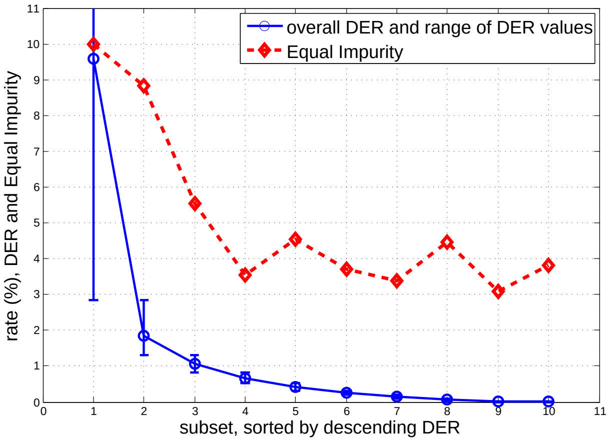


Figure 8.3: *DER and EI for ten subsets of the same size from the NIST SRE 2008 summed dataset, sorted by descending DER.*

from the one considered for speaker verification ($DER < 10\%$). In order to determine this threshold, we rank all the recordings in the NIST SRE 2008 *summed* dataset according to their DER. We split the ranked dataset into ten subsets of the same size, so that the variations in the accuracy of the clustering task observed across different subsets will not be due to their size. Note that it is known that the accuracy of the clustering task depends on the size of the problem, so we need to compare results for subsets of equal size [van Leeuwen, 2010]. Thus, we obtain ten subsets of the same size containing recordings with different DER values.

Figure 8.3 shows the accuracy of the diarization system in terms of DER and the accuracy of the speaker clustering system in terms of EI for every subset. The solid curve represents the overall DER for every subset when they are sorted by descending DER, as well as the range of DER values for the recordings in the corresponding subset. We can see that only the first subset obtains high DER values, and that all recordings belonging to the last seven subsets obtain a DER below 1%.

The dashed curve represents the EI value when performing the clustering task on every subset. The values are below those presented in Table 8.6 since the subsets are smaller (one tenth of the dataset) and contain fewer speakers than the whole dataset considered in Table 8.6 (around 240 per subset, note that a single speaker can be present in more than one subset). As we could expect, the EI decreases as the DER decreases, oscillating around 4% for the last seven subsets. It is interesting to note that even for low DER values, the EI can be very high, as it can be seen in for the second subset. This subset contains recordings with DER below 2.84%, but the EI is as high as 8.83%.

From this results we can conclude that the accuracy of the speaker clustering task is very sensitive to the DER obtained for the recordings in the dataset. According to figure 8.3, we see that the threshold in the DER value to obtain a performance not affected by the diarization error is very low. In this study, we consider $th_{DER} = 1\%$. This th_{DER} value is quite different from that considered in Chapter 7 for the task of speaker verification, which was as high as 10%. Note that the clustering system considered makes use of the same PLDA speaker verification system utilized in previous chapter to validate the quality assessment methodology, so we could expect both thresholds to be related. The significant difference in the threshold values is due to the fact that the clustering problem is solved using an iterative bottom-up clustering strategy, so slight degradations in the accuracy of speaker verification due to diarization errors will propagate through the speaker clustering process, degrading the final accuracy.

8.2.4 Speaker clustering with detection of *reliable* conversations

In order to evaluate the accuracy of the speaker clustering system when only those recordings detected as correctly diarized are fed into the clustering process, we make use of the quality assessment strategy presented in Chapter 7. The quality assessment strategy is trained to detect those diarization hypothesis obtaining a DER below 1%.

Segments	DER	EI	C	S	$\frac{S_{miss}}{S}$
All(2894)	1.31%	16.97%	1289	1040	0.00%
Correct(2146)	0.26%	12.18%	980	926	10.96%
Incorrect(748)	4.56%	15.29%	486	476	54.23%
Accept(1996)	**0.39%**	**11.85%**	**934**	**896**	**13.85%**
Reject(898)	3.50%	14.92%	565	517	50.29%

Table 8.7: *Accuracy of the clustering task using quality assessment for speaker diarization, on the NIST SRE 2008.*

Table 8.7 shows the accuracy of the clustering task using quality assessment to detect those recordings correctly diarized (those with $DER < 1\%$), for the NIST SRE 2008 *summed* dataset. The number of clusters obtained C, the number of speakers present in the detected recordings S and the ratio of speakers missed during the discarding process ($\frac{S_{miss}}{S}$) are shown for comparison. The accepted segments ("accept") are those that belong to a conversation detected by our system as correctly diarized, while the correct segments ("correct") are those that belong to a conversation that is actually correctly diarized ($DER < 1\%$).

We can see that the proposed approach for quality assessment is able to detect 1996 segments with an overall DER as low as 0.39%, obtaining a reduction in the EI from 16.97% to 11.85%, at the expense of missing 13.85% of the speakers in the dataset. If we could select those correctly diarized recordings, the EI would increase slightly to 12.18%, but fewer speakers would be missed (10.96%). Looking at the rejected subset of recordings ("reject"), we can see that the detection approach is working as desired, discarding a set of conversations whose overall DER is as high as 3.50% and would obtain an EI of 14.92%, which is very high given that the EI decreases as the subset is smaller.

Segments	DER	EI	C	S	$\frac{S_{miss}}{S}$
All(3198)	2.45%	13.21%	771	461	0.00%
Correct(756)	0.32%	6.55%	332	309	32.97%
Incorrect(2242)	3.19%	13.55%	674	447	3.04%
Accept(2382)	**1.58%**	**12.00%**	**654**	**440**	**4.56%**
Reject(816)	5.14%	14.52%	371	325	29.50%

Table 8.8: *Accuracy of the clustering task using quality assessment for speaker diarization, on the NIST SRE 2010.*

Since the logistic regression model considered for quality assessment is trained on the NIST SRE 2008 *summed* dataset, we consider the NIST SRE 2010 *summed* subset described in Section 8.2.2 to validate the proposed system. Table 8.8 shows the performance of the clustering task using quality assessment to detect those recordings correctly diarized (those with $DER < 1\%$), for the NIST SRE 2010 *summed* dataset.

In this case, the quality assessment approach detects many more conversations as correctly diarized than those that actually are: a total of 2382 out of 3198 segments are accepted as correctly diarized, while only 756 out of 2382 are actually correctly diarized. Because of this, the overall DER obtained for the accepted recordings is over 1%, and the EI is far over the EI obtained for the correct segments, 12.00% against 6.55%. However, the detection of correctly diarized recordings is still helping in this case: the DER of the accepted recordings is below the DER obtained for all recordings, and the EI is reduced from 13.21% to 12.00%, missing only 4.56% of the speakers.

Note that the difference between the performance obtained for the whole dataset and for the correct subset is in part due to the significant difference in the size of both datasets, and that 32.97% of the speakers are missed in the correct subset. The detection of correctly diarized recordings is introducing a high number of false alarms, but it is still discarding very *unreliable* recordings (DER of 5.14%), so it produces a little improvement in the EI, missing a small number of speakers.

8.2.5 Conclusion

The problem of speaker clustering on datasets containing two-speaker conversations has been addressed. It has been shown that the accuracy of the diarization system is critical for this problem, since the speaker clustering system is very sensitive to diarization errors. In fact, the DER should be below 1% in order to ensure that no degradation is obtained in the task of speaker clustering.

Quality assessment can be very helpful for this task since the detection of those diarization hypotheses with low DER enables us to avoid feeding into the speaker clustering system segments containing two speakers that will degrade the accuracy of the clustering task significantly. The results confirm that quality assessment is helping, but better and more robust confidence measures are needed to ensure that most of the incorrect recordings are rejected.

On the other hand, it has been shown there exists a trade-off between the accuracy of the clustering algorithm and the number of speakers kept in the dataset when detecting correctly diarized recordings. Thus, the approach for quality assessment proposed in Chapter 7 should be trained depending on the application needs.

Part V

Discussion and Conclusions

9

Conclusions

In this Chapter we analyze the speaker diarization techniques proposed in this thesis comparing their accuracy and computational cost with the accuracy and cost of the traditional BIC AHC diarization system. This analysis enables us to extract the main conclusions and contributions of this work, which along with the future lines of work are also presented in this Chapter.

9.1 Discussion

The techniques proposed in this thesis have shown to increase significantly the accuracy of a speaker verification task, and this increase has shown to affect the accuracy of a speaker characterization system, considering a speaker verification application. In this section, we analyze the increase in accuracy obtained by the proposed approaches for speaker diarization taking into account the computational cost. As accuracy measures we consider the DER for speaker diarization, obtained on the NIST SRE 2008 *summed* dataset and also the degradation of a speaker verification task that makes use of the NIST SRE 2008 as testing dataset, in the *stereo-mono* scenario. The degradation is measured against the results obtained with an ideal diarization system, in terms of EER.

As measure of the computational cost we consider the average time it takes to process an audio of certain duration, five minutes in this study. It is not our intention to perform an exhaustive analysis on the computational cost of every approach, but simply to compare them not only in terms of accuracy but in terms of computational cost as well. Also, it is interesting to identify the main bottlenecks in every approach.

9.1.1 Computational cost

In order to estimate the computational cost of the techniques proposed in this thesis, we measure the total time it takes to process the NIST SRE 2008 *summed* dataset, and we compute the average time it takes to process a single recording $\bar{T}_{process}$. This dataset is composed of 2213 with the same duration (five minutes), so the average processing time will be estimated simply as the total time divided by 2213. The average processing time is estimated for every stage of every approach considered for speaker diarization, in order to identify the main bottlenecks in every approach. In addition, the Real Time Factor (RTF) is computed as the ratio between the average processing time and the average duration of the recordings considered ($RTF = \frac{\bar{T}_{process}}{\bar{T}_{recording}}$). The total processing time is measured running the diarization process in a single core of an *Intel Nehalem 2.33 GHz (64 bits)* with 24Gb RAM.

The speaker diarization systems under analysis are listed next:

- **BIC AHC:** As baseline speaker diarization system, we consider the BIC AHC system described in Section 3.4.1.

- **Speaker Factor *light-weight*:** The Speaker Factor system proposed in Chapter 4 is studied, considering the *light-weight* configuration (see Section 4.2).

- **Speaker Factor *heavy-weight*:** The *heavy-weight* configuration for the speaker factor system is also analyzed.

- **Speaker Factor with intra-session variability compensation:** The best configuration available including intra-session variability compensation is considered. This configuration computes speaker factor vectors with dimension $R = 100$, and then reduces the dimensionality to 50 by means of LDA. Finally, the speaker factors are normalized using WCCN and fed into the initial clustering stage (see Chapters 4 and 5).

- **Speaker Factor with intra-session variability compensation, assuming unknown number of speakers:** The previous system is also analyzed under the assumption of not knowing the number of speakers, considering the architecture proposed in Section 6.5.1, and using BIC AHC and a threshold in the ΔBIC value to determine the number of speakers.

- **Speaker Factor with intra-session variability compensation and Quality Assessment:** Finally, the same speaker factor system with intra-session variability compensation is considered, assuming that the number of speakers is known and equal to two, using quality assessment in order to keep only those recordings whose DER is below 10%, as proposed in Chapter 7. Note that this system increases the overall diarization accuracy at the expense of missing some recordings that are detected as difficult to diarize.

The stages of the different systems considered in this analysis are those presented in Figure 6.2. These stages are listed below:

- **Front End:** The Front End includes feature extraction for all systems and speaker factor computation for all the systems based on speaker factors.

- **Variability Compensation:** The Variability Compensation stage includes the LDA and WCCN transformations on the speaker factor vectors. Note that only those systems considering intra-session variability compensation make use of this stage.

- **Initial Clustering/Segmentation:** This stage includes the PCA+K-means initialization in case of speaker factor diarization systems and the BIC segmentation process using a sliding window when the BIC AHC baseline system is considered.

- **Core Segmentation:** The Core Segmentation stage comprises the segmentation process carried out in the speaker factor systems when speaker factor vectors are considered as features.

