Speaker Diarization and Tracking in Multiple-Sensor Environments

Jordi Luque Serrano

Dissertation presented for the degree of Doctor of Philosophy

TALP Research Center

Department of Signal Theory and Communications

Universitat Politècnica de Catalunya

Barcelona, Spain

Advisor: Prof. Javier Hernando Pericas

October 2012

A mis padres y a mis hermanos.

"Un camino de mil millas comienza con un paso."
– Benjamin Franklin

Abstract

This thesis verses about the research conducted in the topic of speaker recognition in real conditions like as meeting rooms, telephone quality speech and radio and TV broadcast news. The main objective is concerned to the automatic detection and the classification of speakers into a smart-room scenario.

Acoustic speaker recognition is the application of a machine to identify an individual from a spoken sentence. It aims at processing the acoustic signals to convert them in symbolic descriptions corresponding to the identity of the speakers. For the last several years, speaker recognition in real situation has been attracting a substantial research attention becoming one of the spoken language technologies adding quality improvement, or enrichment, of recording transcriptions. In real conditions and particularly, the human activity that takes place in meeting-rooms or class-rooms, compared to other domains exhibits an increased complexity and a challenging problem due to the spontaneity of speech, reverberation effects, the presence of overlapped speech, room setup and channel variability or a rich assortment of acoustic events, either produced by the humans or by objects handled by them. Therefore, the determination of both the identity of speakers and their position in time may help to detect and describe that human activity and to provide machine context awareness.

We first seek to improve traditional modeling approaches for speaker identification and verification, which are based on Gaussian Mixture Models, through multi-decision and multi-channel processing strategies, in smart-room scenario. We put emphasis in studying speaker and channel variability techniques such as Maximum a Posteriori Adaptation, Nuisance Attribute Projection, Joint Factor Analysis, or score normalization; aiming to find out strategies and techniques to deal with such drawback. Moreover, we describe a novel speaker verification algorithm that makes use of adapted features from automatic speech recognition.

In a second line of research, related to speaker detection in continuous audio stream, where the optimum number of speakers of their identities are unknown a priory. We developed and adapted some of the previous speaker recognition techniques to a baseline speaker diarization system based upon Hidden Markov Models and Agglomerative Hierarchical Clustering. We evaluate the application of TDOA feature dynamics and other features in order to improve clustering initialization in the AHC or the detection and handling of speaker overlaps; we assess the impact and synergies with technologies like as Speech Activity Detection and Acoustic Event Detection integrated with the diarization system; and we propose and compare new methods as spectral

clustering. Moreover, the adaptation of the diarization system to broadcast news domain and to the speaker tracking task is also addressed.

Finally, the fusion and combination with video and image modalities is also highlighted across this thesis work, in both speaker identification and tracking approaches. Techniques such as Matching Weighting or Particle Filter are proposed in order to combine scores and likelihoods from different modalities. Results provided demonstrate that these information sources can play also an important role in the automatic person recognition task, adding complementary knowledge to the traditional acoustic spectrum-based recognition systems and thus improving their accuracy. This thesis work was performed in the framework of several international and national projects, among them the CHIL EU project and the Catalan founded project Tecnoparla; and in the participation in technology evaluations such as CLEAR, NIST Rich Transcription (RT), NIST Speaker Recognition Evaluation (SRE) and the Spanish tracking evaluation Albayzin.

Resumen

Esta tesis resume el trabajo realizado en en el área de reconocimiento de hablantes en condiciones reales tales como reuniones en salas, en conversaciones de calidad telefónica y en el dominio de programas de tv y radio. El principal objetivo se centra en la detección automática y clasificación de hablantes en una sala inteligente.

El reconocimiento automático del hablante se define como el uso de una máquina para identificar a un individuo a través de su voz. El objetivo es el procesamiento de la señal acústica para convertirla en descripciones simbólicas que se correspondan con las identidades de los hablantes. Durante los últimos años, el reconocimiento del hablante en situaciones reales ha atraído una sustancial atención de los investigadores convirtiéndose en una de las tecnologías del habla capaz de aportar calidad, o enriquecer, las transcripciones de grabaciones de audio. En condiciones reales y en concreto, la actividad humana que tiene lugar en salas de reuniones o clases docentes, comparada con la de otros dominios exhibe una mayor complejidad y es un problema arduo debido a la espontaneidad del habla, los efectos reververantes, la presencia de solapamientos entre locutores, la configuración de la sala y la varibilidad de canal o la gran cantidad de eventos acústicos, tanto producidos por las personas como por objetos. Es evidente que discernir tanto la identidad del hablante como su posición en tiempo puede ayudar a describir la actividad y proporcionar el conocimiento y percepción de la situacón por parte de la máquina.

En el inicio se busca la mejora de los sistemas tradicionales de modelado para las tareas de identificación y verificación, basados en modelos de mezcla de Gaussianas, a través de estrategias de decisión múltiple y procesamiento multi-canal en salas inteligentes. El estudio se centra en técnicas de variabilidad del hablante y de canal tales como adaptación Maximum a Posteriori, proyecciones Nuisance Attribute, análisis factorial, o normalización de puntuaciones; intentando encontrar estrategias para atacar dicha problemática. Además, se describe un original método para la tarea de verificación del hablante que utiliza características adaptadas a través de un reconocedor automático del habla.

Una segunda línea de investigación se relaciona con la detección automática en audio de múltiples hablantes, dónde tanto su número y sus identidades son desconocidas de antemano. En ella se desarrollan y adaptan

algunas de las anteriores técnicas a un sistema estándard de diarización basado en modelos ocultos de Markov y clustering jerárquico aglomerado de los datos. Evaluamos la aplicación de la dinámica dada por características basadas en retardos entre sensores (TDOA) con intención de mejorar el clustering o la detección y tratamiento de los solapamientos entre hablantes; evaluamos el impacto y las sinergias creadas con tecnologías como la detección del habla y la detección de eventos acústicos, integrándolas con el diarizador y se propone un nuevo método basado en clustering espectral. Además se adapta el sistema de diarización tanto para el procesamiento de programas de radio y televisión como para el seguimiento de locutores específicos.

A lo largo del trabajo se resalta la fusión y combinación con las modalidades de vídeo e imagen, tanto en diarización como en seguimiento de hablantes. Técnicas basadas en ponderación seg ún acierto o en filtros de partículas se proponen para combinar puntuaciones y probabilidades generadas por cada modalidad.

Esta tesis se realizó en el contexto de varios proyectos internacionales y nacionales, entre los que se encuentra el proyecto europeo CHIL y el proyecto Catalán Tecnoparla; y en la participacion en evaluaciones de tecnología como CLEAR, NIST Rich Transcription (RT), NIST Speaker Recognition Evaluation (SRE) y la evaluación española Albayzin en seguimiento.

Acknowledgments

”Es tan grande el placer que se experimenta al encontrar
un hombre agradecido que vale la pena arriesgarse
a no ser un ingrato.”
– Séneca

Mientras que el resto de la tesis se encarga de alumbrar los detalles técnicos, este es el único lugar dónde uno tiene la libertad para narrar la historia personal que hay detrás.

Este camino comienza en el verano de 2004, tiempo por el cual fui adoptado por parte de la gran familia del Departament de Teoria del Senyal i Comunicacions (TSC). Una beca de formación me permitió por aquel entonces realizar mi proyecto final de carrera en Ingeniería de Telecomunición. A la vez, realizaba mis primeras incursiones en las tecnologías del habla al integrar el reconocimento de voz con el procesado de lenguaje natural en un demostrador para el proyecto ALIADO. En el siguiente año 2005, empezé mis estudios de tercer ciclo gracias al proyecto europeo CHIL y a una beca de formación de personal investigador (FPI) asociada al proyecto ACESCA que me fue concedida en el verano de 2006. La mayor parte de esta tesis se ha realizado gracias a la financiación otorgada por parte de tales proyectos.

En este punto es inevitable agradecer a todos aquellos profesores del TSC que de una manera u otra confiaron en mí y creyeron en el sueño de este “protozoo” de científico que sólo le interesaba aprender a investigar aunque a veces parezca que únicamente le encanta programar. José B Mariño, Climent Nadeu, Enric Monte, Josep R. Casas, Josep Ramon Morros, Dusan Macho, Joachim Neumann, Ferran Marqués, Toni Bonafonte, Javier Ruiz Hidalgo, Francesc Vallverdú, Albino Nogueiras, ... – y los que seguramente me dejo – todos me enseñaron algo, me escucharon o me dieron aliento, muchas gracias. Especial mención merece mi director de tesis. Javier Hernando, una persona con una paciencia a prueba de mí, que ha confiado y ha resistido los vaivenes de varios años para ayudarme y llevar a buen puerto esta disertación. Sinceramente, sin él este trabajo no hubiera sido posible. Gracias Kavier.

Inevitable recordar a todos los compañeros y alumnos que por una u otra razón compartieron parte de mi viaje. Compañeros de despacho, administradores de software y hardware y eternos desarrolladores en la sombra

para demostraciones en la sala CHIL, coautores de artículos o compañeros de viajes en defensa de nuestras contribuciones a congresos o evaluaciones. Alberto Abad, Xavier Anguera, Carlos Nistal, Andrey Temko, Pere Pujol, Pablo Agüero, Marta Casar, Jordi Adell, Joan Isaac, Carlos Segura, Cristian Canton, Tanya P. Vitale, Eduardo Mendonça, Martin Zelênak, Martin Wolf, Taras Butko, Mireia Farrús, Maxim Khalilov, Mariella Dimicolli, Marti Umbert, Mateu Aguiló Bosch, Marta Ruiz Costa-jussa, Jaume Gallego,

Incalculable recuerdo y experiencia atesoro por la estancia que realizé en 2010 en el INESC ID's Spoken Language Systems Laboratory (L^2F - Laboratório de sistemas de Língua Falada) en Lisboa. Allí pude aprender de la mano de uno de mis mejores amigos, compañero y colega Alberto Abad. Gracias a todos sus miembros por acogerme y tratarme como uno más del grupo. Isabel Trancoso, Hugo Meinedo, Miguel Bugalho, Thomas Pellegrini, Tiago Luís, Oscar Koller y a todos los que me olvido: *"The Cylons never asked us what we wanted"*.

Quiero agradecer también a todos los compañeros de viaje con los cuales he compartido experiencias de trabajo, me han inspirado y siempre me apoyaron para que la travesía terminara en buen puerto: Alejandro Martín "Truji", Ivan Amat, Elisenda Bonet, Álvaro Pérez y gente de Transmural Biotech, a toda la familía de Telecogresca, a la família Micros y en especial medida a todos los amigos y amigas que me acompañaron, aguantaron mis frustaciones, disfrutaron mis victorias o escucharon mis penas y en definitiva, confiaron en que el libro gordo de Jordete se realizara: Miriam, Alberto, Vane, Adriana, Ana, Xavi, Carla. Todos a vuestra manera me habéis inspirado y como musas tenéis un gran espacio en mi corazón.

No quiero terminar sin agradecer el gran año y medio que he pasado en el Departamento de Matemática Aplicada y Estadística de la Escuela de Aeronáuticos en la UPM. Gracias por adoptarme como uno más de la familia, por no dejarme ganar al ping-pong y por permitirme poner las cenefas que he creído oportunas en la tesis. Miguel Chávez, Ángel M. Núñez, Eusebio Valero, Gonzalo Rubio, Fernando Meseguer, José Olarrea, Carlos Redondo, Javier de Vicente, Ignacio E. Parra, Marta Cordero, Mariola Gómez, M. Teresa – gracias de nuevo. Especial mención merecen mi hermano Bartolomé Luque y Lucas Lacasa por aguantarme mientras realizaba esta disertación – gracias al proyecto MODELICO – y sobretodo, por haberme permitido aprender de ellos.

"Gracias a todos."

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This PhD thesis presents research in the field of speaker recognition in real conditions such as interactive meetings, conferences or TV and radio broadcast news. Speech is the natural human way to communicate ideas, opinion or to express our feelings to others. Automatic speaker identification in a real scenario could bring us a huge range of applications and an improvement of the already existing technologies. Some of them include:

- Speaker identification - *Who said that?*
- Speaker verification - *Is that voice from Anna?*
- Speaker segmentation - *When did he say it?*
- Speaker clustering - *Was it again the same speaker?*
- Speaker diarization - *Who spoke when?*
- Speaker tracking - *Anna spoke when?*

In this PhD thesis we will face these and other tasks, as example of the same general problem - *understand speech*.

Speech processing by computer involves diverse fields such as computer science, speech communication or linguistics. At the same time, this discipline can be divided into several sub areas: speech recognition, speech coding, speech synthesis, speech enhancement and speaker classification. Speech recognition deals with the analysis of the linguistic content of a speech signal. Speech coding and speech enhancement are focused on data compression and on the increasing of the intelligibility of the speech signal respectively. Speech synthesis focus on the creation of artificial voices which usually means computer-generated speech.

Speaker classification is mainly concerned with extracting information about individuals from their voices. This includes gathering different idiosyncratic characteristics such as the gender of the speaker [Peterson and Barney, 1951], the language or dialect of the person who is speaking [Atkinson, 1968] and the age, analysis of their emotional expressions [Hecker *et al.*, 1968], health, the cultural and education level, the social status or even the speakers nationality. The speaker identity can also be determined by the speech signal and such task is known as speaker recognition, which is one of the most widely investigated sub areas in speaker classification [Kersta, 1962], [Doddington, 1971], [Furui, 1996], and the focus of this PhD thesis.

Despite all the research conducted in last decades, a continuous speaker recognition in real conditions is far to be a solved problem. In a real situation which presents a typical multi-speaker environment together with continuous interactions between speakers, it becomes a really hard puzzle. For instance, in a meeting where several people interact with each other to exchange information, the speaker recognition system faces a lot of troubles which degrade the performance of actual speaker recognition algorithms. Non-speech events such as silence, steps, chair moving, laughs, reverberation effects, the mismatch among environment conditions in different scenarios; or speech events in itself as speech overlapping between speakers or simply the accurate detection of the speaker turn, are some examples that can be reeled off from a long list.

We can find two main tasks in the literature related the identification of people across time. The speaker diarization task generally answers to the question *Who spoke when?*. It is performed without any prior knowledge of the identity of the speakers in the audio stream or how many are there. Hence the output of the diarization are labels which identify regions in the recording from the same speaker without take care about his identity, *who is?*. In contrast, speaker tracking task attempts to put a name to such labels, identifying the speakers from a set of known target speakers.

1.1 Motivation and objectives

The tasks of speaker tracking and speaker diarization in continuous audio streams involve several processing stages. The two issues are really so close to each other and generally share some main components. Diarization or tracking of speakers involve various technologies such as audio and speaker segmentation, speaker clustering, speaker identification and speaker verification. The techniques from all of these areas are usually applied together.

We can find a high number of approaches in the literature for the continuous acoustic person identification but two of them deserve a preferential treatment. The first one handles the segmentation and identification in separated steps, and the other one deals with them in an integrated approach. In the *step-by-step* approach, also called sequential systems, Generalized Likelihood Ratio (GLR), Bayesian Information Criterion (BIC) and speaker verification techniques are employed for speaker change detection and clustering distance. In the *integrated* approach, Hidden Markov Models (HMM) are in charge of carry out the clustering in an iterative

strategy, performing the segmentation and the identification at the same time. The latter approach has obtained the best results as reported in literature [Anguera *et al.*, 2011] becoming the most employed method in the task.

The main objective of this PhD thesis is to benchmark and to improve the performance of actual state-of-the-art diarization/tracking systems mainly based on low-level information also knows as spectral features systems. The idea is studying speaker verification/identification techniques to select appropriate characteristics related to the human speech and to take advantage of several information sources like as different sensors and modalities coexisting in the speaker's environment like as multi-microphones or video modality.

In order to test and compare the proposed algorithms and methods with other implementations, they were developed for participation in different national and international speaker recognition and diarization evaluations. Such participation not only allowed the author to receive feedback from other researchers and institutions but also getting access to a huge amount of transcribed audio material in order to strengthen the obtained results.

Several databases have been employed to elaborate this thesis. They cover a wide range of domains, ranging from the meeting and the conference domain to radio and television broadcast news or conversational telephonic speech. In addition, and due the quantity of speech data, the speaker's characteristics are also well represented. Male and female, native and non-native, languages as English, Spanish, Catalan, Portuguese, spontaneous, text-read speech, ... are some examples of them. This great quantity of speech material will allow us to tune different algorithms, being as independent as possible to any room distribution, number of microphones and placement, cellular/land phone type, the style of speech or the speaker characteristics itself.

Finally, a real speaker verification application was developed mainly linked to the participation in the European CHIL project [chil, 2006]. Different systems and algorithms were implemented in the UPC CHIL room laboratory. For this purpose a internal speaker database, around 30 speakers including colleges, students and professors, was recorded twice during different years in order to implement the target speaker models.

In what follows, a set of points which summarizes the objectives and deserve an special attention are listed:

- The speaker diarization system was built using the expertise accumulated at International Computer Science Institute (ICSI) in the research done in broadcast news and at the previous work conducted by Xavier Anguera during his PhD. Such algorithms were taken as starting point for the work of this thesis.