- **Clustering:** The Clustering stage performs AHC on a initial segmentation considering ΔBIC as distance metric until a stopping criterion is met. Only the BIC AHC baseline system, which uses the output od the Initial Segmentation stage as initial segmentation, and the speaker factor system under the assumption of unknown number of speakers make use of this stage. The former assumes that the stopping criterion is met when the number of clusters is equal to two (the number of speakers is priorly known). The later uses a threshold in the ΔBIC value as stopping criterion.

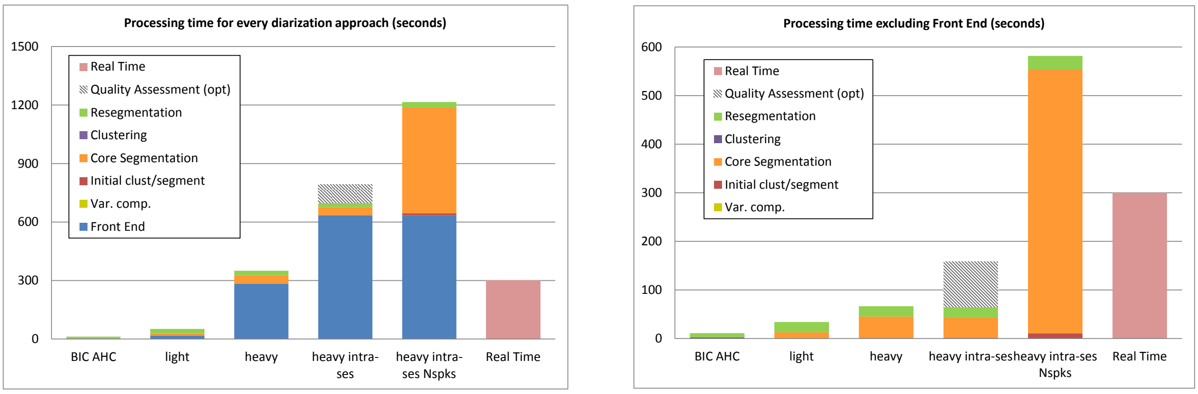
- **Resegmentation:** The Resegmentation stage comprises the Viterbi resegmentation and the soft-clustering passes in the MFCC space. All systems perform both the Viterbi and the soft-clustering passes except the BIC AHC baseline systems, which only considers a Viterbi resegmentation pass.

- **Quality Assessment (optional):** The last stage, not included in Figure 6.2, computes the confidence measures and determines whether or not a hypothetical two-speaker segmentation must be considered. Only the system with Quality Assessment considers this option. The tag “optional” is displayed since the speaker factor system with quality assessment is identical to the speaker factor system with intra-session variability compensation when this last module is added. Thus, this system can include this stage to perform Quality Assessment or not.

Stage	Diarization System				
	BIC AHC	light	heavy	intra-ses (+ Q.)	intra-ses Nspks
Front End	1.03 s	17.56 s	283.22 s	633.75 s	633.75 s
Var. comp.	N/A	N/A	N/A	0.38 s	0.38 s
Initial clust/seg	0.30 s	0.52 s	0.86 s	0.82 s	10.30 s
Core Seg	N/A	12.21 s	44.32 s	42.35 s	542.64 s
Clustering	2.20 s	N/A	N/A	N/A	0.01 s
Resegmentation	8.53 s	21.30 s	21.31 s	21.35 s	28.23 s
Q. Assess. (opt)	N/A	N/A	N/A	93.45 s	N/A
Total time T	12.06 s	51.59 s	349.71 s	698.65 (+ 93.45) s	1215.31 s
Real Time Factor	0.04	0.17	1.17	2.33 (+ 0.31)	4.05

Table 9.1: Average processing time for the proposed diarization systems, stage by stage. The Quality Assessment stage for the speaker factor system with intra-session variability compensation is considered optional.

Table 9.1 shows the average processing time required by every diarization system for processing a five minute conversation containing two speakers. The average processing time is also presented for every stage in every diarization approach. For clarity sake, the average processing times are also presented in Figure 9.1. It can be seen that the proposed approaches based on the JFA paradigm are significantly more costly than the BIC AHC baseline approach.



(a) *Average processing time (in seconds) for every speaker diarization approach, segregated by stage.*

(b) *Average processing time (in seconds) for every speaker diarization approach, segregated by stage. The Front End stage is excluded.*

Figure 9.1: *Average time in seconds to process a recording with a duration of 300 seconds, for every speaker diarization approach, segregated by stage. The average duration of a recording to process is also shown for comparison (Real Time).*

9.1.1.1 Analysis by stage

It is interesting to notice that, for the speaker factor systems, most of the processing time is invested in the Front End stage. In fact, the Front End stage includes the speaker factor extraction which is a costly operation and comprises the main bottleneck of these approaches. It seems interesting to analyze this stage deeply, in order to determine the most costly step and to explore techniques to reduce the overall processing time.

In this study, the Front End stage can be divided into three operations:

- **Feature Extraction:** This operation is shared by all systems, however, the BIC AHC and the *light-weight* configuration for the speaker factor diarization system consider 12 MFCC with no Δ features ($D = 12$), while the *heavy-weight* configuration considers 19 MFCC + Δ features ($D = 38$). We do not expect the feature extraction operation to be decisive regarding computational cost. This is the only step needed in the Front End for the BIC AHC system. That is the main reason why the BIC AHC diarization system is the fastest among all analyzed.

- **Sufficient Statistics:** In the JFA paradigm, once the input features are available, the first step to extract the speaker factor vectors is to compute the sufficient statistics on the UBM [Kenny *et al.*, 2007]. This is usually a costly operation, that depends on the dimension of the feature vector D and the number of components C in the UBM. Note that the *light-weight* configuration considers $C = 256$ while the *heavy-weight* configuration considers $C = 1024$.

- **Speaker Factors:** The last step is to compute the speaker factor vectors from the sufficient statistics, frame by frame over a sliding window. This operation is usually not very costly compared with the computation of the sufficient statistics in the field of speaker recognition. However, in the proposed approaches, this operation is performed frame by frame, and thus this operation becomes critical. The cost of this operation is related to the dimension of the GMM supervectors considered $C \times D$ (see Section

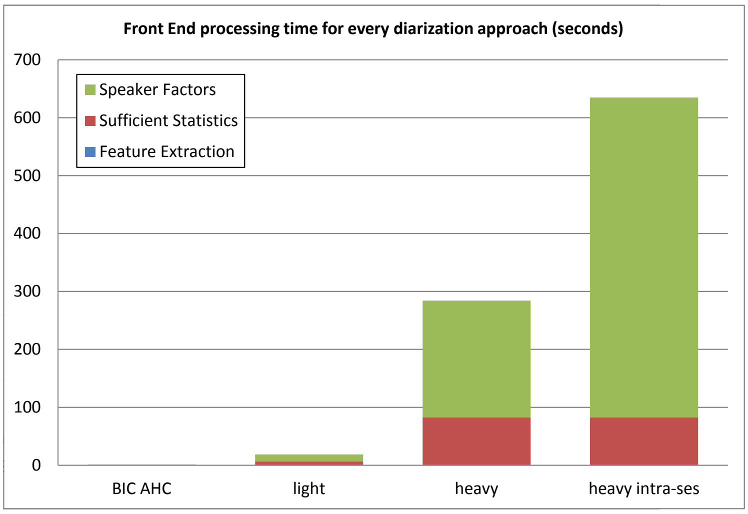


Figure 9.2: *Average processing time (in seconds) of the Front End stage for every speaker diarization approach. The processing time is segregated by Front End operation.*

4.1), and to the dimension of the speaker factor vectors R. Note that the *light-weight* configuration considers $R = 20$, *heavy-weight* configuration considers $R = 50$, while this last configuration with intra-session variability compensation considers $R = 100$.

Note that the speaker factor system with intra-session variability performs exactly the same operations in the Front End when the number of speakers is known or unknown. The same applies when considering quality assessment or not. Thus we study the Front End for four systems (BIC AHC, *light-weight*, *heavy-weight*, and *heavy-weight* with intra-session variability compensation).

Front End op.	Diarization System			
	BIC AHC	light	heavy	intra-ses
Feature Extraction	1.03 s	1.03 s	1.12 s	1.12 s
Sufficient Statistics	N/A	5.58 s	81.39 s	81.39 s
Speaker Factors	N/A	11.98 s	201.83 s	552.36 s
Total time T_{FE}	1.03 s	17.56 s	283.22 s	633.75 s

Table 9.2: Average processing time for the Front End of the proposed diarization systems, operation by operation.

Table 9.2 and Figure 9.2 show the average time the Front End stage takes to process a recording of five minute duration, segregated by operation. As expected, the Feature Extraction takes insignificant time compared to the other operations. The computation of the Sufficient Statistics is a costly operation, but it is not the most costly for any of the speaker factor diarization systems. Note that the cost for the *heavy-weight* configuration is identical using or not intra-session variability compensation since both approaches consider the same features and the same UBM. Also note the significant difference when computing the Sufficient Statistics for the *light-weight* and *heavy-weight* configurations.

But the most costly operation in all cases is the computation of the speaker factors. This is the main bottleneck in the *heavy-weight* configurations for the speaker factor diarization system. Note how the processing time increases as higher dimension for the speaker factors is considered. The reason of this high cost is the fact that a speaker factor vector is computed for every frame, every 10 ms. This time could be reduced taking into account that the speaker factor vectors are computed over a window of 1 second (see Chapter 4). This means that two consecutive speaker factor vectors share 99% of the frames considered to compute them, and thus will be very similar.

Therefore, two ideas can be proposed to reduce the processing time when computing the speaker factors, assuming that the computational cost is critical for our application. Firstly, the overlap between two consecutive windows considered for computing speaker factors can be reduced. Let us assume that the speaker factors are computed every t_{step} frames, and thus the overlap between two consecutive windows is $\frac{100-t_{step}}{100}$ and the processing time will be $\frac{1}{t_{step}}$ of the original time. In the limit, we can consider no overlap, and compute a speaker factor vector every second, reducing the computational time down to 1% of the original time. This implies that the subsequent stages should take into account that there will be only a single speaker factor for every 100 frames. The second idea is to find an approximation to compute most of the speaker factor vectors, given some speaker factor vectors computed exactly. For example, only one speaker factor vector could be computed every 10 frames, and the remaining speaker factors could be interpolated. This way, the processing time of the speaker factor computation would be roughly reduced to one tenth of the time, and there would be no need to modify the subsequent stages.

Note that these two proposed approaches will reduce the computational cost at the expense of losing diarization accuracy. A further study is needed in order to determine an optimal t_{step} or a costless way to approximate the speaker factors accurately.

Coming back to Figure 9.1, we can analyze the cost of the remaining stages. First, note that the variability compensation stage is really cheap. The problem of intra-session variability compensation by means of LDA + WCCN is that the dimension of the initial speaker factor vectors is higher, and thus the cost in the Front End stage is much higher as studied previously. However, after intra-session variability compensation, the dimension of the speaker factor vectors is reduced to 50, so the cost of the subsequent stages is not expected to increase. Note that in Chapter 5 it is shown that WCCN can be considered on its own for intra-session variability compensation, obtaining significant improvement in terms of DER. Thus, given the low cost of intra-session variability compensation it seems interesting to use it for all applications.