- Different speaker recognition algorithms were adapted or developed from scratch aiming to participate in several national and international evaluations. The Albayzin Spanish evaluation for speaker tracking [Segundo, 2006], the NIST Rich Transcription (RT) evaluations for 2007 and 2009 [Fiscus and et al., 2007a], the CLEAR evaluations [Mostefa and et al., 2006], [Mostefa and et al., 2007] and the NIST Speaker Recognition (SRE) Evaluation [Martin, 2010] were selected in order to benchmark the performance of the technology and implemented algorithms.

- The algorithms were adapted both to specific tasks and the characteristics of speech data. Adaptation to speaker tracking task, algorithms developed to deal with multiple training conditions, meeting and broadcast radio speech are some examples. These works were performed in the framework of several national and international projects. Among them: the Catalan project TECNOPARLA the Spanish project SAPIRE (TEC2007-65470), the European project CHIL (IST-2004-506909) or the Spanish project ALIADO (TIC2002-04447-C02).

- Multichannel information processing and fusion with other cues of information will be also highlighted in several chapters. The use of various audio channels or the information fusion coming from other modalities, like as video and image, will be assessed as a promising future way of improvement in the framework of smart-room environments.

1.2 PhD Thesis Overview

This PhD thesis is split into five main parts related to different topics of speaker recognition. A brief description follows of how it is structured:

In part I, the state of the art in speaker recognition is reviewed. This part includes several sections with a brief description of the recognition systems and their areas of application, the speaker information that can be found in the speech signal and some of the most common automatic speaker recognition techniques. Then follows a review of previously proposed diarization algorithms and implementations. Finally and depending on the multi-sensor capabilities of the environment, the main techniques for fusion of modalities are introduced and reviewed in order to process and to fuse multiple microphones, images, prosody parameters and several cues of information usually present in a smart-room environment.

Part II is divided into two main chapters. Chapter 3, put emphasis on the work developed on speaker identification algorithms in the meeting and conference domain. It pays special attention on the participation in the CLEAR 2006 and 2007 evaluations for the audio and the audio-video speaker recognition categories. The algorithms developed for such evaluations which include audio and video fusion and decision fusion strategies are also reviewed. The following chapter focus on the various speaker verification systems developed for participation on the NIST SRE 2010 Speaker Recognition Evaluation. Chapter 4, gives an overview of such algorithms with especial care on the state-of-the-art speaker recognition techniques related to speaker and channel variability and, as in the previous chapter, on the use of Automatic Speech Recognizer (ASR) technology and prosodic features as new cues of information to gather the speaker identity.

Part III leads the reader through the implementation of diarization and tracking algorithms, their adaptation and their assessment in several audio conditions. An initial review of the ideas behind the system and the implementation of speaker diarization and tracking is followed by an analysis of the differences and needs in

order to adapt it to the different database conditions: meetings or conference domain, broadcast news data or the fusion with other sources of information. As in the previous part, the work was mainly developed in the framework of the participation in technology evaluations: NIST Rich Transcription (RT) evaluation in 2007 and 2009 in chapter 5 or in the JTH Spanish Tracking Challenge for Spanish broadcast news radio, in chapter 6.

In the same chapter 6 speaker diarization algorithms were developed to deal with Catalan TV broadcast news in the framework of the Catalan government founded project Tecnoparla. Chapter 7 reports the work developed in the integration and development of a fully automatic person recognition algorithm. The proposed system makes uses of the combination of information from several audio and video modalities into a smart-room aiming to identify and to track several persons at the same time.

Finally, part IV is devoted to report the conclusions. It summarizes the major contributions and results obtained during the elaboration of this PhD thesis, a review of objectives completion and proposals for future work.

Part I

State of the Art

Chapter 2

Speaker Recognition

In speech processing, recognition usually refers to speech recognition. It tries to determine the linguistic content of an utterance on the basis of information obtained from the speech wave whereas speaker recognition determines the talker's identity. Speaker recognition may be further divided into speaker identification and speaker verification. The aim of a speaker identification system is to identify the person who spoke the utterance from a pool of possible speakers or to identify the speaker as unknown. In contrast to identification, speaker verification is used to authenticate a person's claimed identity. Hence the main difference between identification and verification is the number of decision alternatives. In identification, the number of alternatives is equal to the size of the population whereas in verification the solution becomes in a binary decision: accept or reject the claimed identity.

An specific case is the "open-set" speaker identification which involves representing a given set of speakers using their corresponding statistical model descriptions $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_N$, where N is the number of speakers in the set. If the system is provided with the information that all possible test utterances belong to one of the speakers λ_i, we have a "closed set" of training speakers. If a test utterance may be originating by a person that has not been shown to the system before, it is know as an "open set" of speakers. In this case the system should be able to make a rejection.

For a given test utterance, the process of "open-set" speaker identification can be divided into two successive stages of identification and verification. Firstly, it is required to identify the speaker model in the training set, answering the question: *which is the model that best matches the test utterance?* Secondly, it must be determined, verified, whether the test utterance has actually been produced by the speaker associated with such best-matched model or, in the end, by some unknown speaker outside the speaker's pool.

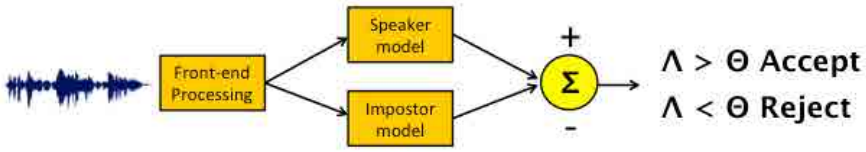


Figure 2.1: *Example of state-of-the-art speaker verification system.*

$$\max_{1\leq i\leq N} \Pr\left(\mathbf{O}|\lambda_i\right) \gtrless \theta \tag{2.1}$$

$$\rightarrow \mathbf{O} \in \begin{cases} \lambda_n, \quad n = \underset{1\leq i\leq N}{\operatorname{argmax}} \Pr\left(\mathbf{O}|\lambda_i\right) \\ \text{unknown speaker model} \end{cases}$$

where $\mathbf{O}$ denotes the feature vector sequence extracted from the test utterance and θ is a pre-determined threshold. In other words, $\mathbf{O}$ is assigned to the speaker model that yields the maximum likelihood over all other speaker models in the set, in the case this maximum likelihood score is greater than the threshold θ. Otherwise, it is declared as uttered by an unknown speaker.

Speaker recognition methods may be also divided into two approaches depending on the application: text dependent and text independent. In the former, the speaker is required to pronounce a specific phrase in both training and recognition steps. When the phrase (or transcription) is known then it is also possible for the system to model learned characteristics of the client's voice, such as speaking rate and accent. The text-dependent methods are usually based on template/model-sequence-matching techniques in which the time axes of an input speech sample and reference templates or reference models of the registered speakers are aligned, and the similarities between them are accumulated from the beginning to the end of the speech. Since a text dependent system models more variability in a person's voice then it is generally more accurate than a text independent system. In the latter, the applications do not rely on any specific spoken message hence a verification phrase is not required. A text independent system is capable of authenticating claimants independently of what is spoken. This approach, therefore, ignores the linguistic variability and it models only the sound of the client's voice which is determined by the physical characteristics of the client's vocal tract. The open-set identification in the text-independent mode is the most challenging class of speaker recognition with various applications including surveillance, and constant authorization control in smart environments and in communications.

There are several applications, such as forensics and surveillance applications, in which predetermined key words cannot be used. Moreover, human beings can recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have attracted more attention. Another advantage of text-independent recognition is that it can be done sequentially, until a desired significance level is reached, without

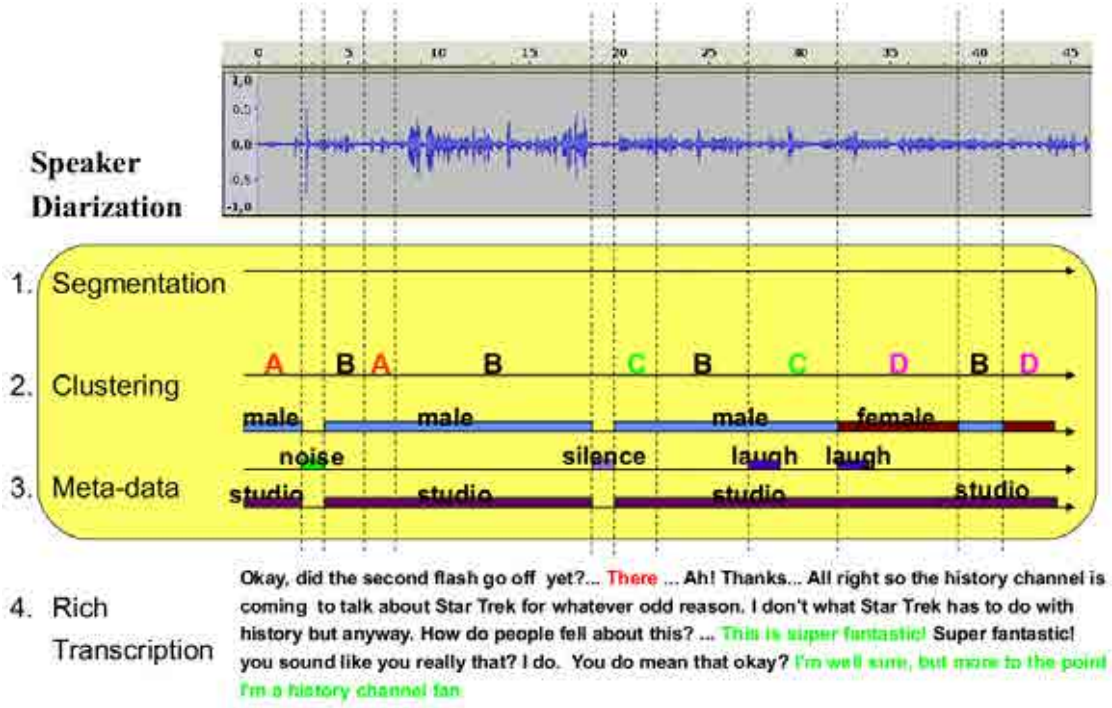


Figure 2.2: *Example of speaker diarization output.*

the annoyance of the speaker having to repeat key words again and again.

In classical speaker identification and verification tasks the systems decide about the speaker's identity by analyzing an excerpt of audio from which is supposed to be just one speaker. This is the usual way for control access applications in which the system forces the claimant to speak an specific utterance or in a two-side telephonic conversation in which each side is provided independently from each other. In such situations the system do not take into account the speaker boundaries or the estimation of the conversation turns.

Nowadays, a huge amount of data is accessible from a wide number of sources like as mass media radio and television, internet, recorded meetings and conferences and so forth. The actual power of computers give us the possibility to analyze and process such data. In any case, to detect and to identify the talkers in a conversation is a real hard challenge. Speaker diarization face such complicated task relying on similar techniques and methods as automatic speech transcription does. Speaker detection and identification in continuous audio streams involves *audio segmentation and classification* in order to discriminate between speech and non-speech events, *speaker segmentation* to arrange the speaker turns and *speaker clustering* aiming to create homogeneous speech segments produced by an unique speaker. In addition, speaker tracking provides a label to the speech segments in order to follow a particular talker among the target speakers from a database. In sharp contrast, speaker diarization do not relies in any prior information about the speakers or how

many of them are in the recording. Both diarization and tracking tasks not only have in common the speaker recognition techniques employed in speaker identification and verification tasks but also those well-know techniques from automatic speech recognizers. The use of speech/non-speech segmentation, bandwidth and gender detectors or segmentation methods as the application of Hidden Markov Models (HMM) are some examples that can be found in the literature.

2.1 Speaker Recognition and Diarization: Applications

Speaker Recognition is becoming a task with a raising interest during last decades among the scientific community. A great amount of work from several disciplines are continuously appearing focusing on this issue. The correct estimation of the person identity is one of the goals from technologies as information indexing and information retrieval, people surveillance, automatic minutes transcription of meetings and , in general, the interaction between computer and humans, the so called context-awareness. The use of audio technologies has found a large acceptance within the scientific community since it is the natural human way of communication and due the actual chance of accessing a wide ocean of audio documents as those coming from radio and television, telephone conversations, meetings, lectures, internet, etc. In addition, this sort of technology has achieved a high performance together with a low grade of intrusion on the human environment.
Such information is of interest for several speech and audio applications [Reynolds and Torres-Carrasquillo, 2005]. For instance, in automatic speech recognition systems the information about *who is speaking?* can be applied for unsupervised speaker adaptation [Anastasakos and et al., 1996], [Matsoukas and et al., 1997] improving the performance of speech recognition in large-vocabulary continuous-speech-recognition (LVCSR) systems [Gauvain and et al., 2002], [Beyerlein and et al., 2002] and [Woodland, 2002]. Other possible use of the speaker labels provided by the speaker identification systems is to aid the transcription task. In this way the transcription is annotated with different label for each speaker. That kind of transcriptions is more readable and useful and could also be used for automatic speaker indexing of audio documents [Makhoul and et al., 2000].
Therefore, nowadays, there is a wide variety and continued growth of applications based on speaker recognition technologies. The most common areas where these applications can be found are listed next.

- **Access control applications** related to secure access to physical and electronic sites are probably the most popular ones. These applications have the advantage that, unlike personal passwords and keys, voices cannot be stolen. However they can be copied by using, for instance, recording devices. In order to protect security systems from this risk, the speaker wishing to access to a secure place is usually asked to pronounce a specific text. In this case, both speaker and the linguistic content of the speech are taken into account.

- **Caller verification in banking and telecommunications** Furthermore to access control, higher levels of verification may be needed for telephone banking in order to achieve more secure transactions. Some

recent applications are also focused on user authentication tasks for remote electronic purchases and both fixed telephone and mobile shopping.

- **Surveillance, law enforcement and forensics** Security agencies have several means of collecting information. One of these is electronic eavesdropping of telephone and radio conversations. As this results in high quantities of data, filter mechanisms must be applied in order to find the relevant information. One of these filters may be the recognition of target speakers that are of interest for the service. Law enforcement includes several applications such as home-parole monitoring, where parolees are called in order to check that they are staying at homeprison call monitoring, border controls, etc. Forensic applications are highly related to law enforcement applications, especially those concerning location of missing people and criminal identification.

- **Speech data management** The use of speaker recognition is also incipient in several applications such as voice mail browsing or intelligent answering machines, where incoming voices are labeled with the speaker name. Speech data management can also be found in smart rooms to automatically track who said what, for example, in a boardroom meeting in order to produce rich transcription and minutes.

- **Speaker indexing and information retrieval** Speaker diarization allows searching for words spoken by a speaker or aiding speaker adaptation techniques for a speech recognition system. Sources in an audio document may also be non-speech events like music, commercials, noise, etc, where diarization could help find the structure of a broadcast program detecting the presence of music, locating commercial to eliminate unwanted audio or be used by speech recognition systems to skip sections for faster processing. As illustrated in the examples, the output audio annotations from diarization may be used directly for applications or as input to assist some downstream human language technology system.

- **Speech transcription** Audio diarization is a useful preprocessing step for an automatic speech transcription system. By separating out speech and non speech segments, the recognizer only needs to process audio segments containing speech, thus reducing the computation time and avoiding word insertions in these portions.

- **Speaker adaptation** Clustering segments of the same acoustic nature, condition-specific models can be used to improve the quality of the transcription. By clustering segments from the same speaker, the amount of data available for unsupervised speaker adaptation is increased, which can significantly improve the transcription performance.

- **Personalization** More and more, a wide variety of devices and smart systems can be found to organize and facilitate or daily life. Presumably, those devices controlled by voice will perform better with a good personal customization. Furthermore, there is also an incipient interest in using speaker characterization in order to provide personal information to be used in advertisements or other services.

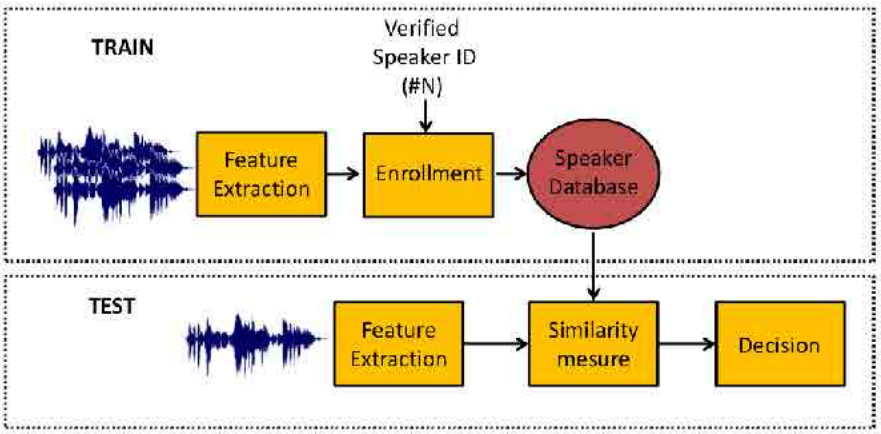


Figure 2.3: *General speaker recognition scheme composed by two main phases: training and testing. Upper, train phase in a common speaker recognition application. Bottom, the template matching corresponding to the testing stage.*

2.2 Speaker Identification and Verification

Automatic speaker recognition is the use of a computer to identify an individual from a spoken utterance. As human beings, we are able to recognize someone just by hearing him or her talk. Usually, a few seconds of speech are sufficient to identify a familiar voice. The idea to teach computers how to recognize humans by the sound of their voices is quite evident, as there are several fruitful applications of this task as mentioned in previous section.

In a human speaker recognition process, the better one knows a person, the easier is to identify others by their speech. Like humans, automatic speaker recognition systems need a training period to learn how this speech is. For this purpose some data from the target speaker has to be collected and, like humans, as more data to create the speaker template more accurate is the recognition. The figure 2.3 presents the general scheme applied in speaker recognition. It involves two main stages: the *training* or *enrollment* of the speaker models and the *testing* phase in which a decision about the speech's identity is taken straight afterwards matching the voice against the speaker database.

The pattern-matching task of speaker recognition involves computing a match score which is a measure of the similarity between the input feature vectors and some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is generated and stored. Then, in order to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user.

One of the central questions addressed by this field is what is it in the speech signal that conveys speaker identity. Traditionally, automatic speaker recognition systems have relied mostly on short-term features as the wide-used Mel Frequency Cepstrum Coefficients (MFCC) or Linear Prediction Coefficients (LPC) all of them related to the spectrum of the voice. However, human speaker recognition relies on other sources

of information; therefore, there is reason to believe that these sources can play also an important role in the automatic speaker recognition task, adding complementary knowledge to the traditional spectrum-based recognition systems and thus improving their accuracy. The first box in the figure 2.3, *Feature Extraction*, both in training and testing, represents the extraction of such characteristics from the speech wave. The extracted features are modeled during the enrollment stage, for example, by means statistical methods such as Gaussian Mixture Models (GMM) [Reynolds, 1995] in order to create a speaker template/model and to update the speaker database.
In the testing stage, the prior trained models are matched against the observations (features) extracted from the test utterance and a similarity/likelihood measure is obtained expressed as in equation 2.2.

$$L = \Pr\left(\lambda_i|\mathbf{O}\right) \tag{2.2}$$

Speaker Identification

In speaker identification applications, given a set of S speakers $\{s_1, s_2, ..., s_N\}$, and a set of models $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, the aim is to find the speaker whose model is assigned to the maximum likelihood:

$$\hat{L} = \underset{1\leq i\leq N}{\operatorname{argmax}} L \tag{2.3}$$

Bayes rule can be applied to statistical models combining equations 2.2 and 2.3 obtaining the following expression:

$$\hat{L} = \underset{1\leq i\leq N}{\operatorname{argmax}} \frac{\Pr(\mathbf{O}|\lambda_i)\Pr(\lambda_i)}{\Pr(\mathbf{O})} \tag{2.4}$$

Due $\Pr(\mathbf{O})$, the probability of the feature $\mathbf{O}$ coming from any speaker, remains constant for all the speakers, the maximization is not affected by this probability. Moreover, all the speakers are assumed to be equally possible, $\Pr(\lambda_i) = \Pr(\lambda_j) \quad \forall i, j = 1, \ldots, N$, it means the prior probabilities for each of the speakers are assumed to be equal. Therefore and assuming equal prior probabilities per speaker, the decision rule can be simplified as follows:

$$\hat{L} = \underset{1\leq i\leq N}{\operatorname{argmax}} \Pr(\mathbf{O}|\lambda_i) \tag{2.5}$$

where the probability $\Pr(\mathbf{O}|\lambda_i)$ or likelihood depends on the statistical technique selected to model the observations. The figure 2.4 summarizes this procedure. Once the feature extraction is carried out, a one-to-one comparison between the observed features $\mathbf{O}$ and the set of speakers models Λ is performed. Maximum likelihood criterion as in 2.5 is taken into account to pick out the speaker with highest chance, in the case of "close-set". In the "open-set" scenario, the final selection depends upon a threshold which decides whereas the speech can be attributed to any of the target speakers in the database or to an unknown speaker.
Finally, the likelihood of a sequence of independent samples $\mathbf{O} = \{o_j\}$ is given by $\prod_j \Pr(o_j)$. The speaker to

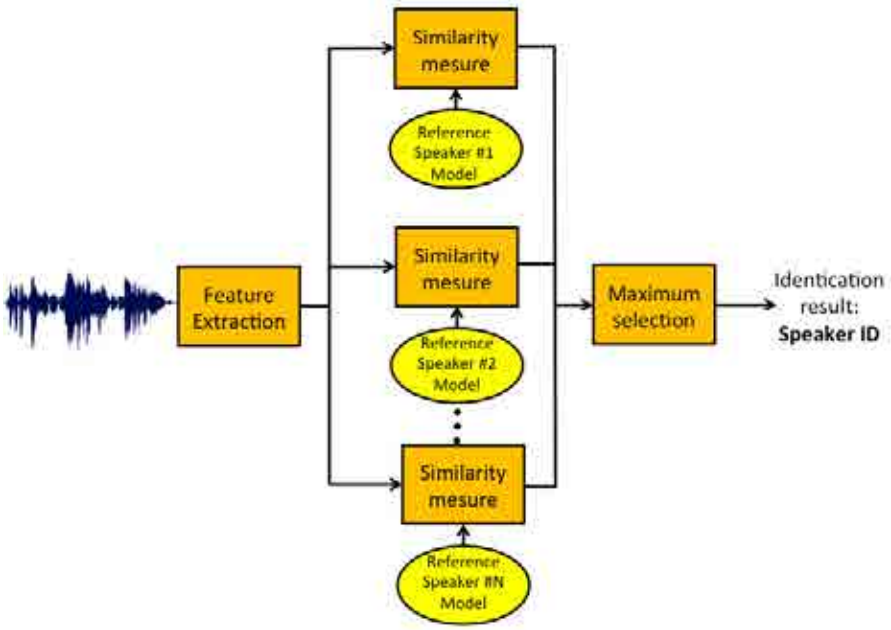


Figure 2.4: *General speaker identification scheme as in [Furui, 1996]*

choose will simply depend on which speaker has the highest likelihood. This will be the most likely speaker given the observed feature, and is known to result in the minimum error strategy [Duda and Hart, 1973]. The work in [Schwart *et al.*, 1982] were first to apply this statistical modeling to the speaker identification task.

Speaker Verification

The speaker verification procedure, as far as a user of such a system is concerned, is similar to that of a password entry system. The user, from here to be known as the claimant, claims the identity of a client, someone authorized to access the system. The verification system requests the claimant give a sample of speech into a microphone. The speech may be prompted or it may be a predefined verification phrase. The system then processes the recorded speech and compares it to a model of the client's voice stored in its database. If it matches then the claimant is authenticated. The length of the verification phrase may affect the system accuracy. Usually, longer verification phrases yield more accurate systems. Depending upon the type of security, the system may request another sample of speech if the match is borderline. For extra security, the system may also pass the recorded speech to a speech recognizer to ensure that the correct response is given. In a system where speech is prompted, this extra security measure prevents the use of recordings.
The speaker verification task, also referred to as detection, mainly consists in to assess whether the test segment **O** was spoken by a hypothesized speaker S usually with the assumption that **O** contains speech from only one

speaker. The single-speaker detection task can be stated as a basic hypothesis test between two hypothesis [Bimbot *et al.*, 2004]:

- H_0: **O** was uttered by the hypothesized speaker S
- H_1: **O** was **not** uttered from the hypothesized speaker S

In order to decide between these two hypotheses a threshold Θ is introduced as a confidence measure on the likelihood values of the hypothesis,

$$\begin{cases} \Pr(\mathbf{O}|H_0) \geq \Theta, & \text{accept } H_0 \\ \Pr(\mathbf{O}|H_1) \geq \Theta, & \text{accept } H_1 \end{cases}$$

where $\Pr(\mathbf{O}|H_0)$ is the probability density function for the hypothesis H_0 evaluated for the observed speech segment **O**, also referred to as the likelihood of the hypothesis H_0 given the speech segment. The likelihood function for H_1 is likewise $\Pr(\mathbf{O}|H_1)$. The decision threshold for accepting or rejecting H_0, Θ, is estimated depending upon the application. Θ controls the trade-off between false detections (Type I errors) and false alarms (Type II errors) in the system hence the value of Θ is estimated depending on the error cost we consider for a given application.
The equation 2.6 can be written as a likelihood ratio (LR) between these two hypotheses, see equation 2.6, "ratio" since it is usually computed in the logarithm domain,

$$\Pr(\mathbf{O}|H_0) - \Pr(\mathbf{O}|H_1) \geq \Theta, \quad \text{accept } H_0 \tag{2.6}$$

The model for H_0 is well defined and is estimated using training speech from S. However the model for H_1 is less well defined since it potentially should represent the entire space of possible alternatives to the hypothesized speaker. From the area of speaker recognition, two main approaches have been taken for this alternative hypothesis modeling. The first approach is to use a set of other speaker models to cover the space of the alternative hypothesis. In various contexts, this set of other speakers has been called cohorts models or background speakers [Auckenthaler *et al.*, 2000] [Zheng *et al.*, 2005]. Given a set of N background speaker models $\{\lambda_1, \ldots, \lambda_N\}$, the alternative hypothesis model is represented by

$$\Pr(\mathbf{O}|\lambda_p) = \Psi\big(\Pr(\mathbf{O}|\lambda_1), \ldots, \Pr(\mathbf{O}|\lambda_N)\big) \tag{2.7}$$

where $\Psi()$ is some function, such as average or maximum, of the likelihood values from the background speaker set. The selection, size, and combination of the background speakers has been the subject of much research [Zheng *et al.*, 2005], [Reynolds, 1997], [Auckenthaler *et al.*, 2000]. In general, it has been found that to obtain the best performance with this approach requires the use of speaker-specific background speaker sets.

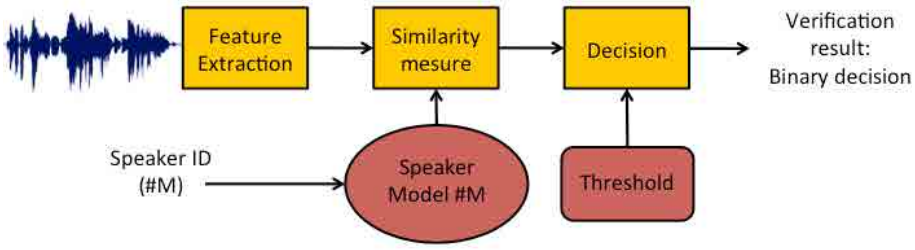


Figure 2.5: *General speaker verification scheme.*

This can be a drawback in an applications using a large number of hypothesized speakers, each requiring their own background speaker set.

$$\max_{1\leq i\leq N} \Pr\left(\mathbf{O}|\lambda_i\right) \gtrless \Theta \tag{2.8}$$

$$\rightarrow \mathbf{O} \in \begin{cases} \lambda_n, \quad n = \underset{1\leq i\leq N}{\operatorname{argmax}} \Pr\left(\mathbf{O}|\lambda_i\right) \\ \text{unknown speaker model} \end{cases}$$

The second major approach to alternative hypothesis modeling is to pool speech from several speakers and train a single model which represents the whole speaker space of characteristics. Such model is known as the Universal Background Model (UBM) in the literature [Reynolds, 1997]. The UBM is trained given a collection of speech samples from a large number of speakers, usually some hundreds, representative of the population expected during verification. The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task.
As conclusion, the output of a verification system is a binary decision: to accept or not to accept the claimed identity, that is the key difference with identification systems.

2.2.1 Speech Features

Next sections introduce to the reader a brief description of the speech production system, the speech signal parametrization and its modeling to compress and analyze the useful parts which convey speaker characteristic information from the speech waveform. Finally, some of the state-of-the-art learning algorithms applied in automatic speaker recognition systems are also reviewed.

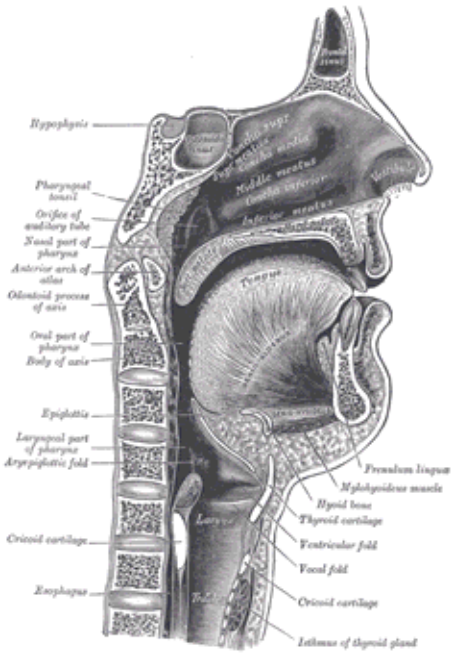


Figure 2.6: *Sagittal section of nose, mouth, pharynx, and larynx. From Gray's Anatomy 1918 edition.* http://commons.wikimedia.org/wiki/File:Sagittalmouth.png

Speech Production

The physical and learned are the two main sources of speaker-specific characteristics. The vocal tract shape is an important physical distinguishing factor of speech. The vocal tract is generally considered as the speech production organ above the vocal folds. As shown in figure 2.6, this includes the following: laryngeal pharynx, oral pharynx, oral cavity, nasal pharynx, and the nasal cavity.

Speech is considered as the response of a slow time varying system. There are two main types of excitations: periodic and noise like which simplify the modeling, but in more detail, excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these. The speech production mechanism consists of a series of pressure changes in acoustic tube, the vocal tract that is excited to generate the desired sound. The excitation is generated by airflow from the lungs, carried by the trachea (also called the wind pipe) through the vocal folds. The vocal folds (formerly known as vocal cords) are shown in figure 2.6. The larynx is composed of the vocal folds, the top of the cricoid cartilage, the arytenoid cartilages, and the thyroid cartilage, also known as "Adams apple". The vocal cords constrict the path from the lungs to the vocal tract. As the lung pressure is increased, air flows out of the lungs and through the opening between vocal cords. The area between the vocal folds is called the glottis. If tension in vocal cords is properly adjusted, the reduced pressure allows the cords to come together, thereby completely constricting the airflow. As a result, pressure increases behind vocal cords, this pressure force it to open and allow the air to pass. Again the air

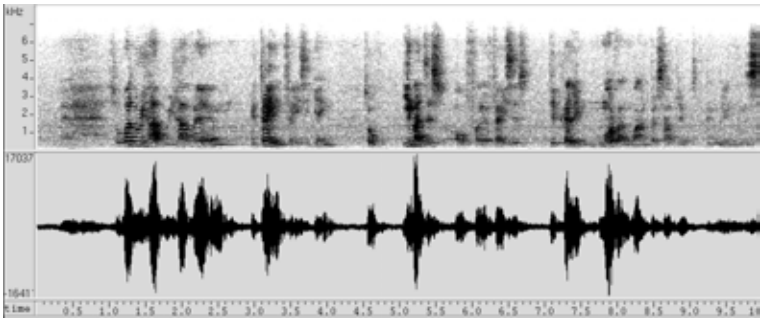


Figure 2.7: *A speech waveform and its corresponding spectogram.*

pressure in the glottis falls and the cycle is repeated. As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by its resonances. Vocal tract resonances are called formants. Thus, the vocal tract shape can be estimated from the spectral shape of the voice signal. Automatic recognition systems typically use features derived only from the vocal tract. An adult male vocal tract is approximately 17 cm long [Flanagan *et al.*, 2008] and the fundamental frequency of general female voice is roughly 225 Hz, male voice is 120 Hz and a small childs voice is around 300 Hz [Kent and Read, 1992]. The physical size and shape of speakers vocal tract determine the range of sounds that can be produced by humans and since each person has their own vocal characteristics there is reason to believe that an individual can be uniquely identified by voice alone [Wolf, 1972], [Rabiner and Schafer, 1978].

There are three main representations of the speech signal: The oscillogram or waveform which represents air pressure changes in a speech wave as a function of time, the spectrum which plots amplitude against frequency and the spectrogram a 3-dimensionally plot of amplitude, frequency and time. The figure 2.7 shows the waveform and the spectrogram corresponding to a 10 seconds sentence uttered by a male speaker. Note that, in the spectrogram, amplitude is represented as a third dimension by dark shades.

The speech production system is usually modeled as the response of linear time varying system (vocal tract) with properly excitation, see scheme in figure 2.8. Speech production is described as two separate and independent processes: the sound generation in the larynx (source) on the one hand and the acoustic filtering of the speech sounds in the vocal tract (filter). The human vocal mechanism is driven by such excitation source, which also contains speaker-dependent information. This representation of the speech production is known as the source-filter approach. If the vocal cords are tensed then voiced sound like vowels are produced by vibration and modulation of the air flow. In case of unvoiced sounds, vocal cords are spread apart and one or two conditions are possible. Either a turbulent flow is produced or a brief transient excitation occurs.