It can be observed that the cost of the Initial Clustering/Segmentation stage is insignificant in all systems. However, the cost of the Core Segmentation stage is quite significant, especially for the solution for unknown number of speakers. This stage estimates a GMM model for the speaker factor vectors and reassigns the speaker factor vectors among the available speakers using Viterbi decoding. As we can expect, the higher the dimension of the speaker factor vectors, the higher the processing time. However, when the number of speakers is known, since it is equal to two, the GMM only has two components and only two HMM are considered during Viterbi decoding. When the number of speakers is unknown, we consider a maximum of ten speakers. Thus the cost of the GMM estimation and the Viterbi decoding will increase significantly. In addition, the Core Segmentation runs iteratively until convergence, and convergence is reached when two consecutive diarization outputs are identical. We have observed that, as we could expect, fewer iterations are needed when the number of speakers S is small (around 4.7 in average for all systems when $S = 2$). For a high number of speakers, the number of iterations needed to reach convergence is much higher (around 10.2 for $N = 2$), increasing the total processing time. A solution for this problem in those applications where the computational cost is critical could be to limit the number of iterations or even to remove the Core Segmentation stage. It is known that the Initial Clustering stage provides very accurate diarization outputs (see Tables 4.2 and 4.9 in Chapter 4). Thus, the processing time can be significantly reduced at the expense of a slight reduction of the diarization accuracy.

The cost of the Clustering stage is related to the initial number of segments, as observed in Chapter 6. The BIC AHC baseline system generates a large amount of segments (around one every three seconds of speech, eighty for every recording in average) in the Initial Segmentation stage, while the speaker factor system when the number of speakers is unknown is forced to generate ten segments in the Initial Clustering and Core Segmentation stages. Thus, the Clustering stage is slower for the BIC AHC system than for the speaker factor system.

The Resegmentation stage is not significantly costly in the speaker factor diarization systems, but it might be costly if the cost of the other stages is reduced. The most costly step within the Resegmentation stage is the soft-clustering resegmentation. The diarization accuracy is not significantly increased by the soft-clustering pass as it was observed in Table 4.2, and a very high accuracy can be obtained simply considering a Viterbi resegmentation pass, in case the computational cost is critical for our application.

Finally, the Quality Assessment stage is also significantly costly. To reduce the computational cost of this stage some confidence measures could be selected and others discarded. The most costly confidence measures are those that need to compute speaker factors or i-vectors since they need to compute the sufficient statistics. This time the computation of the speaker factors or the i-vectors themselves given the sufficient statistics is not critical, since only two vectors need to be computed for each recording. In fact most of the processing time required for Quality Assessment is invested in the computation of the sufficient statistics needed to extract the i-vectors for the PLDA LLR confidence measure. The sufficient statistics needed to extract the speaker factors for the cosine distance and the eigenvalue spread have been computed previously in the Front End, so there is no need to compute them again (see Section 7.2). Thus, the PLDA LLR confidence measure could be removed for those applications where the computational cost is critical.

9.1.1.2 Importance of computational cost: online diarization

It has been seen that there are diarization systems faster than others, but the question is how fast a diarization system must be. The requirements of a diarization system in terms of processing time are completely dependent on the application. Some applications will require fast diarization systems while for others the computational cost may not be an issue. As an example, we can think about applications that provide information in *Real Time*, and thus they need to perform diarization online. This applications will require a diarization system as fast or faster than *Real Time* (five minutes in this case), with $RTF \leq 1$. Thus, those diarization systems obtaining $RTF > 1$ must be optimized or modified in order to perform online diarization.

Note that there are several issues when building and online diarization system, even when considering techniques that are faster than *Real Time*. In fact, online diarization is a new technology that has not been deeply studied by the research community. If we think about the main stages of the diarization systems under study, we can identify some stages that are directly performed over a fixed and small amount of audio information, as the Front End and the Variability Compensation, and other stages that need the complete conversation acquired up to the moment to be performed correctly. The former stages do not present a problem for online diarization as far as they are faster than *Real Time*, but the later stages must be modified in order to work properly.

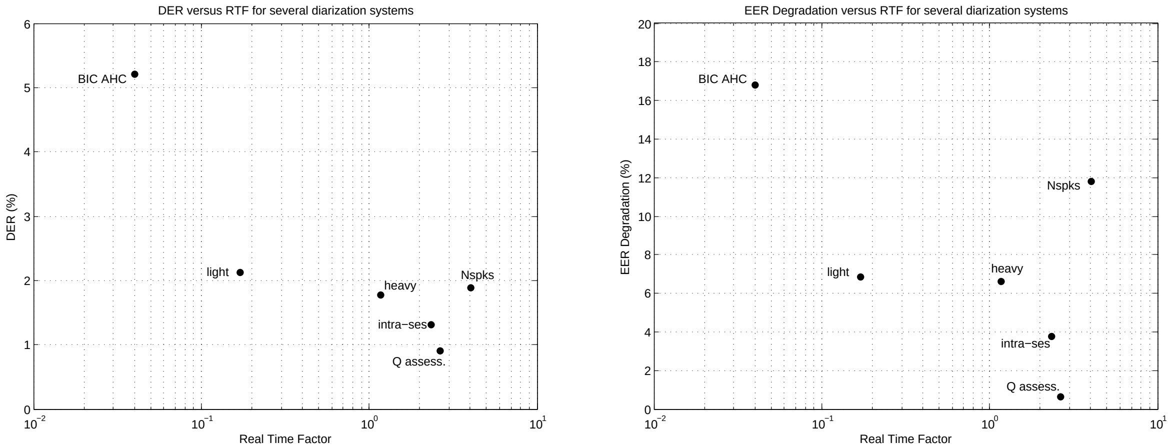
Two methodologies are usually taken into account when designing online diarization systems. The first one is a streaming approach, where the audio signal is processed in large chunks, and the diarization output of every new chunk is combined with the outputs obtained for the previous chunks. This approach is considered in [Castaldo *et al.*, 2008]. In this technique, there is a trade-off between the accuracy of the diarization output on every chunk and the latency of the system. It is usual to consider chunks of around one minute duration, so that the diarization system can process the chunks accurately, and every time a chunk is processed, a clustering algorithm merges the new clusters obtained with those obtained in the previous chunks. Thus the latency of this approach is the size of the chunk plus the time required to process a chunk.

Therefore, this approach can not be strictly considered as an online diarization system, since the latency is very high to produce an *Real Time* output. The main advantage of this approach is that general purpose diarization systems as those studied in this thesis can be directly considered to perform diarization on every chunk, as far as they obtain $RTF \leq 1$. Those diarization systems that do not obtain $RTF \leq 1$, could be speed up with the techniques proposed previously, in order to be used in this streaming approach.

Assuming that there are more than two speakers in the conversation, the size of the chunk also affects the maximum number of speakers that could be found in a single chunk. If the chunk is very small, a two-speaker diarization system could be used. However, for large chunks in which we expected to find more than two speakers, the system must look for the maximum number of expected speakers in the chunk and determine the actual number of speakers. Note that most of the processing time of the speaker factor system when the number of speakers is unknown is (ignoring the Front End stage) is invested in the Core Segmentation stage. We explained that this high cost is due to the fact that the Core Segmentation looks for ten speakers instead of two, and this increases the cost of the Core Segmentation models and also the number of iterations needed to converge. Considering shorter chunks will enable us to reduce the number of speakers, and thus to reduce the computational cost.

The second methodology for online speaker diarization is based on the use of the information available in every moment to make online speaker identification on an audio stream. Once a diarization hypothesis is available, speaker models can be build for every hypothetical speaker and the input audio stream can be decoded online. In order to obtain reliable speaker models, an accurate offline speaker diarization system must be running in background, considering all available information for the current audio stream, or a chunk large enough to obtain an accurate diarization output. The diarization output provided by the accurate offline system is used to build speaker models and then these models are considered to decode the audio stream online. The offline diarization system is always running in background in order to provide more reliable speaker models as more information is available and also to detect new speakers in the audio stream.

This hybrid solution was proposed in [Vaquero *et al.*, 2010c], where it is also shown that in order to perform online diarization in this fashion, the background offline diarization system must be as accurate as possible but also as fast as possible. In fact, the faster the



(a) *DER against RTF for the diarization systems under analysis.*

(b) *EER degradation against RTF for the diarization systems under analysis. The stereo-mono scenario is considered.*

Figure 9.3: *Accuracy against computational cost for the diarization systems under analysis, measured on the NIST SRE 2008 summed dataset. The accuracy os measures in terms of DER and degradation of the EER in a speaker verification task that makes use of the diarization system to process the testing dataset. The computational cost is measure in terms of RTF.*

background diarization system, the earlier the online system can detect new speakers and include new information in the current speaker models. An interesting fact of this approach is that even a diarization system with $RTF > 1$ can be considered as background offline diarization system. The problem is that the latency introduced will increase and the time required to include new speakers and to retrain speaker models will be much higher.

9.1.2 Comparison of diarization approaches

Once we known the influence of the accuracy of a speaker diarization system in a speaker characterization application and the importance of the computational cost in diarization systems, we can compare the approaches studied in this thesis.

System	DER	$Deg(EER)$	RTF
BIC AHC	5.21%	16.78%	0.04
Spk Fact. light	2.12%	6.86%	0.17
Spk Fact. heavy	1.77%	6.62%	1.17
Spk Fact. intra-ses	1.31%	3.78%	2.33
Spk Fact. Nspks	1.88%	11.81%	4.05
Spk Fact. Q. assess.	0.90%	0.62%	2.64

Table 9.3: DER, $Deg(EER)$ and RTF for the speaker diarization systems under study, obtained on the NIST SRE 2008 *summed* condition. The $Deg(EER)$ is obtained on a speaker verification task that considers the NIST SRE 2008 *summed* condition as testing dataset

Table 9.3 and Figure 9.3 compare the accuracy and computational cost for the speaker diarization systems under study. The accuracy is measured in terms of DER and degradation of the EER in a speaker verification task that makes use of the diarization system to process the testing dataset. The computational cost is measure in terms of RTF.

It can be seen that those systems obtaining higher accuracy are those obtaining higher computational cost. In the previous section some methodologies to reduce the computational cost of the proposed systems have been commented. Those methods will reduce the computational cost at the expense of reducing the accuracy. However, the relative ordering by computational cost of the diarization approaches would not change significantly when applying those methods. All speaker factor based systems can be speed up, but not enough to achieve the low cost of the BIC AHC system.

Therefore, as we could expect, there is a trade-off between accuracy and computational cost, and the best diarization system will be the one that fulfills the application requirements on each case. We can think on applications where the computational cost is not an issue, for example, an application that diarizes recordings that will be used later for speaker model enrollment in a speaker verification system. This application may have to process a dataset offline, and without special time requirements. On the other hand, if we think in a surveillance application for monitoring thousands of conversations online, the computational cost will be critical, since the available computational resources and the computational cost will limit the number of conversations to monitor in parallel. We also can think in an application that would require high accuracy and low computational cost, for example, online rich transcription for meetings or broadcast news.

Note also that the most costly system is the only one among all presented that could be used for conversations including more than two speakers, unless the conversations are processed in short chunks so that it is not expected to find more than two different speakers in every chunk.

9.2 Conclusions and Future Lines

In previous Chapters we have presented new techniques for speaker diarization, which have shown to outperform the traditional approaches and to improve the the accuracy of s speaker characterization application that make use of the diarization labels. In this section we present the main conclusions and contributions of this thesis along with the future research lines that can be started from this work.

9.2.1 Conclusions

In this section we summarize the main conclusions of this work.

In the first part of this thesis the importance of speaker diarization for speaker characterization applications has been analyzed. We have shown that in those cases when a speaker characterization application operates on recordings containing multiple speakers, the use of speaker diarization is mandatory. Otherwise, huge degradation in the accuracy of the speaker characterization application is obtained.

The accuracy of the diarization system required in order to obtain no degradation due to the presence of multiple speakers in a speaker characterization application depends on the application. In Chapter 3 we have studied the maximum diarization error that is acceptable in order to obtain no degradation in a speaker verification task. It has been observed that those diarization hypotheses obtaining a DER below 5% obtain no significant degradation and that the degradation in the speaker verification task increases dramatically whenever the DER exceeds 10%.

We also have studied a traditional diarization system based on state-of-the-art techniques (BIC AHC) as processing diarization system for the speaker verification task. It has been seen that, even though this diarization approach increases significantly the accuracy of the speaker verification task compared to that obtained when no diarization system is considered, there is still significant degradation due to diarization errors in certain scenarios.

Therefore, in the second part of this thesis, we have explored new approaches for speaker diarization with the objective of reducing the diarization error and thus the degradation obtained in the speaker verification task, when considering audios containing only two speakers. In Chapter 4 a new technique for speaker diarization based on the JFA paradigm has been proposed. This technique makes use of prior models on inter-speaker variability to enhance speaker separability. It has shown significant improvement in terms of DER when compared to the traditional BIC AHC system. This improvement in diarization accuracy has also been reflected in a significant reduction of the degradation obtained in the speaker verification task, in all scenarios considered.

However, inter-speaker variability is not the only type of variability present in those audio signals that contain more than a single speaker. In Chapter 5, we have analyzed the three types of variability that may appear in a set of audio signals: inter-speaker, inter-session and intra-session variability. We have shown that the modeling of inter-speaker variability increase the accuracy of speaker diarization, while the presence of intra-session variability degrades the diarization accuracy, so it seems appropriate to remove it. Two techniques (LDA and WCCN) have been proposed to remove intra-session variability, and both of them have shown significant increase in diarization accuracy. This increase has also been translated into an increase in the accuracy of the speaker verification task, up to a point where the degradation in the speaker verification task due to diarization errors is negligible in some scenarios.

In addition, it has been seen that capturing inter-session variability in the inter-speaker variability model enhances speaker separability since, in most cases, different speakers use different communication channels. However, inter-session variability could be removed in the same fashion as intra-session variability in those scenarios where it might be harmful.

In the third part of this thesis we have studied the task of speaker clustering in large datasets. New techniques to solve this task have been proposed in Chapter 6. Most of these techniques are based on prior work on the use of speaker verification systems for speaker partitioning of datasets. The use of a state-of-the-art (i-vector PLDA) speaker verification system combined with a stopping criterion based on a threshold for the output LLR seems to be the best solution.

Moreover, the proposed approaches for speaker clustering in large datasets have been applied to the task of clustering segments obtained from the same recording, which is the traditional speaker clustering task considered for diarization. It has been shown that these approaches can be used for clustering in speaker diarization, but they have not been able to outperform the traditional BIC AHC approach for clustering. Thus, the accurate diarization system presented in the second part of this thesis has been expanded with a BIC based clustering approach in order to cluster the speech segments obtained and to determine the actual number speakers. This system has shown little degradation compared to the same system under the assumption of knowing the number of speakers, but as it has been shown in section 9.1.1, it needs much longer time to process a recording.

Finally, in the fourth part we have proposed an approach for quality assessment for speaker diarization that enables us to validate a given diarization hypothesis. This methodology, presented in Chapter 7, is useful to segregate those audio signals from a given dataset that can be considered in a speaker characterization application obtaining no significant degradation due to diarization errors. In fact, it has been shown that the proposed methodology for quality assessment enables us to select a subset of audio signals whose diarization hypotheses obtain a DER below a desired threshold, that will depend on the speaker characterization application. The selected subset obtains a DER significantly lower than the complete dataset, and thus the speaker verification task obtains lower degradation for the selected subset than for the complete dataset.

Therefore, we can ensure a good behavior of the speaker characterization application regardless the accuracy obtained in speaker diarization at the expense of discarding a subset of audio signals from the dataset. The discarded subset can be processed manually, in a semi-supervised fashion, and thus considered for the speaker characterization application, or simply ignored depending on the importance of the discarded data and on the amount of available data.

In addition, the proposed strategy for quality assessment can be useful to increase the diarization accuracy by means of generating several diarization hypothesis and selecting the one obtaining highest quality, as shown in Chapter 8.

9.2.2 Contributions

In this section the main contributions of this thesis are enumerated.

- **Speaker Factor based Diarization:** Although it is not a concept introduced in this thesis, the use of the JFA paradigm and speaker factors for diarization has been studied extensively in Chapter 4. We have introduced new methods for taking advantage of the increased speaker separability provided by the speaker factors, and we have analyzed the best configuration for every stage involved in the diarization process. As a result, the proposed system significantly outperforms the traditional BIC AHC diarization system and the first approach for speaker factor based diarization proposed in [Castaldo *et al.*, 2008].

- **Variability Compensation for Speaker Diarization:** The different types of variability sources involved in the task of speaker diarization has been studied in Chapter 5. This is a novel work since this study had never been done, even though the task of speaker diarization is not new. As a result of this study, we have observed that the use of inter-speaker variability models enhance speaker separability, but also that the presence of intra-session variability degrades the accuracy of a speaker diarization system. The techniques proposed to remove intra-session variability have shown to increase the diarization accuracy, proving the harmful effect of this type of variability. Finally, the presence of inter-session variability has not shown to degrade the accuracy of the speaker diarization system, but to enhance it.

- **Speaker Clustering in Large Datasets:** Several strategies for speaker clustering in large datasets have been proposed and studied in Chapter 6, showing that they can also be applied for speaker clustering within the task of speaker diarization.

- **Quality Assessment for Speaker Diarization:** We have proposed a quality assessment technique for speaker verification. The concept of confidence measures and quality assessment has been studied in other fields of speech technologies, but never in the field of speaker diarization. The set of confidence measures and the methodology proposed in Chapter 7 enables us to validate any diarization hypothesis for a two-speaker conversation.

As proof of the quality of this work, we can enumerate several contributions to journals, book chapters and international conferences, submitted during the development of this work:

- C. Vaquero, O. Saz, E. Lleida, J.M. Marcos, C. Canalís. Vocaliza – A Computer-Aided Application for Spanish Speech Therapy. IV Jornadas en Tecnología del Habla. Zaragoza, Spain, 2006. [Vaquero *et al.*, 2006].

- O. Saz, C. Vaquero, E. Lleida, J.M. Marcos, C. Canalís. Study of Maximum A Posterior Speaker Adaptation for Pathological Speech. IV Jornadas en Tecnología del Habla. Zaragoza, Spain, 2006. [Saz *et al.*, 2006].

- W. Ricardo Rodríguez, C. Vaquero, O. Saz, E. Lleida. Aplicación de las Tecnologías del Habla al Aprendizaje del Prelenguaje y el Lenguaje. IV Congreso Latinoamericano de Ingenieria Biomedica, Isla Margarita, Venezuela, 2007. [Rodriguez *et al.*, 2007].

- C. Vaquero, O. Saz, E. Lleida. Tecnologías del habla para el desarrollo del lenguaje. TELECOM I+D 2007, Valencia, España, 2007. [Vaquero *et al.*, 2007].

- C. Vaquero, O. Saz, E. Lleida and W. R. Rodríguez. Human language Technologies for speech therapy in Spanish language. LangTech 2008, Rome, Italy, 2008. [Vaquero *et al.*, 2008a].

- C. Vaquero, O. Saz, E. Lleida and W. R. Rodríguez. E-Inclusion technologies for the speech handicapped. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, USA, 2008. [Vaquero *et al.*, 2008b].

- W. R. Rodríguez, C. Vaquero, O. Saz, E. Lleida. Speech Technology Applied to Children with Speech Disorders. IV International conference on Biomedical Engineering. Kuala Lumpur, Malaysia 2008. [Rodriiguez *et al.*, 2008].

- Jesús A. Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, José E. García, Luís Buera, Óscar Saz. I3A submission for the NIST SRE 2008. NIST Speaker Recognition Evaluation 2008. Montreal, Canada, 2008. [Villalba *et al.*, 2008b].

- Jesús A. Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, José E. García, Luís Buera, Óscar Saz. Experiencia del I3A en la Evaluación de Reconocimiento de Locutor NIST 2008. IV Jornadas de Reconocimiento Biométrico de Personas. Valladolid, España, 2008. [Villalba *et al.*, 2008a].

- W. R. Rodríguez, O. Saz, E. Lleida, C. Vaquero, Antonio Escartín. COMUNICA — Tools for Speech and Language Therapy. Workshop on Child, Computer and Interaction (ICMI'08 post-conference worshop). Chania, Crete, 2008. [Rodriguez *et al.*, 2008].

- Oscar Saz, W.-Ricardo Rodríguez, Eduardo Lleida, Carlos Vaquero, Antonio Escartín. COMUNICA - Plataforma para el Desarrollo, Distribución y Evaluación de Herramientas Logopédicas Asistidas por Ordenador. V Jornadas en Tecnologías del Habla. Bilbao, Spain, 2008. [Saz *et al.*, 2008].

- O. Saz, E. Lleida, C. Vaquero. Analysis of Acoustic Features in Speakers with Cognitive Disorders and Speech Impairments. Special Issue, 2009. [Saz *et al.*, 2009b].

- O. Saz, V. Rodríguez, E. Lleida, W.R. Rodríguez, C. Vaquero. An Experience with a Spanish Second Language Learning Tool in a Multilingual Environment. Workshop on Spoken Language Technology for Education (SLaTE). Wroxall Abbey, UK, 2009. [Saz *et al.*, 2009c].

- Carlos Vaquero, Nicolas Scheffer, and Sachin Kajarekar. Impact of Prior Channel Information for Speaker Identification. III IAPR/IEEE International Conference on Biometrics: Advances in Biometrics, volume 5558 of Lecture Notes in Computer Science, pages 443–453. Springer Berlin / Heidelberg. Alghero, Sardinia, 2009. [Vaquero *et al.*, 2009].

- Jesús Villalba, Eduardo Lleida, Alfonso Ortega, Carlos Vaquero, and Antonio Miguel. I3A System for Evalita 2009 Speaker Verification Application Evaluation. In Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, Italy, 2009. [Villalba *et al.*, 2009].

- Carlos Vaquero, Alfonso Ortega, Jesús Villalba, Antonio Miguel, and Eduardo Lleida. Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification. In Proc Interspeech 2010, volume 2010, pages 2310–2313, 2010. [Vaquero *et al.*, 2010a].

- Carlos Vaquero, Oriol Vinyals, and Gerald Friedland. A Hybrid Approach to Online Speaker Diarization. In Proc Interspeech 2010, volume 2010, pages 2638–2641, 2010. [Vaquero *et al.*, 2010c].

- Carlos Vaquero, Alfonso Ortega, Eduardo Lleida. Intra-session Variability Compensation for Speaker Segmentation. FALA 2010. Vigo, Spain, 2010. [Vaquero *et al.*, 2010b].

- Diego Castán, Alfonso Ortega, Carlos Vaquero, Antonio Miguel, Eduardo Lleida. VIVOLAB-UZ Audio Segmentation System for Albayzin Evaluation 2010. FALA 2010. Vigo, Spain, 2010. [Castán *et al.*, 2010].

- Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, W. Ricardo Rodrgúez. Tools and Technologies for Computer-Aided Speech and Language Therapy. Speech Communication, Vol 51(10), 948–967, 2009. [Saz *et al.*, 2009a].

- Oscar Saz, Eduardo Lleida, Victoria Rodríguez, W. Ricardo Rodríguez, Carlos Vaquero. The Use of Synthetic Speech in Language Learning Tools: Review and a Case Study. Computer Synthesized Speech Technologies: Tools for Aiding Impairment (book). IGI Global Publishing, Hershey (PA), EE.UU. In press, 2010. [Saz *et al.*, 2010a].

- Oscar Saz, Victoria Rodríguez, Eduardo Lleida, W. Ricardo Rodríguez, C. Vaquero. The Use of Multimodal Tools for Pronunciation Training in Second Language Learning of Preadolescents. Language Teaching: Techniques, Developments and Effectiveness (book), Nova Science Publishers, Hauppauge (NY), EE.UU, 2010. [Saz *et al.*, 2010b].

- Carlos Vaquero, Alfonso Ortega, Jesús A. Villalba, Eduardo Lleida. Confidence Measures and Hypothesis Selection Strategies for Speaker Segmentation. V Jornadas de Reconocimiento Biométrico de Personas. Huesca, Spain, 2010. [Saz *et al.*, 2010b].

- Jesús A. Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel. I3A NIST SRE2010 System Description. V Jornadas de Reconocimiento Biométrico de Personas. Huesca, Spain, 2010. [Vaquero *et al.*, 2010b].