The discrete linear speech production model introduced by [Flanagan *et al.*, 2008], assumes that, for a voiced speech $s(n)$ generated in the larynx, the source (excitation) is a periodic delta train. Unvoiced sounds are represented as noise, see figure 2.8, hence for an unvoiced speech $s(n)$, the source is represented by a random white noise. The speech signal $s(n)$ is filtered through $V(z)$ which represents vocal tract, glottal and lips linear filtering. A common estimation of the signal $s(n)$ is the all-pole linear prediction model, a linear combination

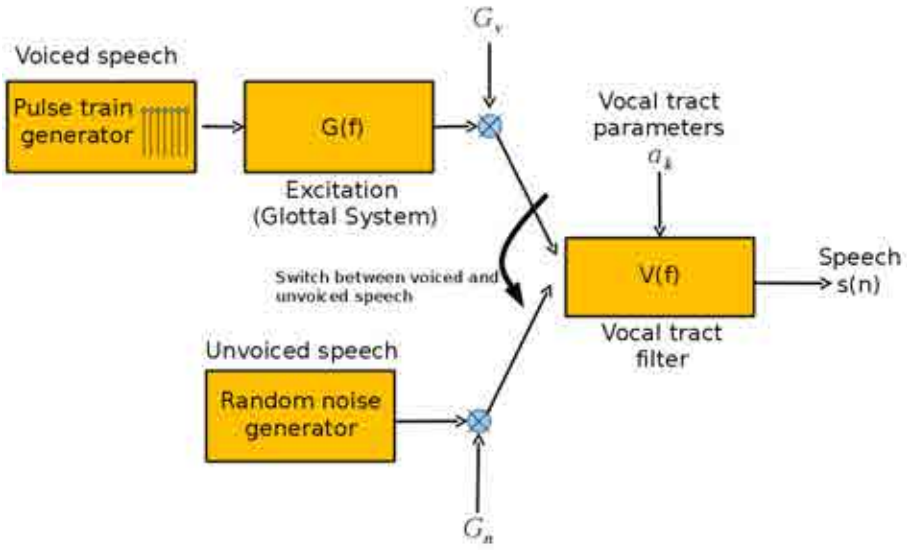


Figure 2.8: *In the linear acoustics model of speech production [Flanagan et al., 2008], the speech signal is produced by filtering an excitation signal (produced in the subglottal system) with a time-varying linear filter (the vocal tract). It should be noted that this model is not valid for all classes of speech sounds, such as frication, where excitation occurs above the glottis. The vocal tract parameters a_k are linear coefficients and they may be estimated by linear prediction*

of its past values and a scaled present input [Picone, 1993],

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gu(n), \tag{2.9}$$

where $s(n)$ is the present output, p is the prediction order, a_k are the model parameters called the predictor coefficients (PCs), $s(nk)$ are past outputs, G is a gain scaling factor, and $u(n)$ is the present input. This modeling yields to the LP transfer function,

$$H(z) = V(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-1}}. \tag{2.10}$$

LP analysis determines the PCs that minimize the prediction error $\hat{s}(n) - s(n)$. Speech features are constructed from the speech model parameters a_k reported in equation 2.9. These LP coefficients are typically non linearly transformed into perceptually meaningful domains suited to the application.

Human voice is characterized by a high degree of variability within the same speaker known as intra-speaker variability. Different emotional states of speakers, colds, time of the day, age, etc., or other external factors such as environmental noise, type of microphone, channel distortion and so forth, make that two speech signals

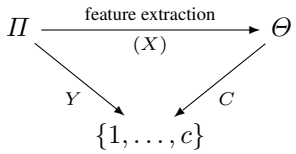
uttered by the same speaker are rarely equal, even when the speaker tries to make them identical. One of the milestones of the speaker recognition technologies is to find those features that can properly characterize a single speaker; i.e. those features whose intra-speaker variability is smaller whereas it presents a high variability between speakers, also known as inter-speaker variability.

Feature Extraction

Most of the parametrization applied in speaker recognition are well known in the speech recognition field [Furui, 2005]. Likewise automatic speech transcription (ASR) the use of parameter based on short-term, frame-based cepstral coefficients, as the well known Mel Frequency Cepstral Coefficients (MFCC) [Furui, 1981], has become the most popular feature domain for signal representation in speaker recognition technologies ,[Reynolds, 1994],[Campbell, 1997], [Kurematsu *et al.*, 2005]. Latest NIST Rich Transcription [Fiscus and et al., 2007a], [Fiscus and et al., 2009a] and NIST Speaker Recognition Evaluations [Martin, 2010] show that most of the presented systems applied MFCC both in speaker diarization and speaker identification/verification tasks. Nevertheless, other kind of parameters such as Frequency Filtering (FF) [Nadeu *et al.*, 2001], Linear Frequency Cepstral Coefficients (LFCC), Perceptual Linear Predictive (PLP), Linear Predictive Coding (LPC) have their small place in the literature.
In some classification problem we need to define

- **Elements**, the things we want to classify
- **Classes**, the labels we want to give at each element
- **Descriptors or features**, how an element is represented in our system

Let suppose there is a correct classification, i.e. a function which ties one class to each element, therefore a learning algorithm associates a class to a list of descriptors in order to fit the classification function defined previously. The following diagram gives a general overview of the learning process,



where Π is the element population, Θ is the feature/descriptor domain and the set {1,...,c} represents the classes (speakers). The function $X : \Pi \rightarrow \Theta$ associates one feature vector to each element and it represents the feature extraction procedure. The Θ space is composed of several dimensions, each of them characterizes one attribute A, logical or numerical. In the speech processing task, $a_0, a_1, \ldots$ are usually the coefficients of the acoustic vectors, the cepstral coefficients for example. From the user point of view, the function $Y : \Pi \rightarrow \{1, \ldots, c\}$ is the classification function and the function $C : \Theta \rightarrow \{1, \ldots, c\}$ is the real classification in the system space. The aim of automatic learning is to find a function C with $C \circ X = Y$, in

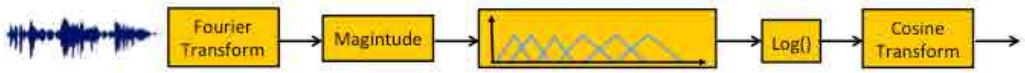


Figure 2.9: *Feature extraction scheme for the Mel-Frequency Cepstral Coefficients (MFCC)*

fact, what we want to find is C with $C \circ X$ being a good approximation of Y. Hence in order to describe one element into the Π space we need to give a value associated to an element in the feature space.
In speech processing the mapping of the population Π (speech domain) to the descriptor domain Θ is known as feature extraction. The speech signal is converted into an electrical signal by a microphone. The obtained electrical signal is then sampled and quantized by an A/D which converts the captured analogical signal from a sound pressure wave to a digital signal. Usually follows a digital filtering that emphasizes important frequency components. It is most often executed using a Finite Impulse Response (FIR) filter of one coefficient digital filter, known as a pre-emphasis filter:

$$H_{pre}(z) = 1 + a_{\text{pre}} z^{-1} \tag{2.11}$$

The pre-emphasis filter enhances the signal spectrum approximately 20 dB per decade. There are two common explanations of the advantages of using this filter [Picone, 1993]. First, voiced sections of the speech signal naturally have a negative spectral slope (attenuation) of approximately 20 dB per decade due to physiological characteristics of the speech production system. An alternate explanation is that hearing is more sensitive above the 1 kHz region of the spectrum and the pre-emphasis filter amplifies this area.
The short-term frame-based approach applied in current speaker recognition systems consists in a segmentation of the digital speech signal into regular segments. The speech signal is usually chopped in frames of $20-30$ms at a rate of $10-20$ms. The frames are commonly weighted by a Hamming window. Values in this range represent a trade-off between the rate of change of spectrum and system complexity. The proper frame duration is ultimately dependent on the velocity of the articulators in the speech production system (rate of change of the vocal tract shape). While some speech sounds (such as stop consonants or diphthongs) exhibit sharp spectral transitions which can result in spectral peaks shifting as much as 80 Hz/ms [Picone, 1993], frame durations less than approximately 8 ms are normally not used. Equally important, however, is the interval over which the power is computed. The number of samples used to compute the summation N , is known as the window duration (in samples). Window duration controls the amount of averaging, or smoothing, used in the power calculation. The frame duration and window duration together control the rate at which the power values track the dynamics of the signal. Finally, a power spectrum estimation is computed in each segment, and speech parameters are extracted. This is the general process to compute short-term frame-based spectral coefficients. In addition, MFCC coefficients, also known as mel-warped features, are based on the non-linear perception of the frequency of sounds and they can be computed as follows:

- Window the signal, usually a Hamming window is applied with overlapping.

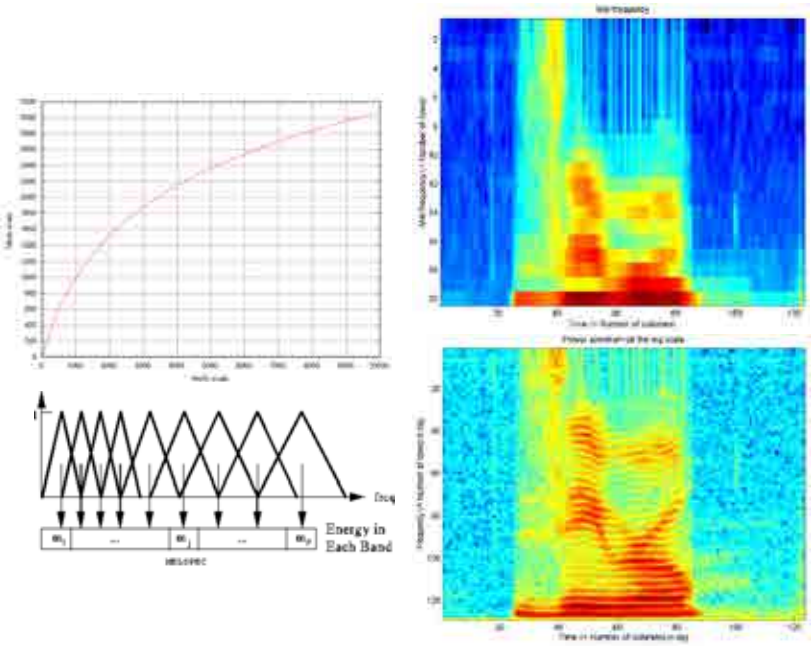


Figure 2.10: *At the top left, plots of pitch mels versus hertz.* $f_{mel} = 2595 \log_{10}(1 + \frac{f}{700})$. *Bottom, overlapped triangular filter bank. At the top right, power spectrum without Mel-frequency wrapping. Bottom, Mel-frequency wrapping of power spectrum. Considering the full image with the mel frequency wrapping set, there is less information than the one without the mel frequency. But instead if we looking in details,we see that the image with the mel frequency wrapping keeps the low frequencies and removes some information. To summarize,the Mel frequency wrapping set allows us to keep only the part of useful information.*

- Compute de Discrete Fourier Transform (DFT)
- Take the square-magnitude
- Warp the frequencies (Apply mel-scale by means a filter bank)
- Take the logarithm (compression) of the filter bank energies
- Compute the Discrete Cosine Transform (DCT) of the log filter-bank energies

The mel scale [Stevens, 1937] attempts to map the perceived frequency of a tone, or pitch, onto a linear scale. This scale is displayed in figure 2.10. The mel warping transforms the frequency scale to place less emphasis on high frequencies. It is often approximated as a linear scale from 0 to 1000 Hz, and then a logarithmic scale beyond 1000 Hz. Mel warping is implemented by passing the DFT coefficients through a filter-bank of Q overlapped triangular bandpass filters , see figure 2.10, about $12 - 20$, compacting the information and reducing the variance by averaging the DFT samples in each filter.

The cepstrum can be considered as the spectrum of the log spectrum. Removing its mean (Cepstral Mean Normalization, CMN) reduces the effects of linear time-invariant filtering (e.g., channel distortion). The density of the cepstrum has the benefit of being modeled well by a linear combination of Gaussian densities as used in the Gaussian Mixture Model [Reynolds, 1995] and together with the well performance in speaker-recognition systems have popularized their use in speaker recognition tasks.

Often, the time derivatives of the mel cepstra (also known as delta cepstra) are used as additional features to model trajectory information [Furui, 1986]. The case of the velocity and acceleration parameters, computed from the first and second order derivatives of the spectral parameters is commonly applied in speaker identification and verification tasks though in the speaker tracking and diarization this kind of parameters degrades the speaker segmentation. It suggest the use of different sets of parameter depending on they are applied to. Delta coefficients are computed by means the following regression formula,

$$\Delta \mathbf{o}_t(i) = \frac{\sum_{d=1}^{D} d\big(\mathbf{o}_{t+d}(i) - \mathbf{o}_{t-d}(i)\big)}{2\sum_{d=1}^{D} d^2} \tag{2.12}$$

where $\Delta \mathbf{o}_t(i)$ is the delta coefficient i at time t computed in terms of the corresponding static coefficients $\mathbf{o}_{t-d}$ to $\mathbf{o}_{t+d}$. The same formula is applied to the delta coefficients with another window size to obtain acceleration coefficients.

Frequency Filtering (FF) parameters [Nadeu *et al.*, 2001] has been also applied to speaker recognition with successful results [Hernando, 1997], [Luque and Hernando, 2008a], and [Luque and Hernando, 2008b]. These parameters are computed in the same fashion as the MFCC but replacing the final Discrete Cosine Transform of the logarithmic filter-bank energies by the following filter:

$$H(z) = z - z^{-1} \tag{2.13}$$

These features have several interesting characteristics: they are uncorrelated, computationally simpler than MFCCs, have frequency meaning and they have generally shown an equal or better performance than MFCCs in both speech and speaker recognition.

The spectral features are so far the parameters more widely applied to perform automatic speaker recognition, nevertheless some recent papers have introduced some novels parameters taking benefit from the multi-microphone conditions. [Pardo *et al.*, 2007] and [Pardo *et al.*, 2012] report that the use of the time-delays between microphones is useful for speaker diarization and [Koh and et al., 2008] submitted a novel diarization system to NIST RT'07 evaluation based upon both segmentation and clustering by means the Direction of Arrival (DOA) information.

Nonetheless, such spectrum-based parametrization obtain an acceptable performance in speaker recognition, they are not focused on representing the useful information to distinguish among speakers and to discriminate from other sources like background noises or music. This approach, while highly successful in clean or matched acoustic conditions, suffers significant performance degradation in the presence of variability. It ignores long-term information that can convey supra-segmental information, such as prosodic and speaking

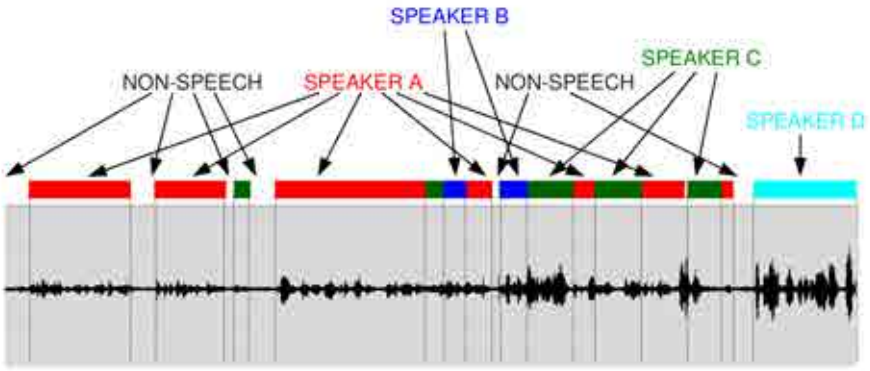


Figure 2.11: *Speech and non-speech labeling for further processing in a recognition task.*

style. That is the reason why a considerable amount of work about this issue attempt to center on the speaker characteristics and the particular conditions of the task that they are applied to. In [Chan *et al.*, 2006] they propose the use of vocal source features for the task of speaker segmentation and [Yamashita and Matsunaga, 2005] proposes a speaker segmentation system based on energy, peak-frequency centroid, peak-frequency bandwidth and other spectral features.

Furthermore, there are other levels of information that convey useful characteristics about the speaker. Because spectral slices are not modeled in sequence, the spectrum-based approach fails to capture longer-range stylistic features of a persons' speaking behavior, such as lexical, prosodic, and discourse-related habits. Recent studies [Reynolds *et al.*, 2003], [Shriberg *et al.*, 2005], [Dehak *et al.*, 2007] have demonstrated that these level can add complementary knowledge to the traditional spectrum-based recognition systems, improving their accuracy. Research on additional information sources in speaker recognition has been mainly focused on the use of the fundamental frequency and energy trajectories to capture long-term information. One of the reasons is that fundamental frequency appears to be more robust to acoustic degradations from channel and noise effects. Other studies suggest the use of different information sources: number of phonemes per word, number of frames per word, pause rate and duration, etc. [Dehak *et al.*, 2007], [Shriberg *et al.*, 2005]. In [Abad and Luque, 2010] and [Abad *et al.*, 2011] the phonemes output by an hybrid ANN/HMM speech recognition system using acoustic models of phonemes in various languages are applied to obtain a high-dimensional Parallel Transformation Network feature vector for speaker characterization.

Voice Activity Detection

Voice activity detector (VAD), as illustrated in figure 2.11, aims at locating the speech segments from a given audio signal. VAD is an important sub-component for any real-world recognition system and usually it is applied as a dedicated speech/non-speech detector in a pre-processing step. Even though a seemingly trivial binary classification task, it is, in fact, rather challenging to implement a VAD working robustly across different

domains. In the context of meetings non-speech segments may include silence, but also ambient noise such as paper shuffling, door knocks, keyboard typing or non-lexical noise such as breathing, coughing and laughing, among other background noises. In BN data, commercials, street noise, or background/overlapped music are common events among the several non-speech sources. For CTS audio, typically some form of standard energy/spectrum based speech activity detection is used since non-speech tends to be silence or noise sources [Abad *et al.*, 2010].

Therefore, depending on the domain data being used, the non-speech classes can consist of silence, music, room noise, street noise, etc. For broadcast news audio, e.g, usually five class models are trained: speech, music, noise, speech+music, and speech+noise. The extra speech models are used to help minimize false rejects of speech occurring in the presence of music or noise. It is more important to minimize speech miss rates since these are unrecoverable errors in most systems. ineffective. For meeting domain, the model-based approaches also tend to have better performances and rely on a two-class detector such as a HMM, with models pre-trained with external speech and non-speech data [Wooters and et al., 2004], [Zhu *et al.*, 2008] and optionally adapted to specific meeting conditions. Discriminant classifiers such as Linear Discriminant Analysis (LDA) [Rentzeperis *et al.*, 2006] or Support Vector Machines (SVM) [Luque and Hernando, 2008b], [Temko *et al.*, 2007] have also been proposed in the literature. The main drawback of model-based approaches is the lack of robustness against unseen conditions due their reliance on external data for the training of speech and non-speech models.

2.2.2 Modeling

One of the key issues in speaker recognition is the technique applied for speaker modeling. Several modeling strategies has been applied in the speaker recognition task. Most of them rely on assuming some data structure to a greater or lesser extent like as its probability density function. Such structure is usually characterized by a collection of parameters. The negative aspects stem from the structure assumption itself, one which may not be adequate to the modeling task by limiting the form that the multivariate probability density can take. The positive aspects derive from the succinctness of the representation and the fact that the more limited the form, fewer data that are needed to specify the density.

Following, we introduce a brief summary of the most applied modeling techniques applied to speaker recognition task.

First attempts to speaker recognition tasks employed methods based on template matching and can still be useful under constrained circumstances. By template matching, we mean the comparison of an average computed on test data to a collection of stored averages developed for each of the speakers in training. In more complex probabilistic models, the pattern matching is probabilistic and it results in a measure of the likelihood, or conditional probability, of the observation given the model. The observation is assumed to be an imperfect replica of the template, and the alignment of observed frames to template frames is selected to minimize a distance measure d. The likelihood L can be approximated by exponentiating the utterance match scores and assuming in that way that scores are proportional to log-likelihoods:

$$L = \exp^{-ad} \tag{2.14}$$

where a is a positive constant.

Statistical feature averaging or also termed **Long-term averaging** [Markel *et al.*, 1977] is an approach which employs the mean of some feature over a relatively long utterance to distinguish among speakers. For text-independent recognition, ideally one has utterances of several seconds or minutes in order to ensure that a voice is modeled by mean features of a broad range of sounds, rather than by a particular sound or phone. Test utterances are compared to training templates $\mathbf{X}$ by the distance between feature means, $\bar{x}$. Several metrics can be used for minimum distance classifiers like as the Euclidean ($W = I$) or Mahalanobis ($W = \Sigma^{-1}$) distances.

$$d(x_i, \bar{x}) = (x_i, \bar{x})^T W (x_i, \bar{x}) \tag{2.15}$$

The **nearest neighbor** (NN) method for estimating the density from a sample $R = \{r_i\}$ at point x is to measure the distance between and the point in the sample closest to x, x's nearest neighbor:

$$d_{NN}(x, R) = \min_{r_j \in R} |x - r_j| \tag{2.16}$$

Let $X = \{x_i\}$ and $R = \{r_i\}$ denote the collections of feature vectors extracted from test and reference utterances, respectively. R is used to estimate the speaker's density. There exists an inversely proportional relationship between the probability density function and the distance at the point x_i:

$$\begin{aligned} \hat{\Pr}(x_i) &= \frac{1}{V_n\big(d_{NN}(x_i, R)\big)} \\ \ln \hat{\Pr}(x_i) &\approx -\ln d_{NN}(x_i, R) \end{aligned} \tag{2.17}$$

where $V_n(\rho)$ represents the volume of sphere in n-dimensional space with radius ρ which is proportional to ρ^n. For the complete collection of test feature vectors X the log-likelihood is expressed as:

$$\hat{\ell}(X) = -\sum_{x_i \in X} \ln d_{NN}(x_i, R) \tag{2.18}$$

Since more than one neighbor is normally taken into account, the technique is commonly referred to as k-nearest neighbor (kNN), where k nearest neighbors are used in the identification process. Majority voting and sum rules are the most commonly used approaches in NN classification (Campbell, 1997; Kouand Gardarin, 2002; Cunningham and Delany, 2007).

Vector quantization (VQ) modeling constructs representatives of the data. VQ modeling is identical to nearest neighbor modeling except that distances to nearest data representatives are measured. The need to reduce the computation and memory demands of the nearest neighbor approach is a chief motivation behind VQ modeling [Soong *et al.*, 1985]. In practice the features are multi-dimensional (usually greater, nominally

between $12 - 20$) and hundreds of vectors are employed for characterizing the speaker. Selecting the data representatives can be approached as a problem of grouping the training feature vectors into clusters. All vectors falling inside a cluster are represented by a centroid, perhaps the cluster mean or a member of the cluster. The feature space is quantized by mapping every vector to one of the cluster centroids. The clusters, however, create unrealistically rigid boundaries in the sense there can be no overlap in the features generated by two different acoustic classes. Each vector belongs to one and only one cluster. The matching score for L frames of speech is given by the following expression:

$$z_{VQ} = \sum_{i=1}^{L} \min_{c_j \in \mathbf{C}} d(x_i, c_j) \tag{2.19}$$

where $C = \{c_j\}$ represent the collections of centroids also known as VQ codebook.
The corresponding distance measure in VQ modeling is perhaps the most intuitive method. This technique can be independent of time, as previous highlighted methods in which all temporal variation is ignored and global averages (e.g., centroids) are all that is used, or time-dependent. The time-dependent model is more complicated because it must accommodate human speaking rate variability.
Dynamic time warping (DTW) is the most popular method to compensate for speaking-rate variability in text-dependent systems. A text-dependent template model consists of a sequence of templates $(X_1, ..., X_N)$ that must be matched to an input sequence $(x_1, ..., x_M)$ usually of different length. A DTW algorithm does a constrained, piece-wise linear mapping of one or both time axes to align the two speech signals while minimizing the accumulated distance z, which is expressed as follows:

$$z_{DTW} = \sum_{i=1}^{M} d(x_i, X_{i(j)}) \tag{2.20}$$

The template indexes $j(i)$ are given by the DTW algorithm which aligns each x_i from the input sequence to the most likely template at the sequence step i constrained to previous and late templates. At the end of the time warping, this accumulated distance DTW is used as a match score. Dynamic time warping accounts for the variation over time of parameters corresponding to the dynamic configuration of the articulators and vocal tract [Campbell, 1997]

Gaussian Mixture Models

The Gaussian mixture model (GMM)-based approach has been identified as an effective approach for speaker modeling [Reynolds, 1995]. A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system.
The Gaussian Mixture Model [Reynolds, 1995] is a weighted sum of Gaussian distributions:

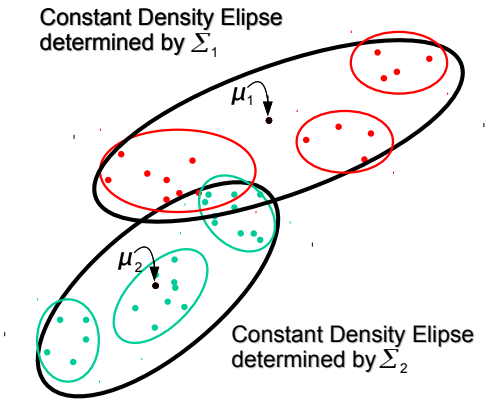


Figure 2.12: *Gaussian Mixture Model (GMM) example for feature samples of two dimensions for two speakers. The Gaussian means μ determine the location and the covariances Σ establish the shape of Gaussian distributions.*

$$\Pr\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) = \sum_{i=1}^{M} w_i \; g\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{2.21}$$

where $\mathbf{x}$ is a D-dimensional continuous-valued data vector, i.e. measurement or features, $w_i, i = 1, \ldots, M$, are the mixture weights, and $g\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), i = 1, \ldots, M$, are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form,

$$g\left(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_i})^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu_i})\right\} \tag{2.22}$$

with mean vector and covariance matrix $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ respectively and the mixture weights w_i satisfying the constraint: $\sum_{i=1}^{M} w_i = 1$. Vertical bars $|\cdot|$ indicates matrix determinant. Therefore the GMM is represented by the parameter set λ,

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}, \quad i = 1, \ldots, M \tag{2.23}$$

For a sequence of T training vectors $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_T}\}$, the GMM likelihood, assuming independence between

the vectors[1], can be written as,

$$\Pr(\mathbf{X}|\lambda) = \prod_{t=1}^{T} \Pr(\mathbf{x_t}|\lambda), \tag{2.24}$$

or in a more common expression by means log-likelihoods, just applying logarithm function to each side in equation 2.24, which allows a more tractable computation,

$$\log \Pr(\mathbf{X}|\lambda) = log \sum_{t=1}^{T} \log \Pr(\mathbf{x_t}|\lambda). \tag{2.25}$$

For each speaker that the system has to recognize, the GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) [Dempster *et al.*, 1977]. Maximum likelihood (ML) parameter estimations are obtained by using a few iterations of the EM algorithm, around $10 - 20$ iterations are usually computed. The basic idea of the EM algorithm is, beginning with an initial model λ_{t_0}, to estimate a new model λ_{t_1}, such that $\Pr(X|\lambda_{t_1}) \geq \Pr(X|\lambda_{t_0})$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

In speaker recognition task, a speaker may be modeled by using either a decoupled GMM from training data or by means a Maximum A Posteriori (MAP) estimation, a form of Bayesian adaptation [Duda and Hart, 1973], from a well-trained prior model [Gauvain and Lee, 1994], [Reynolds, 2002]. In the former case, each model is built independently by using the training utterances provided by the registering speaker. In the latter case, also termed GMM-adaptation, each model is the result of adapting a general model, which represents a large population of speakers, to better represent the characteristics of the specific speaker being modeled. This general model is usually referred to as world model or universal background model (UBM). An UBM is a large GMM (2048 mixtures) model used in a biometric verification systems to represent general, person independent feature characteristics to be compared against a model of person-specific feature characteristics when making an accept or reject decision. For example, in a speaker verification system, the UBM is a speaker-independent Gaussian Mixture Model (GMM) trained with speech samples from a large set of speakers to represent general speech characteristics. Using a speaker specific GMM trained with speech samples from a particular enrolled speaker, a likelihood-ratio test for an unknown speech sample can be formed between the match score of the speaker specific model and the UBM. The UBM may also be used when training the speaker-specific model by acting as a the prior model in MAP parameter estimation in order to gain robustness against newly or incomplete data.

The specifics of the adaptation are as follows. Given a prior model and training vectors from the desired class, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$, we first determine the probabilistic alignment of the training vectors into the prior mixture components. That is, for mixture i in the prior model (UBM), we compute $\Pr(i|\mathbf{x}_t, \lambda_{\text{UBM}})$ as the percentage of the mixture component i to the total likelihood,

[1]The independence assumption is often incorrect but needed to make the problem tractable.

$$\Pr(i|\mathbf{x}_t, \lambda_{\text{UBM}}) = \frac{w_i \; g\big(\mathbf{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\big)}{\sum_{i=1}^{M} w_i \; g\big(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\big)} \tag{2.26}$$

Then compute the sufficient statistics for the weight, mean and variance parameters [2]

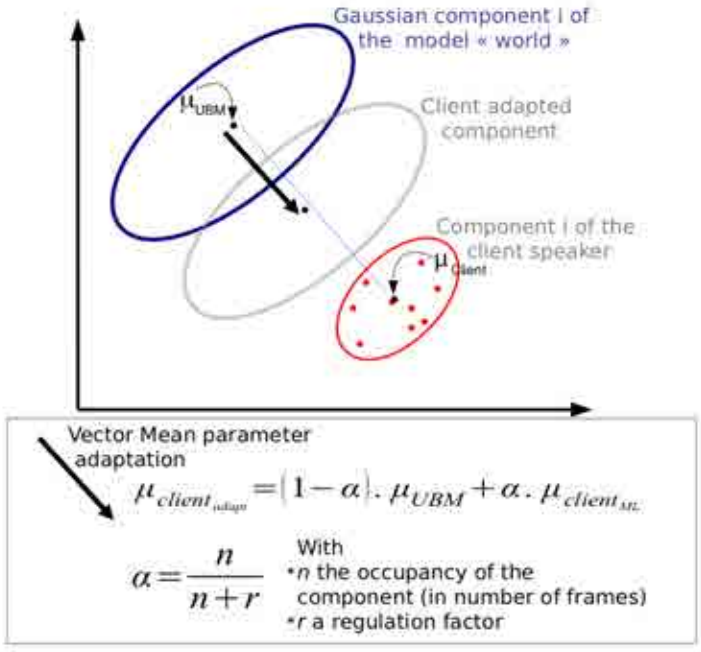


Figure 2.13: *Example of adaptation of the mean component of a Gaussian Mixture Model (GMM) in feature space of two dimensions. This procedure is one of the most popular approaches for model adaptation due its simplicity and good performance. The training vectors (red dots) are probabilistically mapped into the UBM (prior) mixtures and the adapted mixture parameters are derived using the statistics of the new data and the UBM (prior) mixture parameters. The adaptation is data dependent, so UBM (prior) mixture parameters are adapted by different amounts.*

$$n_i = \sum_{i=1}^{T} \Pr(i|\mathbf{x}_t, \lambda_{\text{UBM}}) \tag{2.27}$$

$$E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{i=1}^{T} \Pr(i|\mathbf{x}_t, \lambda_{\text{UBM}}) \; \mathbf{x}_t \tag{2.28}$$

$$E_i(\mathbf{x}^2) = \frac{1}{n_i} \sum_{i=1}^{T} \Pr(i|\mathbf{x}_t, \lambda_{\text{UBM}}) \; \mathbf{x}_t^2 \tag{2.29}$$

Lastly, these new sufficient statistics from the training data are used to update the prior sufficient statistics

[2] $(x)^2$ is shorthand for diag$((xx))$.

(prior model λ_{UBM} for mixture i to create the adapted parameters for such mixture through equations:

$$\hat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i\right]\gamma \tag{2.30}$$

$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\mathbf{x}) + (1 - \alpha_i^m)\boldsymbol{\mu}_i \tag{2.31}$$

$$\hat{\boldsymbol{\sigma}}_i^2 = \alpha_i^v E_i(\mathbf{x}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2 \tag{2.32}$$

The adaptation coefficients controlling the balance between old and new estimates are $\{\alpha^w, \alpha^m, \alpha^v\}$ for the weights, means and variances, respectively. The scale factor, γ, is computed over all adapted mixture weights to ensure they sum to unity. Note that the sufficient statistics, not the derived parameters, such as the variance, are being adapted. For each mixture and each parameter, a data-dependent adaptation coefficient α^p , $p \in w, m, v$, is used in the above equations. This is defined as:

$$\alpha_i^p = \frac{n_i}{n_i + r^p} \tag{2.33}$$

where r_p is a fixed *relevance* factor for parameter p. It is common in speaker recognition applications to use one adaptation coefficient for all parameters ($\alpha_w = \alpha_m = \alpha_v = \frac{n_i}{n_i + r}$) and further to only adapt certain GMM parameters, such as only the mean vectors, as it is depicted in the figure 2.13, while the weights and covariances are shared between all speakers. The relevance factor is a way of controlling how much new data should be observed in a mixture before the new parameters begin replacing the old parameters. This approach is robust to limited training data.

The concept of a UBM is also applied for discriminative systems, such as Support Vector Machines (SVM), where explicit likelihood functions for the two hypothesis (belongs or not belongs) are not used. In this case, the UBM refers to the collection data from the general population used as negative examples when training a person specific discriminate function.

Hidden Markov Models

A Hidden Markov Model (HMM) is a stochastic model, a kind of Bayesian network, commonly used for modeling sequences, in which the observations are a probabilistic function of the state [Rabiner, 1989];[Rabiner and Juang, 1993];[Campbell, 1997]. Gaussian mixtures are usually employed in speech to model probability distributions of each state, where the probability density function associated to the random variables (features) is a multivariate distribution with probability density function: $X \sim N(\mu, \Sigma)$. The HMM, as a finite-state machine, has each state associated with a deterministically observable event and the observations (features) are stochastic function of the state. The states are connected by a transition network, where the state transition probabilities are $a_{ij} = \Pr(s_i|s_j)$. The figure 2.14 illustrates and example of a four-state hidden Markov model commonly applied to text-dependent speaker recognition. Note that a GMM may be considered as an HMM composed by just one state.