- Carlos Vaquero, Alfonso Ortega, and Eduardo Lleida. Intra-Session Variability Compensation and a Hypothesis Generation and Selection Strategy for Speaker Segmentation. In ICASSP, pages 4532–4535, 2011. [Vaquero *et al.*, 2011a].

- Carlos Vaquero, Alfonso Ortega, and Eduardo Lleida. Partitioning of Two-Speaker Conversation Datasets. In Proc Interspeech 2011, pages 385–388, 2011. [Vaquero *et al.*, 2011b].

- Diego Castán, Carlos Vaquero, Alfonso Ortega, David Martínez. Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain. In Proc Interspeech 2011, pages 421–424, 2011. [Castán *et al.*, 2011].

- Carlos Vaquero, Alfonso Ortega, Antonio Miguel and Eduardo Lleida. Quality Assessment of Speaker Diarization for Speaker Characterization. IEEE Transactions on Audio, Speech and Language Processing. Under review, 2011. [Vaquero *et al.*, 2011c].

But the most important proof of the quality of this work is the technology transfer to the industry. In fact, all the work in this thesis is developed with the goal of being useful for real life in the future. Currently, the company *Agnitio S.L.* is providing diarization solutions for speaker verification based on the new approaches presented in this thesis.

9.2.3 Future Lines

Nevertheless and as usual, this thesis has not solved any problem but created new ones. In this section we enumerate the main future lines that could be considered to continue this work.

- **Features for Speaker Factor based Diarization:** In Chapter 2, several features that have shown to be useful for speaker diarization are reviewed. The variability presented by most of these features can be described by a JFA model and thus we can expect them to work with the speaker diarization approach proposed in Chapter 4. In fact, some features that are known to be useful for speaker diarization, as the prosodic features [Shriberg *et al.*, 2005] are known to be hard to model with the JFA paradigm. Recent work [Kockmann *et al.*, 2011] has shown that it is possible to build a PLDA model on these features. This opens the doors to the use of this features with the proposed diarization system. Therefore, it is interesting to study the proposed approach for speaker diarization considering other types of features.

- **Audio Dependent Variability Estimation:** The compensation for intra-session variability proposed in Chapter 5 has been very successful increasing the accuracy of the speaker diarization system. However the proposed approach is based on a linear transformation that assumes that the covariances of the speaker factor vectors for all the speakers are identical. This is not true in general and, even thought the covariances are known to be very similar, it would be interesting to explore techniques to infer the covariances for the speakers present in the recording under analysis, building and adaptive compensation strategy.

- **PLDA for Speaker Clustering:** One of the problems observed in the PLDA based strategy for speaker clustering in large datasets is that it is not possible to take advantage of large clusters (with several recordings) in order to provide robustness to the PLDA. The problem that appears is that the score distribution varies depending on the number of recording sessions available in each cluster. A simple solution to this problem could be the estimation of a single i-vector on the complete cluster, so that the PLDA strategy will see the problem as if there were only one session per cluster. The problem of this strategy is the increased computational cost and that it is difficult to combine it with some stopping criteria, such as those based on the intra-cluster and inter-cluster distributions. Normalization and fusion strategies can also be explored to solve this problem, but there is still a lot of work to do in this line.

- **Stopping Criteria for Speaker Clustering:** Probably the most interesting line in the field of speaker clustering is to find a robust method to determine the number of speakers present in a recording or in a dataset. The best technique among all explored in this thesis is to set a threshold for the clustering metric, which is the stopping criterion traditionally considered in speaker diarization, and it is known to be dependent on the audio signal itself. The solutions based on the intra-cluster and inter-cluster populations seems promising but there is still work to do in this line.

- **Speaker Clustering for Speaker Diarization:** The techniques for speaker clustering in large datasets proposed in Chapter 6 have not been completely successful in the field of speaker diarization, but there is still a lot of work to do. Another interesting line is to study the variability present among the segments to cluster further. In this work the same models considered to compensate for intra-session variability in the speaker factors frame by frame are used to compensate intra-session variability audio segment by audio segment. Usually, the audio segments obtained as output of the diarization system are longer that the one second window considered to estimate the speaker factors frame by frame, and have different lengths. Thus it would be interesting to adapt variability compensation to this new problem in order to improve the accuracy of speaker clustering in speaker diarization.

- **Quality measures:** The quality assessment methodology for speaker diarization proposed in Chapter 7 are designed for two-speaker conversations. An interesting research line is to explore confidence measures that enable us to perform quality assessment for recordings containing more that two speakers. In addition, another interesting line of work is to develop a quality assessment strategy that provides a confidence measures for every speaker in the audio signal, since even when the diarization is incorrect, some of the speakers can be correctly segmented.

- **Overlapped Speech Detection:** In Chapter 7 we have seen that the accuracy of the proposed speaker diarization system is different in two similar databases (NIST SRE 2008 *summed* and NIST SRE 2010 *summed* conditions). It seems that one of the main causes of this difference is the presence of higher overlapped speech in the NIST SRE 2010 than in the NIST SRE 2008 dataset. The detection of overlapped speech is a interesting research line that it is still far to be solved, and would help to improve significantly the accuracy of the proposed diarization system. In fact, the use of inter-speaker variability models could help in the detection of overlapped speech. In addition, it would be interesting to determine the best way to deal with the detected overlapped speech in order to maximize not only the accuracy of the speaker diarization system, but also the performance of the speaker characterization task that will make use of the diarization system output.

- **Computational Cost:** In section 9.1.1 we have observed that the proposed strategies for speaker characterization provide high accuracy at the expense of a high computational cost. Thus, the reduction of the computational cost is a interesting research line that will increase the applicability of these strategies. Several strategies to reduce the computational cost have been suggested in section 9.1.1. These strategies need to be validated and optimized and other approaches can be explored.

- **Online Diarization:** All the approaches for speaker diarization presented in this thesis are designed to work over complete recordings, or over large chunks of audio signal. Ignoring the computational cost, all these approaches can be considered to process a streaming signal chunk by chunk. However, none of them can be used to process an audio signal online. Further research is needed on order to develop online diarization systems based on the proposed approaches.

- **Extension to other domains:** Finally, this thesis is focused on telephone conversations, but there are other domains where the use of speaker diarization is quite interesting. Some of domains for speaker diarization traditionally studied are meetings and broadcast news or parliamentary speeches. The proposed techniques can be easily extended to other domains, but further research is needed in order to keep the high accuracy obtained in the telephone domain when compared to traditional diarization methods.

Part VI

Appendices

A

Baseline Diarization in the *Mono-Mono* Scenario

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	3.05%	3.86%	4.74%	9.45%
$2\% \leq DER < 5\%$	3.87%	4.61%	5.59%	7.29%
$5\% \leq DER < 10\%$	9.27%	6.26%	6.24%	9.71%
$DER \geq 10\%$	7.66%	7.14%	7.74%	11.41%

Table A.1: EER depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario, considering the baseline diarization system.

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	3.06%	3.74%	4.41%	5.06%
$2\% \leq DER < 5\%$	3.74%	5.15%	5.25%	4.19%
$5\% \leq DER < 10\%$	7.67%	5.51%	5.15%	6.51%
$DER \geq 10\%$	5.19%	4.99%	6.14%	6.91%

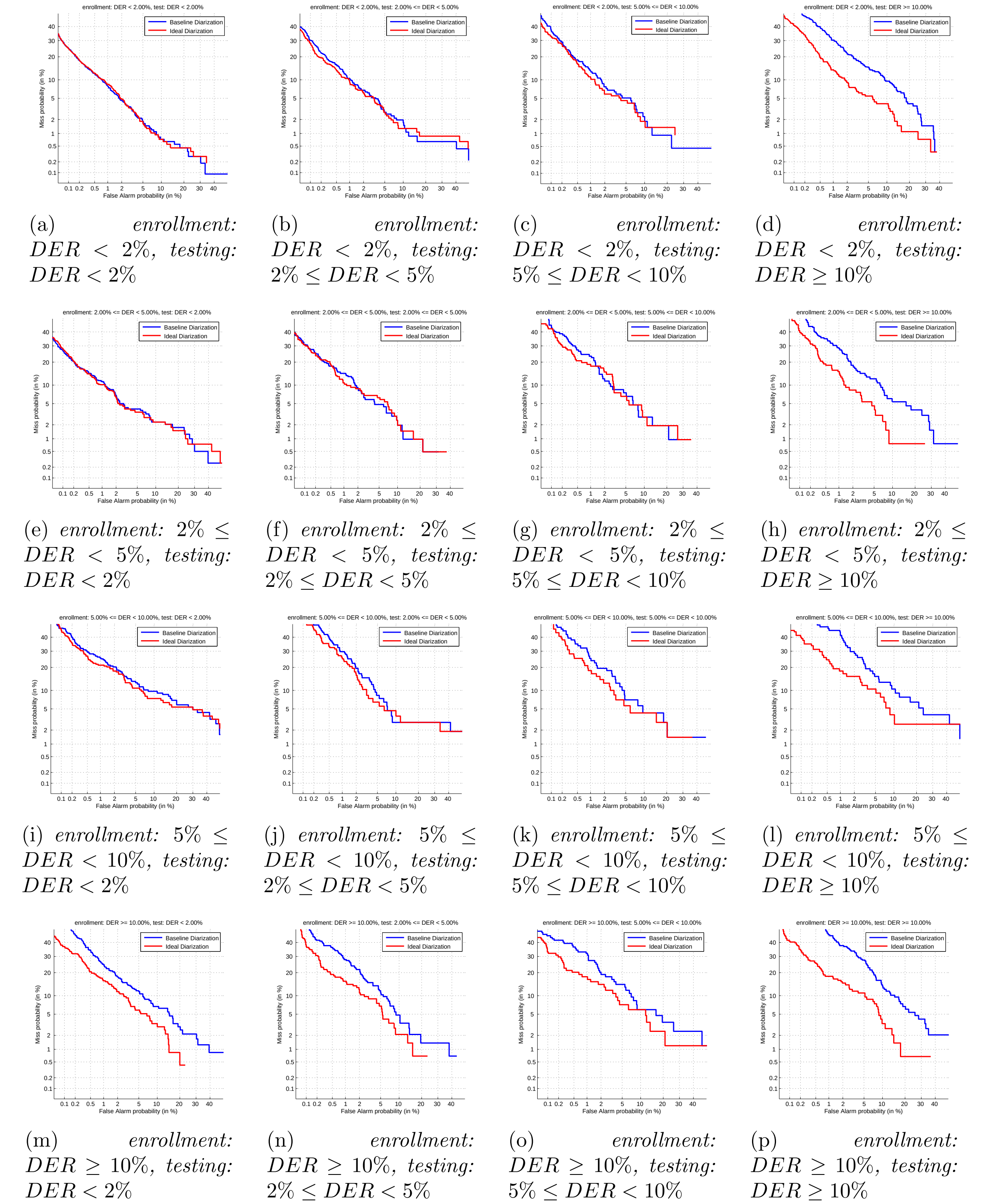
Table A.2: EER depending on the DER obtained by the baseline diarization system for every enrollment and testing subset in the *mono-mono* scenario, considering the ideal diarization system.

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	0.22%	3.09%	7.54%	86.97%
$2\% \leq DER < 5\%$	3.37%	-10.45%	6.50%	74.19%
$5\% \leq DER < 10\%$	20.82%	13.60%	21.07%	49.08%
$DER \geq 10\%$	47.58%	42.98%	26.02%	65.14%

Table A.3: EER degradation introduced by the the baseline diarization system depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario. The degradation is measured with respect to the EER obtained for ideal diarization system for the corresponding subsets.

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	0.1668	0.1924	0.2254	0.3955
$2\% \leq DER < 5\%$	0.1917	0.2217	0.2952	0.3646
$5\% \leq DER < 10\%$	0.3391	0.3824	0.3360	0.4607
$DER \geq 10\%$	0.3365	0.3737	0.3810	0.5486

Table A.4: $min(C_{norm})$ depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario, considering the baseline diarization system.