By using the Baum-Welch algorithm, the probability that a sequence of speech frames was generated by the

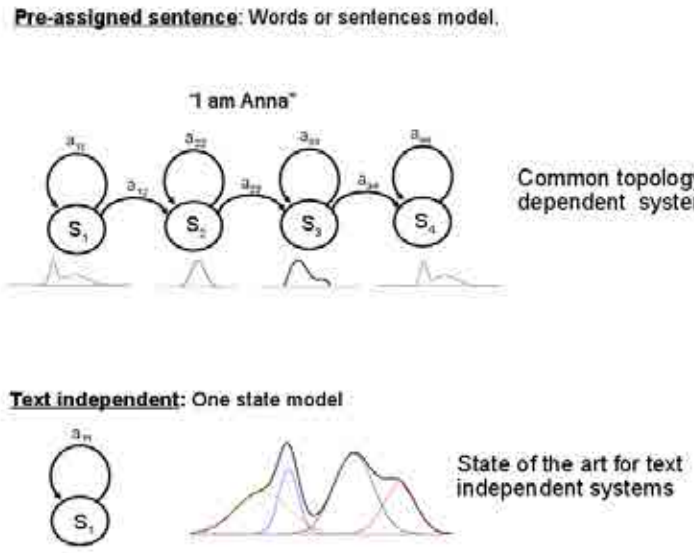


Figure 2.14: *Examples of Hidden Markov Models. At top, a common topology for a text-dependent text recognition system. At bottom, a trivial HMM composed of one state modeled by a GMM.*

model can be determined [Rabiner and Juang, 1986];[Rabiner, 1989]. This probability, or likelihood, is used as a score for L frames of input speech given the model. Given a sequence of speech frames $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_L\}$ the likelihood depending on the model is computed as:

$$\Pr(\mathbf{X}|\lambda) = \sum_{\substack{\text{all state} \\ \text{sequences}}} \prod_{i=1}^{L} \Pr(\mathbf{x}_i|s_i) \Pr(s_i|s_{i-1}) \tag{2.34}$$

Artificial Neural Networks

Artificial neural networks Artificial neural networks (ANN) are also used in speaker recognition applications. The kind of neural networks used are feed-forward neural networks, where the information moves only in one direction (forward) from the input nodes, through the hidden nodes, if any, and to the output nodes. Commonly, a feed-forward neural network is created for each known speaker, and each network contains one output that is trained to be active only for its speaker. In the testing phase, an input feature vector is fed forward through each network, and the identification is determined by the network with the highest accumulated output values. In the speaker verification mode, the input vectors of the unknown user are fed forward through the network belonging to the claimed speaker. If the average output value is bigger than a threshold, the speaker is accepted [Oglesby and Mason, 1990].

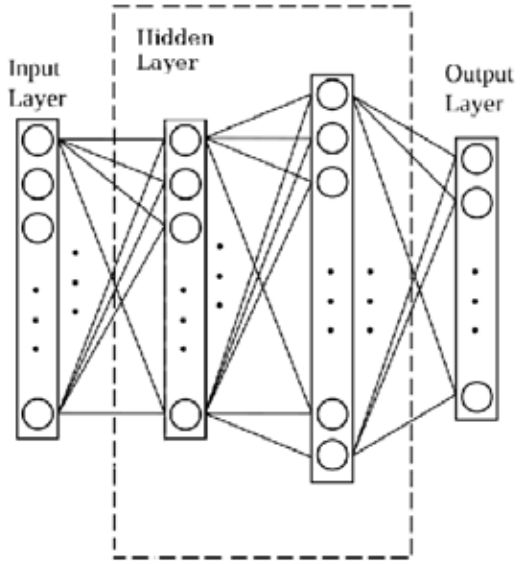


Figure 2.15: *Artificial Neural Network example.*

Another kind of networks, the time-delay neural networks (TDNN) were developed by [Bennani and Gallinari, 1991] to capture transient information using a connectionist approach.
In a recent work [Abad and Luque, 2010; Abad *et al.*, 2011], the authors present a new approach based upon ASR transcriptions and upon an adaptation network. In this case, features carrying out speaker characteristics are obtained from the adaptation transforms applied to the Multi-Layer Perceptrons (MLP) that form a connectionist speech recognizer. Finally, speaker features are modeled by means SVM technique. Such a system is described in more detail in chapter 4.

Support Vectors Machine

Support vector modeling relies on stacking a huge number of speech features in a vector (super vector) which is finally modeled by a Support vector machine (SVM). This strategy is known to be a high performance speaker recognition approach and other tasks thanks to their ability to generalize. Support vector machines (SVM) [Boser *et al.*, 1992], [Cortes and Vapnik, 1995] is a state-of-the-art binary classifier and one of the most currently fusion techniques based on the discriminative approach. Recent works on statistical machine learning have shown the advantages of discriminative classifiers like SVM in a wide range of applications [Cristianini and Shawe-Taylor, 2000].
The SVM model relies on two assumptions. First, transforming data into a high-dimensional space may convert complex classification problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions. Second, SVMs are based on using only those training patterns that are near the decision surface assuming they provide the most useful information for classification. The maximum margin

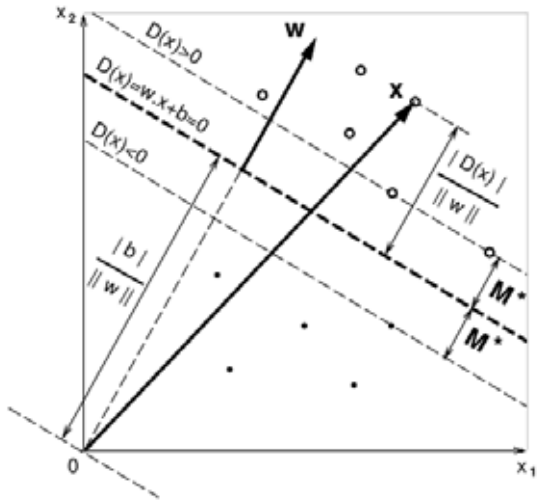


Figure 2.16: *Two-class linear classification by SVM. Image from [Boser et al., 1992].*

training algorithm finds a decision function for pattern vectors x of dimension n belonging to either of two classes A and B which separates through the hyperplane of equation $wx + b = 0$. The input to the training algorithm is a set of p examples x_i with label y_i:

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_p, y_p)\} \tag{2.35}$$
$$\text{where} \begin{cases} y_k = 1 & \text{if } \mathbf{x}_k \in \text{class A} \\ y_k = -1 & \text{if } \mathbf{x}_k \in \text{class B} \end{cases}$$

From these training examples the algorithm finds the parameters of the decision function $D(x)$ during a learning phase. After training, the classification of unknown patterns is predicted according to the following rule:

$$\begin{aligned} x \in A \quad & \text{if } D(x) > 0 \\ x \in B \quad & \text{otherwise} \end{aligned} \tag{2.36}$$

The algorithm corresponds to a linear method in a high-dimensional feature space non-linearly related to the input space. Given a linearly separable two-class training data, the SVM algorithm finds an optimal hyperplane that splits input data in two classes, maximizing the distance of the hyperplane to the nearest data points of each class, see figure 2.16. However, data are normally not linearly separable. In this case,

non-linear decision functions are needed, and an extension to non-linear boundaries is achieved by using specific functions called kernel functions [Boser *et al.*, 1992]. The kernel functions map the data of the input space to a higher dimensional space, the feature space, by a non-linear transformation. The optimal hyperplane is then constructed in the feature space, creating a non-linear boundary in the input space. The mentioned hyperplane for a non-linearly separable [3] data is defined by:

$$D(x) = \sum_{k=1}^{p} \boldsymbol{\alpha}_k t_k K(\mathbf{x}_k, \mathbf{x}) + b \tag{2.37}$$

where $t_k = \{1, -1\}$ are the labels or desired outputs, K is a chosen kernel and the coefficients α_i are such that the following condition is satisfied:

$$\sum_{i=1}^{N} \alpha_i y_i = 0; \quad 0 \leq \alpha_i \leq C \tag{2.38}$$

The vectors $\mathbf{x}_k$ are the support vectors, which determine the optimal separating hyperplane and correspond to the points of each class that are the closest to the separating hyperplane. N is the number of support vectors and C is an adjustable parameter that controls the effect of the misclassified data. In linearly non-separable the SVM replaces the inner product $\mathbf{x} \cdot \mathbf{y}$ by a kernel function $K(\mathbf{x}; \mathbf{y})$, and then constructs an optimal separating hyperplane in the mapped space. The kernel $K(\mathbf{x}; \mathbf{y})$ is constrained to have certain properties (the Mercer condition [Vapnik, 1998]), so that can be expressed as:

$$K(\mathbf{x}, \mathbf{y}) = b(\mathbf{x})^t b(\mathbf{y}), \tag{2.39}$$

where $b(x)$ is a mapping from the input space (where x lives) to a possibly infinite dimensional space. The kernel is required to be positive semi-definite. The Mercer condition ensures that the margin concept is valid, and the optimization of the SVM is bounded. The optimization condition relies upon a maximum margin concept, see figure 2.12. For a separable data set, the system places a hyperplane in a high dimensional space so that the hyperplane has maximum margin. The data points from the training set lying on the boundaries (as indicated in dashed lines in figure 2.12) are the support vectors in equation 2.37. According to the Mercer theorem [Vapnik, 1998], the kernel function implicitly maps the input vectors into a high dimensional feature space in which the mapped data is linearly separable. Possible choices of kernel functions include:

- Polynomial $K(\mathbf{x}; \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d$ where the parameter d is the degree of the polynomial
- Gaussian Radial Basis Function $K(\mathbf{x}; \mathbf{y}) = exp(-\frac{(\mathbf{x} - \mathbf{y})^2}{2\sigma^2})$, where the parameter σ is the width of the Gaussian function
- Multi-Layer Perceptron $K(x; y) = \tanh(k(\mathbf{x} \cdot \mathbf{y}) - \mu)$, where the k and μ are the scale and offset parameters.

[3] In a linearly separable case, $D(\mathbf{x}) = sign(\sum_{i=1}^{t} \boldsymbol{\alpha}_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + \mathbf{b})$ the kernel function is just an inner product

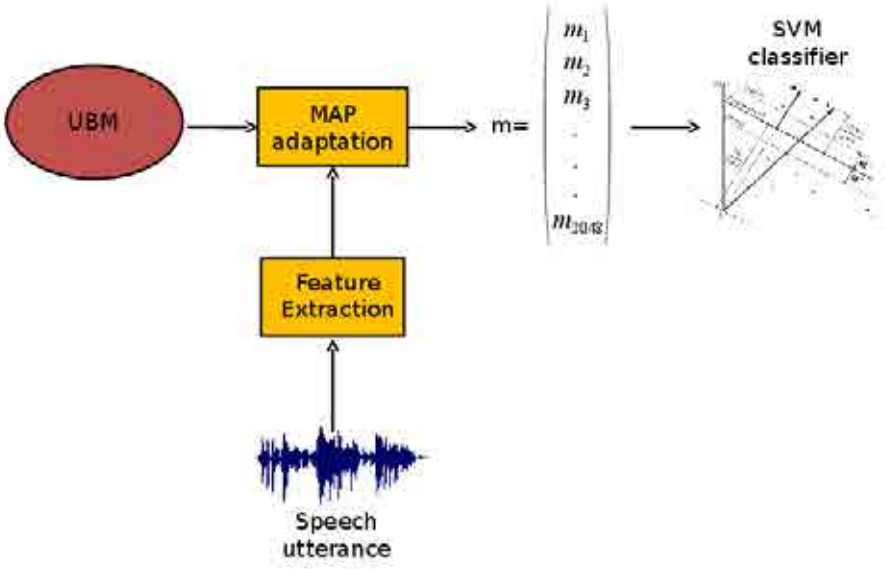


Figure 2.17: *A Gaussian supervector modeling example by stacking the GMM-UBM means which feed a SVM classifier.*

Since the SVM is a two-class classifier, we handle speaker recognition as verification problems. That is, we use a one vs. all strategy [Campbell *et al.*, 2006a]. We train a target model for the speaker and the set of known non-targets are used as the remaining class. We repeat the process and produce models for other speakers. For speaker verification, we train in a manner similar to speaker identification. For each target speaker, we label the target speaker's utterances as class 1. We also construct a background speaker set (class 0) that consists of example impostor speakers. The example impostors should be representative of typical impostors to the system. We keep the background speaker set the same as we enroll different target speakers. In contrast to the speaker identification problem, the non-target set of speakers is not as well defined; we try to capture a representative population of example impostors and, in contrast to GMM-UBM approach, we look for a set of impostor speakers "close" to the target speaker instead of a set representative of the whole population.
Furthermore, combining Gaussian mixture models with Support vector machines [Campbell *et al.*, 2006b], the so-called *Gaussian supervector* approach (GSV), see figure 2.17 surpasses previous approaches based on standard GMM-UBM or SVM classification of spectral features. Supervectors modeling is a robust way to present utterances using a single vector. Such single point representation leads to several advantages i.e. avoids the normalization issues due the length variability of utterances in training and enrollment and on the other hand, supervectors give a new feature domain (the supervector space) in which compensation methods as inter-session variability compensation techniques have a suitable framework for development, see joint factor analysis (FA) [Kenny *et al.*, 2008b] in next section or nuisance attribute projection (NAP) [Solomonoff *et al.*, 2007]. It that sense, it becomes possible to directly quantify and remove the unwanted variability from the

supervectors.

The supervectors has applied as inputs to support vector machine (SVM) as illustrated in figure 2.17 with successful results in speaker recognition tasks. This leads to sequence kernel SVMs, where the utterances with variable number of feature vectors are mapped to a fixed-length vector using the sequence kernel [Wan and Renals, 2005]. The way to obtain an enrolled speaker model is as follows. First Gaussian Mixture Models for each speaker client are obtained with MAP adaptation of the Gaussian means of the UBM, i.e. based on spectral features [Reynolds *et al.*, 2000], as described in the section 2.2.2. The UBM means are adapted with few MAP iterations and usually with a relevance factor of 16. The Gaussian Super Vector (GSV) system stacks the mixture means of the MAP adapted Gaussian speaker models to obtain super vectors of every speech segment. For each speaker, that the system has to recognize, a high dimensional supervector are computed which feeds a discriminative algorithm, i.e. SVM, to perform classification.

In [Campbell *et al.*, 2006b] the authors derive the Gaussian supervector kernel by bounding the Kullback-Leibler (KL) divergence measure between GMMs. Lets the UBM $\lambda_{UBM} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and two utterances a and b which are described by their MAP-adapted GMMs, see section 2.2.2. That is, $\lambda_a = \{w_k, \boldsymbol{\mu}_k^a, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\lambda_b = \{w_k, \boldsymbol{\mu}_k^b, \boldsymbol{\Sigma}_k\}_{k=1}^K$. The KL divergence kernel is then defined as,

$$K(\lambda_a, \lambda_b) = \Sigma_{k=1}^K \left(\sqrt{w_k}\boldsymbol{\Sigma}_k^{(-1/2)}\boldsymbol{\mu}_k^a\right)^T \left(\sqrt{w_k}\boldsymbol{\Sigma}_k^{(-1/2)}\boldsymbol{\mu}_k^b\right), \tag{2.40}$$

which just is a variance normalization of the means. All the Gaussian means μ_k need to be normalized with factor $\sqrt{w_k}\ \Sigma_k^{(-1/2)}$ before feeding the SVM classifier with them.

2.2.3 Compensation Techniques

Segmentation and Score Normalization

Observe that if we decompose a segment X which is a collection of N frames, into K segments of size n_k where $n_k = N$, and let $\log L_k$ denote the log likelihood of the k^{th} segment, then

$$S = \log L = \sum_{k=1}^{K} \log L_k \tag{2.41}$$

That is, the log likelihood (or segment score S) of the full segment is, because of the frame independence assumption, the sum of log likelihood of the smaller segments. The first step is to extract segments from the test utterance. A direct approach is to uniformly chop the dialog into segments of some arbitrary size, with the idea that a subset of these segments will be pure enough to be recognized. The importance of representing the log likelihood as the sum of smaller segments is that it enables us to generalize the scoring of a utterance in several important ways. The first aspect is our ability to select the best model from the collection of models for each speaker for each of the different segments. The second reason for evaluating multiple segments is that partitioning an utterance enables us to discard or de-emphasize segments contaminated by other speakers and noise. The above log likelihood 2.41 provide us with the unnormalized scores. These scores require

normalization. Recall that the log likelihood of the data for a segment is obtained by evaluating a probability model or the data in the segment; other segments have different data scored with different models and therefore there is no basis for comparison of likelihoods from different segments. The line of research into the field of acoustic which faces such issues is speaker normalization. Such sort of techniques have the main purpose of avoid the influence of background, non-linear distortions produced by the channel or noise and other non-speaker events giving robustness against unseen data and variability in the channel or the environment. Normalization techniques can be applied at different levels.