(a) *enrollment: $DER < 2\%$, testing: $DER < 2\%$*

(b) *enrollment: $DER < 2\%$, testing: $2\% \leq DER < 5\%$*

(c) *enrollment: $DER < 2\%$, testing: $5\% \leq DER < 10\%$*

(d) *enrollment: $DER < 2\%$, testing: $DER \geq 10\%$*

(e) *enrollment: $2\% \leq DER < 5\%$, testing: $DER < 2\%$*

(f) *enrollment: $2\% \leq DER < 5\%$, testing: $2\% \leq DER < 5\%$*

(g) *enrollment: $2\% \leq DER < 5\%$, testing: $5\% \leq DER < 10\%$*

(h) *enrollment: $2\% \leq DER < 5\%$, testing: $DER \geq 10\%$*

(i) *enrollment: $5\% \leq DER < 10\%$, testing: $DER < 2\%$*

(j) *enrollment: $5\% \leq DER < 10\%$, testing: $2\% \leq DER < 5\%$*

(k) *enrollment: $5\% \leq DER < 10\%$, testing: $5\% \leq DER < 10\%$*

(l) *enrollment: $5\% \leq DER < 10\%$, testing: $DER \geq 10\%$*

(m) *enrollment: $DER \geq 10\%$, testing: $DER < 2\%$*

(n) *enrollment: $DER \geq 10\%$, testing: $2\% \leq DER < 5\%$*

(o) *enrollment: $DER \geq 10\%$, testing: $5\% \leq DER < 10\%$*

(p) *enrollment: $DER \geq 10\%$, testing: $DER \geq 10\%$*

Figure A.1: *DET curves considering the baseline and the ideal diarization systems in the mono-mono scenario, for several subsets depending on the DER obtained by the baseline diarization system.*

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	0.1682	0.1727	0.1974	0.2267
$2\% \leq DER < 5\%$	0.1805	0.2005	0.2527	0.2274
$5\% \leq DER < 10\%$	0.2922	0.3414	0.2713	0.2620
$DER \geq 10\%$	0.2491	0.2346	0.2459	0.2629

Table A.5: $min(C_{norm})$ depending on the DER obtained by the baseline diarization system for every enrollment and testing subset in the *mono-mono* scenario, considering the ideal diarization system.

DER considered for Enrollment	DER considered for Testing			
	$DER < 2\%$	$2\% \leq DER < 5\%$	$5\% \leq DER < 10\%$	$DER \geq 10\%$
$DER < 2\%$	-0.81%	11.38%	14.19%	74.51%
$2\% \leq DER < 5\%$	6.18%	10.58%	16.84%	60.32%
$5\% \leq DER < 10\%$	16.06%	12.00%	23.84%	75.84%
$DER \geq 10\%$	35.09%	59.28%	54.92%	108.69%

Table A.6: $min(C_{norm})$ degradation introduced by the baseline diarization system depending on the DER obtained for every enrollment and testing subset in the *mono-mono* scenario. The degradation is measured with respect to the $min(C_{norm})$ obtained for ideal diarization system for the corresponding subsets.

B

An Example of the Speaker Partitioning Problem

Given a set $Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ composed by $N = 4$ i-vectors, and a PLDA model Θ for speaker recognition, a total of $K = B_4 = 15$ hypothetical partitions can be evaluated:

- One partition assuming that there is only a single speaker in the set (the coarsest partition):

$$H_1 \ : \ \mathcal{C}_1(1) = \{\phi_1, \phi_2, \phi_3, \phi_4\} \ : \ \mathcal{L}(H_1|\Theta) = \mathcal{L}(\phi_1, \phi_2, \phi_3, \phi_4|\Theta)$$

- Seven partitions assuming that there are two speakers in the set.

$$H_2 \ : \ \mathcal{C}_2(1) = \{\phi_1, \phi_2, \phi_3\}\,, \mathcal{C}_2(2) = \{\phi_4\} \ : \ \mathcal{L}(H_2|\Theta) = \mathcal{L}(\phi_1, \phi_2, \phi_3|\Theta)\mathcal{L}(\phi_4|\Theta)$$

$$H_3 \ : \ \mathcal{C}_3(1) = \{\phi_1, \phi_2, \phi_4\}\,, \mathcal{C}_3(2) = \{\phi_3\} \ : \ \mathcal{L}(H_3|\Theta) = \mathcal{L}(\phi_1, \phi_2, \phi_4|\Theta)\mathcal{L}(\phi_3|\Theta)$$

$$H_4 \ : \ \mathcal{C}_4(1) = \{\phi_1, \phi_3, \phi_4\}\,, \mathcal{C}_4(2) = \{\phi_2\} \ : \ \mathcal{L}(H_4|\Theta) = \mathcal{L}(\phi_1, \phi_3, \phi_4|\Theta)\mathcal{L}(\phi_2|\Theta)$$

$$H_5 \ : \ \mathcal{C}_5(1) = \{\phi_2, \phi_3, \phi_4\}\,, \mathcal{C}_5(2) = \{\phi_1\} \ : \ \mathcal{L}(H_5|\Theta) = \mathcal{L}(\phi_2, \phi_3, \phi_4|\Theta)\mathcal{L}(\phi_1|\Theta)$$

$$H_6 \ : \ \mathcal{C}_6(1) = \{\phi_1, \phi_2\}\,, \mathcal{C}_6(2) = \{\phi_3, \phi_4\} \ : \ \mathcal{L}(H_6|\Theta) = \mathcal{L}(\phi_1, \phi_2|\Theta)\mathcal{L}(\phi_3, \phi_4|\Theta)$$

$$H_7 \ : \ \mathcal{C}_7(1) = \{\phi_1, \phi_3\}\,, \mathcal{C}_7(2) = \{\phi_2, \phi_4\} \ : \ \mathcal{L}(H_7|\Theta) = \mathcal{L}(\phi_1, \phi_3|\Theta)\mathcal{L}(\phi_2, \phi_4|\Theta)$$

$$H_8 \ : \ \mathcal{C}_8(1) = \{\phi_1, \phi_4\}\,, \mathcal{C}_7(2) = \{\phi_2, \phi_3\} \ : \ \mathcal{L}(H_8|\Theta) = \mathcal{L}(\phi_1, \phi_4|\Theta)\mathcal{L}(\phi_2, \phi_3|\Theta)$$

- Six partitions assuming that there are three speakers in the set.

$$\begin{aligned} H_9 \ &: \ \mathcal{C}_9(1) = \{\phi_1, \phi_2\}\,, \mathcal{C}_9(2) = \{\phi_3\}\,, \mathcal{C}_9(3) = \{\phi_4\} \ : \\ \mathcal{L}(H_9|\Theta) &= \mathcal{L}(\phi_1, \phi_2|\Theta)\mathcal{L}(\phi_3|\Theta)\mathcal{L}(\phi_4|\Theta) \end{aligned}$$

$$\begin{aligned} H_{10} \ &: \ \mathcal{C}_{10}(1) = \{\phi_1, \phi_3\}\,, \mathcal{C}_{10}(2) = \{\phi_2\}\,, \mathcal{C}_{10}(3) = \{\phi_4\} \ : \\ \mathcal{L}(H_{10}|\Theta) &= \mathcal{L}(\phi_1, \phi_3|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_4|\Theta) \end{aligned}$$

$$\begin{aligned} H_{11} \ &: \ \mathcal{C}_{11}(1) = \{\phi_1, \phi_4\}\,, \mathcal{C}_{11}(2) = \{\phi_2\}\,, \mathcal{C}_{11}(3) = \{\phi_3\} \ : \\ \mathcal{L}(H_{11}|\Theta) &= \mathcal{L}(\phi_1, \phi_4|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_3|\Theta) \end{aligned}$$

$$\begin{aligned} H_{12} \ &: \ \mathcal{C}_{12}(1) = \{\phi_1\}\,, \mathcal{C}_{12}(2) = \{\phi_2, \phi_3\}\,, \mathcal{C}_{12}(3) = \{\phi_4\} \ : \\ \mathcal{L}(H_{12}|\Theta) &= \mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_2, \phi_3|\Theta)\mathcal{L}(\phi_4|\Theta) \end{aligned}$$

$$\begin{aligned} H_{13} \ &: \ \mathcal{C}_{13}(1) = \{\phi_1\}\,, \mathcal{C}_{13}(2) = \{\phi_2, \phi_4\}\,, \mathcal{C}_{13}(3) = \{\phi_3\} \ : \\ \mathcal{L}(H_{13}|\Theta) &= \mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_2, \phi_4|\Theta)\mathcal{L}(\phi_3|\Theta) \end{aligned}$$

$$\begin{aligned} H_{14} \ &: \ \mathcal{C}_{14}(1) = \{\phi_1\}\,, \mathcal{C}_{14}(2) = \{\phi_2\}\,, \mathcal{C}_{14}(3) = \{\phi_3, \phi_4\} \ : \\ \mathcal{L}(H_{14}|\Theta) &= \mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_3, \phi_4|\Theta) \end{aligned}$$

- One partition assuming that there are four speakers in the set (the finest partition).

$$\begin{aligned} H_{15} \ &: \ \mathcal{C}_{15}(1) = \{\phi_1\}\,, \mathcal{C}_{15}(2) = \{\phi_2\}\,, \mathcal{C}_{15}(3) = \{\phi_3\}\,, \mathcal{C}_{15}(4) = \{\phi_4\} \ : \\ \mathcal{L}(H_{15}|\Theta) &= \mathcal{L}(\phi_1|\Theta)\mathcal{L}(\phi_2|\Theta)\mathcal{L}(\phi_3|\Theta)\mathcal{L}(\phi_4|\Theta) \end{aligned}$$

Bibliography

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

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

[AMI, 2006] AMI. Augmented multi-party interaction, 2006.

[Anguera, 2005] Xavier Anguera. Xbic: Real-time cross probabilities measure for speaker segmentation. *Technical Report ICSI*, 2005.

[Anguera, 2006] X. Anguera. Robust speaker diarization for meetings, 2006.

[Attias, 2000] Hagai Attias. A variational bayesian framework for graphical models. In *In Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[Auckenthaler *et al.*, 2000] Roland Auckenthaler, Michael Carey, and H Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[Barras *et al.*, 2004] C. Barras, J.-L. Gauvain, S. Meignier, and X. Zhu. Improving speaker diarization. In *RT'O4F Workshop*, USA, 7-10 Nov 2004.

[Barras *et al.*, 2006] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1505–1512, 2006.

[Ben *et al.*, 2004] Mathieu Ben, Michael Betser, Frd́éric Bimbot, and Guillaume Gravier. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. In *in Intl. Conf. on Speech and Language Processing*, 2004.

[Bimbot *et al.*, 2004] F Bimbot, JF Bonastre, C Fredouille, G Gravier, I Magrin-Chagnolleau, S Meignier, T Merlin, Javier Ortega-García, D Petrovska-Delacrétaz, and D A Reynolds. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4):430–451, 2004.

[Bishop, 2006] Christopher Bishop. *Pattern Recognition and Machine Learning.* Springer Science+Business Media, LLC, 2006.

[Boakye *et al.*, 2008] Kofi Boakye, B. Trueba-Hornero, Oriol Vinyals, and Gerald Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *ICASSP'08*, pages 4353–4356, 2008.

[Bonastre *et al.*, 2000] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, and C. Wellekens. A speaker tracking system based on speaker turn detection for nist evaluation. In *Proceedings of the Acoustics, Speech, and Signal Processing, 2000. on IEEE International Conference - Volume 02*, ICASSP '00, pages II1177–II1180, Washington, DC, USA, 2000. IEEE Computer Society.

[Brummer and De Villiers, 2010] Niko; Brummer and Edward De Villiers. The Speaker Partitioning Problem. In *Oddyssey Speaker and Language Recognition Workshop*, 2010.