At the *speech waveform level*, e.g. spectral subtraction algorithms [Boll, 1979] which aims to suppress or to reduce the spectral effects of acoustically added noise in the digital waveform. Spectral subtraction relies on the suppression of stationary noise from speech by subtracting the spectral noise bias calculated during non speech activity.

At the *feature level*, e.g. cepstral mean normalization (CMN) [Liu *et al.*, 1993] which removes the mean of the cepstral coefficients in order to avoid non-linear effects due the channel distortions or feature warping techniques [Pelecanos and Sridharan, 2001] that attempts to warp the cepstral distribution into a standardized Gaussian shape. Some examples applied in the diarization task can be found in [Zhu and et al., 2005] and [Sinha *et al.*, 2005]. In these works, feature warping [Pelecanos and Sridharan, 2001], [Ouellet *et al.*, 2004], [Reynolds, 2003] is applied on the distribution of cepstral features mapping it to a standardized distribution over a specified time interval.

At the *model level*, e.g. newly introduced techniques which deals with the intra-speaker variability and the inter-session variability. Some examples are the nuisance attribute projection (NAP) method that works by removing subspaces that causes variability in the kernel of the SVM [Campbell *et al.*, 2006b], or the joint factor analysis (JFA) [Kenny, 2005] modeling which attempts to estimate such variability in order to compensate it directly from the models.

At *score level*, e.g. in speaker recognition task the works from [Zhu and et al., 2005] and [Abad and Luque, 2010] make use of score normalization techniques applied on the models and at the score level. Some of normalization techniques are based on the world model as UBM normalization [Reynolds, 1997] [Reynolds, 1995], which is derived from Bayes' theorem, or perform distribution scaling like as cohort normalization [Rosenberg *et al.*, 1992] which uses a set of cohort speakers who are close to the target speaker, test normalization (also known as T-norm), zero normalization (also known as Z-norm) [Zheng *et al.*, 2005]. Mostly of them coming from the speaker verification field [Liu *et al.*, 1993] [Bimbot *et al.*, 2004]. The cohort can be seen as a replacement for the world model by calculating a probability of the cohort under the conditions of the observation. The selection of the cohort set of speakers can be done during the training stage. If a large set of speakers is chosen for a cohort, it behaves as an impostor-centric normalization. A normalization technique which uses a mean and variance estimation for distribution scaling is zero normalization (Z-norm) [Reynolds, 1997]. The advantage of Z-norm is that the estimation of the normalization parameters can be performed off-line during training. A speaker model is tested against example impostor utterances $Z = \{z_1, z_2, \ldots, z_N\}$ and the log-likelihood scores are used to estimate a speaker specific mean μ_Z and variance σ_Z for the impostor

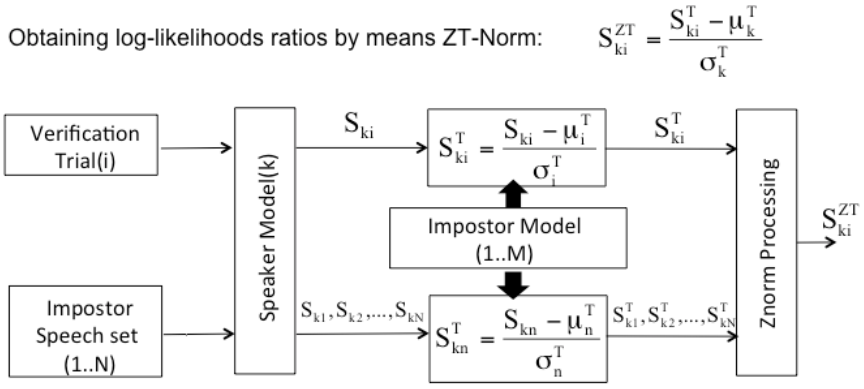


Figure 2.18: *ZT normalization scheme to perform distribution scaling over the scores. Firstly a T-norm is applied over the log-likelihood produced by the test utterance. Following, a zero normalization with parameters estimated from a set of impostors utterances is applied to previous T-scaled log-likelihoods.*

distribution. The normalization has the form,

$$S_Z = \log L_Z = \frac{\log(\Pr(\lambda|\mathbf{O})) - \mu_Z}{\sigma_Z}, \tag{2.42}$$

A normalization method which is also based on a mean and variance estimation for distribution scaling is test normalization (T-norm). During testing, a set of example impostor models is used to calculate impostor log-likelihood scores for a test utterance, similar to a cohort approach. However, unlike the cohort approach, a mean and variance parameter are estimated from these scores. These parameters are then used to perform the distribution normalization in the same fashion as equation 2.42,

$$S_T = \log L_T = \frac{\log(\Pr(\lambda|\mathbf{O})) - \mu_T}{\sigma_T}, \tag{2.43}$$

where μ_T and σ_T are the scores' mean and standard deviation respectively that are obtained by a set of impostors models $\lambda_{T-norm} = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ evaluated over the speech segment. The advantage of T-norm over a cohort normalization is the use of the variance parameter which approximates the distribution of the cohort population more accurately. The estimation of these distribution parameters is carried out on the same utterance as the target speaker test. Therefore, an acoustic mismatch between the test utterance and normalization utterances, possible in Z-norm, is avoided.

The figure 2.18 depicts the scheme to obtain ZT-normalized log likelihoods by combining a test normalization step together with zero normalization. Such score normalization has been applied to speaker recognition task with well performance in spite of the prohibitive computational cost, where impostor models and impostor utterances sets can reach hundreds of speakers. [Auckenthaler *et al.*, 2000].

Nuisance Attribute Projection

Any variation in different utterances of the same speaker (be it due to different handsets, environments, or phonetic content), as characterized by their supervectors is harmful and leads to degradation on the speaker recognition performance. For a given speaker, the supervectors estimated from different training utterances may not be the same especially when these training samples come from different handsets. Channel compensation is therefore necessary to make sure that test data obtained from different channels (than that of the training data) can be properly scored against the speaker models. We will now discuss two different techniques. Former based on switch off those directions of variability, nuisance attribute projection (NAP), and latter based on generative modeling, that is, Gaussian mixture model (GMM) with factor analysis (FA) technique.
Nuisance attribute projection (NAP) is a successful method for compensating SVM supervectors [Solomonoff *et al.*, 2007]. It is not specific to some kernel, but can be applied to any kind of SVM supervectors. The NAP transformation removes the directions of undesired sessions variability from the supervectors before SVM training. The NAP transformation of a given supervector s is [Brummer *et al.*, 2007],

$$\hat{\mathbf{s}} = \mathbf{s} - \mathbf{P}(\mathbf{P}^T s), \tag{2.44}$$

where $\boldsymbol{P}$ is the eigenchannel matrix. The eigenchannel matrix is trained using a development data set with a large number of speakers, each having several training utterances (sessions) and labeled with channel nuisance variables, e.g. electret, carbon button, cell microphones. The training set is prepared by subtracting the mean of the supervectors within each speaker and pooling all the supervectors from different speakers together; this removes most of the speaker variability but leaves session variability. By performing eigenanalysis on this training set, one captures the principal directions of channel variability. The P projection matrix removes the component of a vector in the direction of a specified subspace. Hence NAP attempts to estimate a subspace that contains mostly channel information and build a projection that zeroes that out, by minimizing the figure of merit:

$$\delta = \sum_{ij} W_{ij}\mathbf{P}(\mathbf{s}_i - \mathbf{s}_j)^2, \tag{2.45}$$

where $\mathbf{s}_i$ is the supervector from utterance i, W_{ij} are a weight matrix whose entries correspond to pairs of training observation vectors, weighting the importance of the variability direction, and P is a projection matrix with corank$(\mathbf{P}) \ll dim(\mathbf{X})$ and $\mathbf{P}^2 = \mathbf{P}$. W_{ij} is positive for pairs of vectors that we want to move together, it means, sessions i and j have different background conditions or zero otherwise. Taking into account P matrix can be written as $\mathbf{P} = \mathbf{I} - \mathbf{v}\mathbf{v}^t$ where $\mathbf{v}$ is a matrix with orthonormal columns spawning the directions being removed. The design criterion for P is:

$$\mathbf{v}^* = \underset{\mathbf{v},\|\mathbf{v}\|_2}{\operatorname{argmin}} \sum^{i,j} W_{ij} \parallel \mathbf{P}\mathbf{s}_i - \mathbf{P}\mathbf{s}_j \parallel_2^2, \tag{2.46}$$

whose solution is an eigenvalue problem:

$$\mathbf{A}(\text{diag}(\mathbf{W1}) - \mathbf{W})\mathbf{A}^t\mathbf{v} = \lambda\mathbf{v}, \tag{2.47}$$

where $\mathbf{A}$ is a matrix whose columns are $\mathbf{s}_i$, $\mathbf{W}$ is the matrix consisting of W_ij and $\mathbf{1}$ is the vector of all ones. Summarizing, equation 2.44 then just means subtracting the supervector that has been projected on the channel space.

Joint Factor Analysis

The technique of joint factor analysis (JFA) [Kenny and Dumouchel, 2004] was proposed for modeling explicitly the channel variability aiming to compensate the channel effects. The JFA model considers the variability of a Gaussian supervector as a linear combination of the speaker and channel components. Given a training sample, the speaker-dependent and channel-dependent supervector $\mathbf{M}$ is decomposed into two statistically independent components, as follows

$$\mathbf{M} = \mathbf{s} + \mathbf{c}, \tag{2.48}$$

where $\mathbf{s}$ and $\mathbf{c}$ are referred to as the speaker and channel supervectors, respectively. Let d be the dimension of the acoustic feature vectors and K be the number of mixtures in the UBM. The supervectors M, s and c live in a Kd dimensional parameter space. The channel variability is explicitly modeled by the channel model of the form,

$$\mathbf{c} = \mathbf{U}\mathbf{x}, \tag{2.49}$$

where $\mathbf{U}$ is a rectangular matrix and $\mathbf{x}$ are the channel factors estimated from a given speech sample. The columns of the matrix $\mathbf{U}$ are the eigenchannels estimated for a given data set. During enrollment, the channel factors $\mathbf{x}$ are to be estimated jointly with the speaker factors $\mathbf{y}$ of the speaker model of the following form:

$$\mathbf{M} = \mathbf{m} + \mathbf{V}y + \mathbf{D}z, \tag{2.50}$$

In the above equation, $\mathbf{m}$ is the UBM supervector, $\mathbf{V}$ is a rectangular matrix with each of its columns referred to as the eigenvoices, $\mathbf{D}$ is $Kd \times Kd$ diagonal matrix and $\mathbf{z}$ is a $Kd \times 1$ column vector. In the special case $\mathbf{y} = 0, \mathbf{s} = \mathbf{m} + \mathbf{Dz}$ describes exactly the same adaptation process as the MAP adaptation technique (section 2.2.2). Therefore, the speaker model in the JFA technique can be seen as an extension to the MAP technique with the eigenvoice model $\mathbf{Vy}$ included, which has been shown to be useful for short training samples. The matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{D}$ are called the hyperparameters of the JFA model. These matrices are estimated beforehand on large data sets. One possible way is to first estimate $\mathbf{V}$ followed by $\mathbf{U}$ and $\mathbf{D}$ [Kenny, 2005], [Kenny *et al.*, 2008b]. For a given training sample, the latent factors $\mathbf{x}$ and $\mathbf{y}$ are jointly estimated and followed by estimation of $\mathbf{z}$. Finally, the channel supervector $\mathbf{c}$ is discarded and the speaker supervector $\mathbf{s}$ is

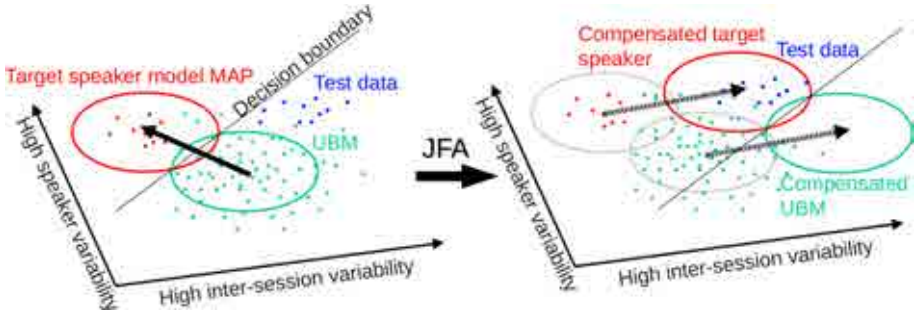


Figure 2.19: *Joint Factor Analysis' key point: At left, the target speaker model is obtained through Bayesian adaptation (Maximum a Posteriori, MAP) of the means of the UBM. Next a decision threshold is chosen based upon the training data. Both the high session variability or intra-speaker variability is depicted as newly target speaker data (blue dots). In that case the decision leads to an error since the trained models does not count for such variability. At right, JFA estimation allows to compensate such variability, shifting the models consequently and compensating the newly data variability. For recognition, move both models along the high inter-session variability direction(s) to fit well the test data (e.g. in ML sense).*

used as the speaker model. By doing so, channel compensation is accomplished via the explicit modeling of the channel component during training. For detailed account of estimation procedure the reader should refer to [Kenny *et al.*, 2005]; [Kenny *et al.*, 2008b]. For comparing various scoring methods, refer to [Glembek *et al.*, 2009]. The JFA approach has become one of the most successful compensation techniques for speaker verification as has been reported in [Kenny and Dumouchel, 2004], [Kenny *et al.*, 2007] and [Kenny *et al.*, 2008a] dominating the latest NIST 2010 and 2008 speaker recognition evaluations (SRE) [Martin, 2010].

2.2.4 Evaluation metrics

After having computed a match score of similarity between the input user and the corresponding template stored in the database, a decision is taken whether the user must be accepted or rejected by the system. However, such decision can be both correct or not correct. If the decision is incorrect, two different types of error can occur [Chollet and Bimbot, 1995]:

- **False rejection (or non detection)**: the system rejects a valid identity claim.
- **False acceptance (or false alarm)**: the system accepts an identity claim from an impostor.

Both types of errors give rise to two types of error rates, which are commonly used to measure the performance of a system:

- **False rejection rate (FRR)**: the system rejects a valid identity claim.
- **False acceptance rate (FAR)**: the system accepts an identity claim from an impostor.

Either of the two types of errors can be reduced at the expense of an increase in the other, so that the trade-off between FRR and FAR depends on a decision threshold. In a real-world system, which is usually not perfect,

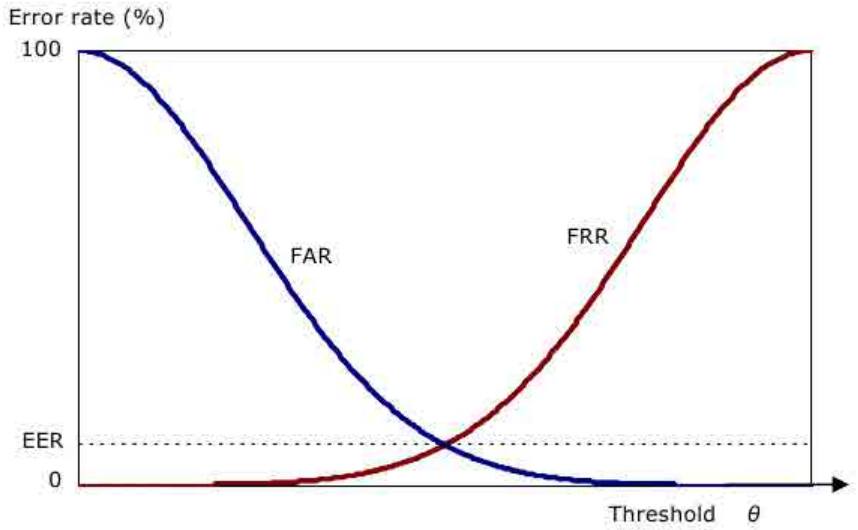


Figure 2.20: *FRR and FAR as a function of a threshold θ. The intersection point between the two curves determines the value of the EER.*

FRR and FAR intersect at a certain point (figure 2.20). Such a value of FRR and FAR at this point is known as *Equal Error Rate* (EER).

If the threshold is set to a low value, the system tends to accept most of the identity claims, giving few false rejection errors but many false acceptances. On the contrary, with a high threshold the system tends to reject most of the identity claims, giving rise to few false acceptance errors and a lot of false rejection.

The *Receiver Operating Characteristic* (ROC) curve plots the FRR versus the FAR [Chollet and Bimbot, 1995]. This curve is monotonous and decreasing, and the better the system is, the closer to the origin the curve will be. Another representation of the ROC curve is used sometimes by plotting the correct detection rate (instead of FRR) versus the false alarms [Duda and Hart, 1973].

It is also common to plot the error curve on a normal deviate scale. In this case, the curve is known as the *Detection Error Trade-offs* (DET) curve [Chollet and Bimbot, 1995]. In an hypothetical system whose clients and impostors scores are Gaussians with the same variance, the DET curve is a linear curve where the slope equals -1, which becomes more easily readable and comparable to other DET curves. As in the ROC curve, better systems are closer to the origin. In a real system, the score distributions are not exactly Gaussians, but they are close enough to them to allow this representation. Figures 2.21 depicts an example of both ROC and DET curves. The intersection of each curve with the diagonal dotted line indicates the value of the EER.