[Brummer and Dupreez, 2006] N Brummer and J Dupreez. Application-independent evaluation of speaker detection. *Computer Speech Language*, 20(2-3):230–275, 2006.

[Brümmer *et al.*, 2007] Niko Brümmer, Jan Honza Cernocky, and Albert Strasheim. Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. In *IEEE Transactions on Audio, Speech and Signal Processing*, 2007.

[Brummer *et al.*, 2010] Niko Brummer, Lukas Burget, Patrick Kenny, Pavel Matejka, Edward Villiers de, Martin Karafiat, Marcel Kockmann, Ondrej Glembek, Oldrich Plchot, Doris Baum, and Mohammed Senoussauoi. Abc system description for nist sre 2010. In *Proc. NIST 2010 Speaker Recognition Evaluation*, pages 1–20. National Institute of Standards and Technology, 2010.

[Brummer, 2005] Niko Brummer. http://sites.google.com/site/nikobrummer/focalbilinear, 2005.

[Brummer, 2010] Niko Brummer. Bayesian PLDA, 2010.

[Campbell *et al.*, 2006] W M Campbell, D E Sturim, D A Reynolds, and A Solomonoff. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, pages I–97–I–100. Ieee, 2006.

[Castaldo *et al.*, 2008] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair. Stream-based speaker segmentation using speaker factors and eigenvoices. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4133–4136, Las Vegas, Nevada, mar 2008. IEEE.

[Castán *et al.*, 2010] Diego Castán, Alfonso Ortega, Carlos Vaquero, Antonio Miguel, and Eduardo Lleida. VIVOLAB-UZ Audio Segmentation System for Albayzin Evaluation 2010. In *FALA 2010*, pages 437–440, 2010.

[Castán *et al.*, 2011] Diego Castán, Carlos Vaquero, Alfonso Ortega, and David Martínez. Hierarchical audio segmentation with hmm and factor analysis in broadcast news domain. In *Interspeech*, pages 421–424, 2011.

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

[Chen *et al.*, 2002] S. S. Chen, E. Eide, M. J. F. Gales, Ramesh A. Gopinath, D. Kanvesky, and Peder A. Olsen. Automatic transcription of broadcast news. *Speech Communication*, 37:69–87, 2002.

[CHIL, 2006] CHIL. Computers in the human interaction loop, 2006.

[Davis and Mermelstein, 1980] S Davis and P Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28(4):357–366, 1980.

[Dehak *et al.*, 2009] Najim Dehak, Redah Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In *Interspeech 2009*, Brighton, UK, 2009.

[Dehak *et al.*, 2010] N Dehak, P Kenny, R Dehak, P Dumouchel, and P Ouellet. Front-End Factor Analysis For Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2010.

[Delacourt and Wellekens, 2000] Perrine Delacourt and Christian Wellekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech Communication*, 2000.

[Do, 2003] M. N. Do. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *Signal Processing Letters, IEEE*, 10(4):115–118, mar 2003.

[Duda and Hart, 1973] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 1973.

[EARS, 2004] EARS. http://projects.ldc.upenn.edu/ears/, 2004.

[FCC, 2011] FCC. Federal communications comission (fcc) media bureau, http://transition.fcc.gov/mb/, 2011.

[Fox *et al.*, 2009] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 2009.

[Friedland *et al.*, 2009a] Gerald Friedland, Hayley Hung, and Chuohao Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *ICASSP*, pages 4069–4072, 2009.

[Friedland *et al.*, 2009b] Gerald Friedland, Oriol Vinyals, Yan Huang, and Christian Müller. Prosodic and other long-term features for speaker diarization. *IEEE Transactions on Audio, Speech & Language Processing*, 17(5):985–993, 2009.

[Garcia-Romero, 2006] Daniel Garcia-Romero. Using quality measures for multilevel speaker recognition. *Computer Speech*, 20(2-3):192–209, 2006.

[Gauvain *et al.*, 1999] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and MichÃ¨le Jardino. The limsi 1998 hub-4e transcription system. In *IN PROC. OF THE DARPA BROADCAST NEWS WORKSHOP*, pages 99–104, 1999.

[Gish *et al.*, 1991] H. Gish, M. H. Siu, and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 873–876, 1991.

[Gonzalez-Rodriguez *et al.*, 2003] Joaquin Gonzalez-Rodriguez, Marta García-Gomar, Daniel Ramos, and Javier Ortega-García. Robust likelihood ratio estimation in Bayesian forensic speaker recognition. In *Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech 2003)*, pages 693–696, Geneva, 2003.

[Hansen *et al.*, 2005] J. H. L. Hansen, Rongqing Huang, Bowen Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul. SpeechFind: advances in spoken document retrieval for a National Gallery of the Spoken Word. *IEEE Transactions on Speech and Audio Processing*, 13(5):712–730, sep 2005.

[Harriero *et al.*, 2009] Alberto Harriero, Daniel Ramos, Joaquin Gonzalez-Rodriguez, and Julian Fierrez. Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification. In Massimo Tistarelli and Mark Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 434–442. Springer Berlin / Heidelberg, 2009.

[Hermansky and Morgan, 1994] H Hermansky and N Morgan. RASTA processing of speech. *Ieee Transactions On Speech And Audio Processing*, 2(4):578–589, 1994.

[Hermansky, 1990] H Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[Huang and Hansen, 2006] R. Huang and J. H. L. Hansen. Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. *IEEE Trans. Speech and Audio Processing*, 14:907–919, 2006.

[Hung *et al.*, 2000] Jeih-Weih Hung, Hsin-Min Wang, and Lin-Shan Lee. Automatic metric-based speech segmentation for broadcast news via principal component analysis. In *Interspeech*, pages 121–124, 2000.

[Imseng and Friedland, 2009] D. Imseng and G. Friedland. Robust speaker diarization for short speech recordings. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, 2009.

[Imseng and Friedland, 2010] David Imseng and Gerald Friedland. Tuning-robust initialization methods for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2028–2037, 11 2010.

[ITU-D, 2010] ITU-D. International telecommunication union - development (itu-d) ict data and statistics (ids), online: http://www.itu.int/itu-d/ict/statistics/at_glance/keytelecom.html, 2010.

[Jain *et al.*, 2008] A.K. Jain, P.J. Flynn, and A.A. Ross. *Handbook of biometrics*. Springer, 2008.

[Jin *et al.*, 1997] Hubert Jin, Francis Kubala, and Rich Schwartz. Automatic speaker clustering. In *DARPA Speech Recognition Workshop*, pages 108–111, 1997.

[Johnson and Woodland, 1998] S. E. Johnson and P. C. Woodland. Speaker clustering using direct maximisation of the mllr-adapted likelihood. In *PROC. ICSLP 98*, pages 1775–1779, 1998.

[Kajarekar *et al.*, 2009] Sachin S. Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet. The sri nist 2008 speaker recognition evaluation system. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4205–4208, 2009.

[Kazman *et al.*, 1995] R. Kazman, W. Hunt, and M. Mantei. Synamic meeting annotation and indexing. In *Pacific Workhop on Distributed Multimeda Systems*, 1995.

[Kemp *et al.*, 2000] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alex Waibel. Strategies for automatic segmentation of audio data. In *in Proc. ICASSP*, pages 1423–1426, 2000.

[Kenny *et al.*, 2007] P Kenny, G Boulianne, P Ouellet, and Pierre Dumouchel. Speaker and session variability in GMM-based speaker verification. *Ieee Transactions On Audio Speech And Language Processing*, 15(4):1448–1460, 2007.

[Kenny *et al.*, 2008] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel. A Study of Interspeaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988, jul 2008.

[Kenny *et al.*, 2010] P. Kenny, D. Reynolds, and F. Castaldo. Diarization of telephone conversations using factor analysis. *IEEE Journal on Selected Topics in Signal Processing*, 4(6):1059–1070, December 2010.

[Kenny, 2008] Patrick Kenny. Bayesian analysis of speaker diarization with eigenvoice priors, 2008.

[Kenny, 2010] Patrick Kenny. Bayesian Speaker Verification with Heavy-Tailed Priors. In *Oddyssey Speaker and Language Recognition Workshop*, 2010.

[Kinnunen and Li, 2010] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.

[Kockmann *et al.*, 2011] Marcel Kockmann, Luciana Ferrer, Lukas Burget, Elisabeth Shriberg, and Jan Cernocky. Recent progress in prosodic speaker verification. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, pages 4556–4559. IEEE Signal Processing Society, 2011.

[Kubala *et al.*, 1996] Francis Kubala, Hubert Jin, Yspyros Matsoukas, Long Nguyen, Rich Schwartz, and John Makhoul. The 1996 bbn byblos hub-4 transcription system. In *In Proc. of DARPA Speech Recognition Workshop*, pages 90–93, 1996.

[Kuhn *et al.*, 2000] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Audio, Speech, and Language Processing*, 8(6):695–707, November 2000.

[Lapidot *et al.*, 2002] I. Lapidot, Hugo Guterman, and A. Cohen. Unsupervised speaker recognition based on competition between self-organizing maps. *IEEE Transactions on Neural Networks*, 13(4):877 – 887, July 2002.

[Lapidot, 2003] Itshak Lapidot. Som as likelihood estimator for speaker clustering. In *Interspeech*. ISCA, 2003.

[Levinson, 1986] S.E. Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29 – 45, 1986.

[Li and Porter, 1988] K P Li and J E Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 595–598, 1988.

[Liu and Kubala, 1999] Daben Liu and Francis Kubala. Fast speaker change detection for broadcast news transcription and indexing. In *EUROSPEECH*, 1999.

[Löffler *et al.*, 2002] Jobst Löffler, Konstantin Biatov, Christian Eckes, and Joachim Köhler. Ifinder: an mpeg-7-based retrieval system for distributed multimedia content. In *ACM Multimedia Conference*, pages 431–435, 2002.

[Lu and Zhang, 2002] Lie Lu and Hong-Jiang Zhang. Real-time unsupervised speaker change detection. *Pattern Recognition, International Conference on*, 2:20358, 2002.

[Martin *et al.*, 1997] A Martin, G Doddington, T Kamm, M Ordowski, and M Przybocki. The DET curve in assessment of detection task performance. In *Fifth European Conference on Speech Communication and Technology*, volume 97, pages 1895–1898. ISCA, Citeseer, 1997.

[Meignier *et al.*, 2006] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean franÃ§ois Bonastre, and Laurent Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech & Language*, 20:303–330, 2006.

[Nguyen *et al.*, 1998] Trung Hieu Nguyen, Engsiong Chng, and Haizhou Li. Deterministic annealing for clustering, compression, classification, regression and related optimization problems. In *Interspeech*, pages 2210–2239, 1998.

[NIST, 1998] NIST. Nist speaker recognition evaluation, 1998.

[NIST, 2004] NIST. http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/index.html, 2004.

[NIST, 2010a] NIST. Nist speaker recognition evaluation 2008, 2010.

[NIST, 2010b] NIST. Nist speaker recognition evaluation 2010, evaluation plan, 2010.

[NIST, 2010c] NIST. NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig/tests/sre/, 2010.

[Pardo *et al.*, 2007] Jose Pardo, Xavier Anguera, and Chuck Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Trans. Comput.*, 56:1189–1224, September 2007.

[Pelecanos and Sridharan, 2001] Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *Oddyssey Speaker and Language Recognition Workshop*, Crete, Greece, 2001.

[Prince and Elder, 2007] Simon J D Prince and James H Elder. Probabilistic Linear Discriminant Analysis for Inferences About Identity. *IEEE International Conference on Computer Vision*, (iii):1–8, 2007.

[Ramirez *et al.*, 2004] J. Ramirez, J.C. Segura, C. Benitez, A. de La Torre, and A. Rubio. Voice activity detection with noise reduction and long-term spectral divergence estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages ii–1093–6. IEEE, 2004.

[Reynolds and Rose, 1995] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *Ieee Transactions On Speech And Audio Processing*, 3(1):72–83, 1995.

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

[Reynolds *et al.*, 1998] Douglas A. Reynolds, Elliot Singer, Beth A. Carlson, Gerald C. O'Leary, Jack McLaughlin, and Marc A. Zissman. Blind clustering of speech utterances based on speaker and language characteristics. In *ICSLP*, 1998.

[Reynolds *et al.*, 2000] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[Reynolds *et al.*, 2009] Doug Reynolds, Patrick Kenny, and Fabio Castaldo. A Study of New Approaches to Speaker Diarization. In *Interspeech 2009*, Brighton, UK, 2009.

[Reynolds, 1995a] Douglas A Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.

[Reynolds, 1995b] Douglas A Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.

[Rodriguez *et al.*, 2007] W.R. Rodriguez, Carlos Vaquero, O. Saz, and E. Lleida. Aplicación de las tecnologiias del habla al desarrollo del prelenguaje y el lenguaje. In *IFMBE PROCEEDINGS*, volume 18, page 1064. Springer, 2007.

[Rodriguez *et al.*, 2008] W. Rodriguez, O. Saz, E. Lleida, Carlos Vaquero, and A. Escartin. COMUNICA-Tools for Speech and Language Therapy. In *Proceedings of the Workshop on Child, Computer and Interaction*, Chania (Greece), 2008.

[Rodriiguez *et al.*, 2008] W.R. Rodriiguez, Carlos Vaquero, O. Saz, and E. Lleida. Speech technology applied to children with speech disorders. In *Proceedings of the 4th International Conference on Biomedical Engineering*, pages 247–250. Springer, 2008.

[Rose, 2008] Kenneth Rose. T-test distance and clustering criterion for speaker diarization. In *Proceedings of the IEEE*, pages 36–39, 2008.

[Rougui *et al.*, 2006] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez. Hierarchical organization of a set of gaussian mixture speaker models for scaling up indexing and retrieval in audio documents. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC '06, pages 1369–1373, New York, NY, USA, 2006. ACM.

[Saz *et al.*, 2006] Oscar Saz, Carlos Vaquero, Eduardo Lleida, J.-M. Marcos, and C. Canalis. Study of Maximum A Posteriori Speaker Adaptation for Automatic Speech Recognition of Pathological Speech. In *Proceedings of the IV Jornadas en Tecnologias del Habla*, pages 8–11, Zaragoza (Spain), 2006.

[Saz *et al.*, 2008] Oscar Saz, W.-Ricardo Rodriguez, Eduardo Lleida, Carlos Vaquero, and Antonio Escartin. Plataforma para el desarrollo, distribución y evaluación de herramientas logopédicas asistidas por ordenador. In *Proceedings of the V Jornadas en Tecnologias del Habla*, pages 37–40, Bilbao (Spain), 2008.

[Saz *et al.*, 2009a] O. Saz, W.R. Rodriguez, S.C. Yin, E. Lleida, R. Rose, and Carlos Vaquero. Tools and Technologies for Computer-Aided Speech and Language Therapy. *Speech Communication*, 51(10):948–967, 2009.

[Saz *et al.*, 2009b] Oscar Saz, Eduardo Lleida, and Carlos Vaquero. Analysis of Acoustic Features in Speakers with Cognitive Disorders and Speech Impairments. *EURASIP Journal on Advances in Signal Processing*, Special Is, 2009.

[Saz *et al.*, 2009c] Oscar Saz, Victoria Rodriguez, Carlos Vaquero, Eduardo Lleida, and W.-Ricardo Rodriguez. An Experience with a Spanish Second Language Learning Tool in a Multilingual Environment. In *Proceedings of the Workshop on Speech and Language Technology in Education*, Wroxall (UK), 2009.

[Saz *et al.*, 2010a] Oscar Saz, Eduardo Lleida, Victoria Rodriguez, W.-Ricardo Rodriguez, and Carlos Vaquero. The Use of Synthetic Speech in Language Learning Tools: Review and a Case Study. In John Wesley Mullinex and David Stern, editors, *Computer Synthesized Speech: Tools for Aiding Impairment*, chapter 12. Information Science Reference, 2010.

[Saz *et al.*, 2010b] Oscar Saz, Victoria Rodriguez, Eduardo Lleida, W.-Ricardo Rodriguez, and Carlos Vaquero. The Use of Multimodal Tools for Pronunciation Training in Second Language Learning of Preadolescents. In Frank Columbus, editor, *Language Teaching: Techniques, Effectiveness and Developments*. Nova Publishing, Hauppenage, NY (USA), 2010.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[Shriberg *et al.*, 2005] Elizabeth Shriberg, Luciana Ferrer, Sachin S. Kajarekar, Anand Venkataraman, and Andreas Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3-4):455–472, 2005.

[Shum *et al.*, 2011] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass. Exploiting intra-conversation variability for speaker diarization. In *Interspeech*, 2011.

[Siegler *et al.*, 1997] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA Speech Recognition Workshop*, pages 97–99, 1997.

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

[Siu *et al.*, 1992] M.-H. Siu, G. Yu, and H. Gish. An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 2:189–192, 1992.

[Solewicz and Koppel, 2005] Yosef Solewicz and Moshe Koppel. Considering Speech Quality in Speaker Verification Fusion. In *Interspeech 2005*, 2005.

[Solomonoff *et al.*, 1998] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish. Clustering speakers by their voices. In *International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[Stolcke *et al.*, 2010] Andreas Stolcke, Gerald Friedland, and David Imseng. Leveraging speaker diarization for meeting recognition from distant microphones. In *ICASSP*, pages 4390–4393, 2010.

[Sugiyama *et al.*, 1993] M. Sugiyama, J. Murakami, and H. Watanabe. Speech segmentation and clustering based on speaker features. In *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing: speech processing - Volume II*, ICASSP'93, pages 395–398, Washington, DC, USA, 1993. IEEE Computer Society.

[Tranter and Reynolds, 2006] S. E. Tranter and Douglas A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1557–1565, 2006.

[Valente and Wellekens, 2004] Fabio Valente and Christian J Wellekens. Variational bayesian speaker clustering. In *Odyssey'2004, The speaker and language recognition workshop, May 31- June 3, 2004, Toledo, Spain*, 05 2004.

[van Leeuwen, 2010] David van Leeuwen. Speaker linking in large data sets. In *Oddyssey Speaker and Language Recognition Workshop*, 2010.

[Vaquero *et al.*, 2006] Carlos Vaquero, O. Saz, E. Lleida, J.M. Marcos, and C. Canalis. VOCALIZA: An application for computer-aided speech therapy in spanish language. In *Proceedings of the IV jornadas en tecnologias del habla*, pages 321–326, Zaragoza (Spain), 2006.

[Vaquero *et al.*, 2007] Carlos Vaquero, Ó. Saz, and E. Lleida. Tecnologias del habla para el desarrollo del lenguaje. In *Proceedings of the Telecom I+D*, pages 1–8, 2007.

[Vaquero *et al.*, 2008a] Carlos Vaquero, Oscar Saz, and Eduardo Lleida. Human language technologies for speech therapy in spanish language. In *Proceedings of the LangTech*, pages 129–132, Rome (Italy), 2008.

[Vaquero *et al.*, 2008b] Carlos Vaquero, Oscar Saz, Eduardo Lleida, and W.-Ricardo Rodriguez. E-inclusion technologies for the speech handicapped. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4509–4512, Las Vegas, NV (USA), March 2008. Ieee.

[Vaquero *et al.*, 2009] C Vaquero, N Scheffer, and S Kajarekar. Impact of Prior Channel Information for Speaker Identification. In Massimo Tistarelli and Mark Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 443–453. Springer Berlin / Heidelberg, 2009.

[Vaquero *et al.*, 2010a] Carlos Vaquero, Alfonso Ortega, Villalba Jesús, Antonio Miguel, and Eduardo Lleida. Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification. In *Proc Interspeech 2010*, volume 2010, pages 2310–2313, 2010.

[Vaquero *et al.*, 2010b] Carlos Vaquero, Alfonso Ortega, Jesús Villalba, and Eduardo Lleida. Confidence Measures and Hypothesis Selection Strategies for Speaker Segmentation. In *JRBP Conference*, pages 67–74, Huesca, 2010.

[Vaquero *et al.*, 2010c] Carlos Vaquero, Oriol Vinyals, and Gerald Friedland. A Hybrid Approach to Online Speaker Diarization. In *Proc Interspeech 2010*, volume 2010, pages 2638–2641, 2010.

[Vaquero *et al.*, 2011a] Carlos Vaquero, Alfonso Ortega, and Eduardo Lleida. Intra-Session Variablity Compensation and a Hypothesis Generation and Selection Strategy for Speaker Segmentation. In *ICASSP*, pages 4532–4535, 2011.

[Vaquero *et al.*, 2011b] Carlos Vaquero, Alfonso Ortega, and Eduardo Lleida. Partitioning of Two-Speaker Conversation Datasets. In *Proc Interspeech 2011*, volume 2011, pages 385–388, 2011.

[Vaquero *et al.*, 2011c] Carlos Vaquero, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Quality Assessment of Speaker Diarization for Speaker Characterization (Under review). *IEEE Transactions on Audio, Speech and Language Processing*, 2011, 2011.

[Villalba *et al.*, 2008a] Jesus Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, JE Garcia, Luis Buera, and Oscar Saz. Experiencia del I3A en la Evaluación de Reconocimiento de Locutor NIST 2008. In *Proceedings of the Cuartas Jornadas en Reconocimiento Biometrico de Personas*, 2008.

[Villalba *et al.*, 2008b] Jesus Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, JE Garcia, Luis Buera, and Oscar Saz. I3A submission for the NIST SRE 2008. In *NIST SRE 2008*, Montreal, 2008.

[Villalba *et al.*, 2009] Jesús Villalba, Eduardo Lleida, Alfonso Ortega, Carlos Vaquero, and Antonio Miguel. I3A System for Evalita 2009 Speaker Verification Application Evaluation. In *Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, 2009.

[Vogt *et al.*, 2009] Robert J. Vogt, Jason Pelecanos, Nicolas Scheffer, Sachin Kajarekar, and Sridha Sridharan. Within-Session Variability Modelling for Factor Analysis Speaker Verification. In *Proc Interspeech 2009*, volume 2009, pages 1563–1566, 2009.

[Willsky and Jones, 1976] A. Willsky and H. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *Automatic Control, IEEE Transactions on*, 21(1):108–112, 1976.

[Wooters *et al.*, 2004] Chuck Wooters, James Fung, Barbara Peskin, and Xavier Anguera. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In *In RT-04F Workshop*, 2004.

[Yamaguchi *et al.*, 2005] Masahide Yamaguchi, Masaru Yamashita, and Shoichi Matsunaga. Spectral Cross-Correlation Features for Audio Indexing of Broadcast News and Meetings. In *Interspeech 2005*, Lisbon, Portugal, 2005.

[Zhou and Hansen, 2000] Bowen Zhou and John Hansen. Unsupervised audio stream segmentation and clustering via the bayesian information criterion. In *in Proc. ISCLP 2000*, pages 714–717, 2000.

[Zhou and Hansen, 2005] Bowen Zhou and John H. L. Hansen. Efficient audio stream segmentation via the combined t2 statistic and bayesian information criterion. *IEEE Transactions on Audio, Speech and Language Processing*, 13:467–474, 2005.

[Zhu *et al.*, 2005] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain. Combining speaker identification and bic for speaker diarization. In *Interspeech'05, ISCA*, lisbon, sept 2005 2005.

[Zhu *et al.*, 2006] Xuan Zhu, Claude Barras, Lori Lamel, and Jean-Luc Gauvain. Speaker Diarization: From Broadcast News to Lectures. In Steve Renals, Samy Bengio, and Jonathan G. Fiscus, editors, *Machine Learning for Multimodal Interaction*, volume 4299, chapter 35, pages 396–406. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[Zizka *et al.*, 2010] Josef Zizka, Igor Szoke, and Honza Cernocky. http://www.superlectures.com, 2010.