Detection Cost

Apart from the DET curve and the EER associated, for the NIST speaker evaluations [Fiscus and et al., 2009a], NIST provides and additional cost function which measure the system performance establishing a fixed cost to

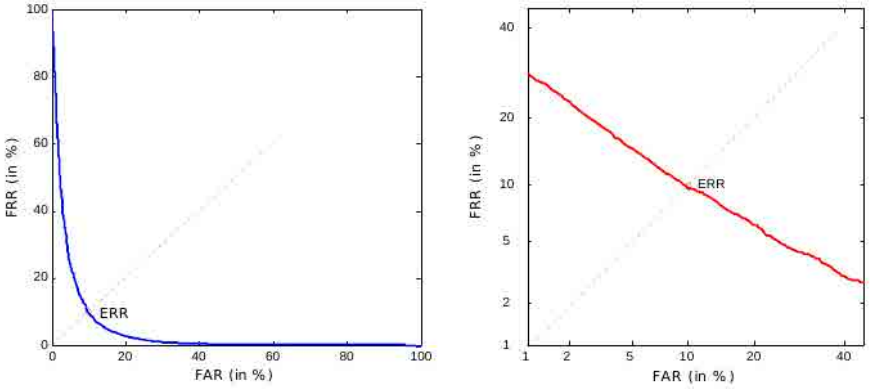


Figure 2.21: *(a) Example of a ROC curve. (b) The corresponding DET curve (b).*

FA and FR errors as well as a priori probability for target and non-target individuals. Likewise, it matches the metric to the particularities of the task by favoring the cost of some error type instead of the other one. This cost is defined for speaker verification as:

$$C_{Det} = C_{FR} \cdot P_{FR|S_T} \cdot P_T + C_{FA} \cdot P_{FA|S_{NT}} \cdot P_{S_{NT}} \tag{2.51}$$

where C_{FR} and C_{FA} are the associated costs to FR and FA errors respectively; $P_{FR|S_T}$ (the probability of false reject given a target speaker) measures the system FR; $P_{FA|S_{NT}}$ (the probability of false acceptance given a non-target speaker) measures the system FA; and finally, P_T and $P_{NT} = 1P_T$ the prior target and non-target probability. In NIST speaker evaluations and by extension, in this work, costs and target probability will be set as follows:

- $C_{FA} = C_{FR} = 1$
- $P_T = 0.001$

2.3 Speaker Diarization and Tracking

Speaker diarization and tracking belong to the speaker-based processing techniques. Therefore the feature representation of the acoustic signal attempts to represent the speaker information and discriminate between different talkers as it is done in the previous mentioned tasks, identification and verification. With the increasing availability of archived audio material comes an increasing need for efficient and effective means of searching and indexing through this voluminous material. Searching or tagging speech based on who is speaking is one of the more basic components required for dealing with audio archives, such as recorded meetings or the audio portion of broadcast shows. Traditional approaches to speaker recognition, however, are designed to

identify or verify the speaker in a speech sample known to be spoken by a single person. For audio indexing or searching, the basic recognition approach needs to be expanded to handle both detection and tracking of speakers in multi-speaker audio [Dunn *et al.*, 2000].

In general, a spoken document is a single-channel recording that consists of multiple audio sources. Audio sources may be different speakers, music segments, types of noise, etc. For example, a broadcast news program consists of speech from different speakers as well as music segments, commercials, and sounds used to segue into reports. Audio diarization is defined as the task of marking and categorizing the audio sources within a spoken document where types and details of the audio sources are application specific.

Audio diarization is the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources. These sources can include particular speakers, music, background noise sources, and other signal source/channel characteristics. Diarization has utility in making automatic transcripts more readable and in searching and indexing audio archives [Reynolds and Torres-Carrasquillo, 2005]. The goal when searching and indexing target speakers is to find and identify the regions in the audio streams that belong to the target speakers and produce an efficient way for accessing these regions in the audio-data archives. The task of finding such speaker-defined regions was first introduced in the Rich Transcription project in *"Who spoke when"* evaluations, [Fiscus and et al., 2004]. The diarization task is also defined by the amount of specific prior knowledge allowed, as example, speech from just a few of the speakers, the number of speakers in the audio, or the structure of the audio recording, like as commercials or music to segue into next show sections.

Whenever there is a speaker of interest, it may be desirable to determine not only whether the speaker appears in a multi-speaker segment, but to identify the specific intervals within the segment corresponding to the speaker through speaker identification methods. The task of identifying the regions associated with particular speakers is known as a speaker tracking task and was defined during a 1999 NIST Speaker Recognition evaluation, [Martin and et al., 2000]. Whereas diarization and tracking procedures serve for the detection of speakers in audio data, the purpose of speaker indexing is the organization of audio data according to detected speakers for efficient speaker-based audio-retrieval. For a more portable and independent speaker diarization system, it is desired to operate without specific prior knowledge of speakers, the number of speakers in the audio or the structure. This is the general task definition used in the Rich Transcription evaluations [Fiscus and et al., 2004].

Most speaker diarization systems and tracking systems share a similar general architecture, see figure 2.22. Firstly, the signal is chopped into homogeneous segments. The segment boundaries are located by finding acoustic changes in the signal and each segment is expected to contain speech from only one speaker. The resulting segments are then clustered so that each cluster corresponds to an unique speaker, a major issue being that the number of speakers is unknown a priori and needs to be automatically estimated. There may be specific prior knowledge via previous example speech from the speakers in the audio, then the task becomes a tracking task. The true identities of the speakers are obtained in a speaker-identification module in the next stage. Here, a multiple-speaker verification of each cluster is performed. A speaker-identification module is capable of

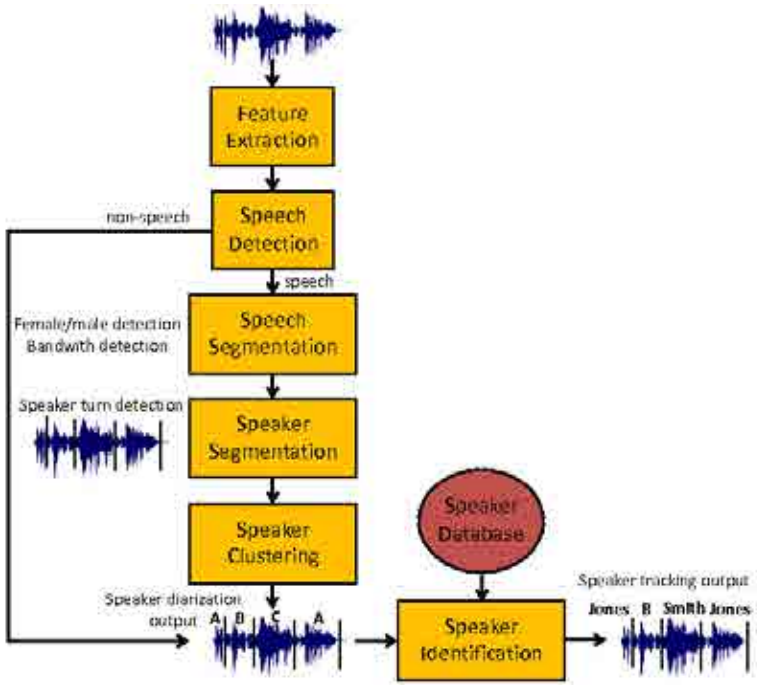


Figure 2.22: *General speaker diarization and tracking schemes.*

recognizing just those speakers, who are present in the database of target speakers and are previously enrolled in the system. The speech data from clusters that do not correspond to any of the speakers in the target group should be marked as unknown speaker data. At the end, a speaker index is derived, which is used as a basis for searching and tracking speakers in the audio database. Each particular system also presents specific aspects which can be classified following different criteria, performing speaker segmentation and clustering at the same time instead of the sequential approach or introducing the prior speaker knowledge at the beginning of the diarization process. Anyway, the figure 2.22 does not try to explain all possible ways to perform speaker diarization/tracking but give the reader a brief summary of the main concepts in such tasks.

Speaker Diarization Domain

There are three primary domains for speaker diarization research and development: broadcast news audio, recorded meetings, and telephone conversations [Reynolds and Torres-Carrasquillo, 2005]. Depending on domain characteristics, diarization algorithms had to be adapted according to such differences. The three domains mainly differs in the nature and quality of the data recorded: bandwidth, kind of microphones, studio or telephonic conditions, background noises, music, the amount and types of non-speech events, the number of speakers, the durations and sequencing of speaker turns and the style or spontaneity of the speech are some examples. Each domain presents unique diarization challenges, although often high-level system techniques

tend to generalize well over several domains. The NIST Rich Transcription speaker evaluations [Fiscus and et al., 2007a], [Fiscus and et al., 2009a] have primarily used both broadcast news and meeting data whereas the NIST speaker recognition evaluations [Martin, 2010] have primarily used conversational telephone speech (CTS) with summed sides, also known as two-wire conversational telephonic speech.

Number of speakers is usually larger in broadcast news ranging among $25 - 50$ in a one-hour duration show, $4 - 10$ for a conference meeting or 2 in CTS data. Speaker turns occur less frequently than they do in conference meeting data or telephone data, resulting in BN having a longer average speaker turn length. For CTS data, a huge variety of cellular and microphones are used to capture the speech signal and, as is well know, with a characteristic bandwidth of 8 KHz. In the case of BN speech data is usually acquired using lavalier microphones with some recordings being made in the studio and others outside. The diarization systems must be able to deal with the non-homogeneous data found in broadcast audio, such as a wide variety of speakers and speaking styles, changing speakers, accents, background conditions, etc.

There are notable differences in speaking style observed in BN, meetings and CTS data. Broadcast speech is much closer to written language than conversational speech or meeting speech are, where different social conventions are observed. Speech may be spontaneous or often read, as in a news reporting or at least prepared in advance as in an interview or telephone speech. At the other side, meetings are usually recorded using desktop or far-field microphones which are less invasive for meeting users than head-mounted or lavalier microphones that are used primarily for annotation purposes. In CTS, the speech quality is affected by a variety of different types of telephone handset, the background noise (other conversations, music, street noise, etc.), as well as a much higher proportion of interruptions, overlapping speech, and third person interjections or side conversations. The challenges for CTS at the acoustic level concern speaker normalization, the need to cope with channel variability, spontaneous speech, and the need for efficient speaker adaptation techniques. Therefore, just differences between meeting room configurations, microphone placement or kind of telephone handsets lead to variations in recording quality [Reynolds, 1996]. Furthermore, speech overlapping among speakers or the Lombard effect [4] results in a signal-to-noise ratio generally better for BN data than it is for meeting recordings. Although BN recordings may contain speech that is overlapped with music, laughter, or applause, in general, the detection of acoustic events and speakers tends to be more challenging for conference meeting data than for BN data or CTS data.

In terms of linguistic content, in CTS data there are many more speech fragments, hesitations, restarts and repairs, as well as back-channel confirmations to let each interlocutor know the other person is listening. The first-person singular form is much more predominant in conversational speech. Another major difference from BN is that some interjections such as ”uh-huh” and ”mhm” (meaning yes) and ”uh-uh” (meaning no) that are considered as non-lexical items in BN. The word ”uhhuh,” which serves both to signal agreement and a back-channel ”I’m listening,” accounts for about 1% of the running words in the CTS data. The most common word in the English CTS data, I, accounts for almost 4% of all word occurrences, but only about 1% of the word occurrences in BN [Matsoukas *et al.*, 2006].

[4] The Lombard effect is the involuntary tendency of speakers to increase their vocal effort when speaking in loud noise to enhance the audibility of their voice

An extensive analysis of BN characteristics is reported in [Anguera, 2006] and a comparison of BN and conference meeting data can be found in [Mirghafori and Wooters, 2006].

Main Algorithms

Automatic speech transcription (ASR) and speaker diarization rely on similar methods for segmentation and clustering. Nonetheless, differences in their objectives leads to different needs, particularly concerning where accuracy is most important. Automatic transcription requires accurate segment boundaries. Although the rejection of non-speech segments is useful in order to minimize insertion of words and to save computation time, it is important that the segment boundaries are located in non-informative zones such as silences or breaths. Indeed, having a word cut by a boundary disturbs the transcription process and increases the word error rate. In speaker diarization or tracking the error is computed by a time-metric which computes the error rate frame-by-frame as the percentage of frames correctly identified. Therefore, fuzzy speaker boundaries or not well detected are not specially harmful as in the ASR case. In typical diarization tasks, the number of speakers in a given audio stream is not a priori known and must be estimated from data. This means that the diarization system has to solve simultaneously two problems: finding the actual number of speakers and pooling together speech from the same speaker. This problem is often cast into a model selection problem. The number of speakers determines the complexity of the model in terms of number of parameters. The model selection criterion chooses the model with the right complexity and thus the number of speakers. It can be found several diarization approaches in the literature but main algorithms can be summarized in the following points:

- **Speech enhancement:** Speech data can be optionally preprocessed using Wiener filtering [Wiener, 1949] to attenuate noise using, for example, [Adami *et al.*, 2002]. The most common approach to multi channel speaker diarization involves acoustic beamforming as initially proposed in [Anguera *et al.*, 2007a]. Many RT participants use the free and open-source acoustic beamforming tool kit known as BeamformIt [Anguera, 2005] which consists of an enhanced delay-and-sum algorithm to correct miss alignments due to the time-delay-of-arrival (TDOA) of speech to each microphone. A reference channel is selected and the other channels are appropriately aligned and combined with a standard delay-and-sum algorithm. The contribution made by each signal channel to the output is then dynamically weighted according to its SNR or by using a cross-correlation based metric.

- **Voice activity detection:** The aim of this step is to find the regions of speech in the audio stream. A simple solution that works satisfactorily on typical telephone-quality speech data, uses signal energy to detect speech. Nevertheless, differences in the kind of speech, microphones or room configurations may result in variable signal-to-noise ratios (SNRs) from one data to another and high differences on energy levels can be observed in the non-speech parts of the signal. Therefore, the characterization and classification of non-speech class becomes one of the most difficult due its natural variability. More complex approaches apply maximum likelihood classification with Gaussian Mixture Models (GMMs)

trained on labeled data for both speech and non-speech events. Viterbi algorithm is a common technique to perform segmentation using the models to identify speech regions. A word or phone decoding step may also be used for finer grain speech boundary detection. Usually non-speech frames detected are discarded for further processing in the diarization algorithm.

- **Modeling strategy:** Each speaker can be modeled by a Gaussian mixture model (GMM) with diagonal covariance matrices composed of several components. As is done in the speaker recognition task, larger models with 2048 components have been proposed [Ben and et al., 2004], [Meignier and et al., 2001], [Ajmera and Wooters, 2003]. In this case, a more robust estimation of the models despite the limited amount of data per speaker can be obtained by performing the maximum a posteriori (MAP) adaptation of a prior model [Gauvain and Lee, 1994]. On the other hand, using a single Gaussian with a full covariance matrix for the modeling of a speaker also provides good results [Moh *et al.*, 2003]. Some approaches, as in [Ajmera and et al., 2004], make use of an automatic model complexity estimation known as model selection criterion, i.e. inference on the suitable number of models' parameters aiming both to avoid manual parameter tuning and to obtain better estimate of the right number of actual speakers.

- **Link between segmentation and clustering:** Segmentation can be done first, followed by clustering with no connection between the two parts [Ben and et al., 2004], inspired from the work presented in [Siegler and et al., 1997], [Chen and Gopalakrishnan, 1998]; alternatively, the segmentation and clustering can be jointly optimized, via, for example, the iterative segmentation and clustering procedures described in [Gauvain *et al.*, 1998], [Meignier and et al., 2001], [Ajmera and Wooters, 2003]. A limitation of the first method is that errors made in the segmentation step are not only difficult to correct later, but can also degrade the performance of the subsequent clustering step.

- **Clustering strategy:** It relies either on an agglomerative clustering [Gauvain *et al.*, 1998],[Ajmera and Wooters, 2003] or on a divisive clustering method [Meignier and et al., 2001], [Tranter and Reynolds, 2004].

2.3.1 Speaker Segmentation

The aim of audio segmentation is to find time-stamps in the audio streams at changes between different speakers or acoustic environments. To detect target speakers in an audio stream, it is best to segment the audio into homogeneous regions according to changes in speaker identity, environmental conditions and channel conditions. Furthermore, if the content is Broadcast News (BN), one would like to segment the audio stream into homogeneous regions according to speaker identity or gender, environmental condition and channel condition so that regions of different nature can be handled differently: e.g., regions of pure music and noise can be rejected; also, one might design a separate recognition system for telephone speech. Since a same speaker may appear multiple times in several conditions it is not easy to create a correct segmentation. Most of the systems are based in the Bayesian Information Criterion (BIC) but there exist various segmentation algorithms

proposed in the literature that can be categorized into three main categories [Chen and Gopalakrishnan, 1998]: *Energy and Decoder-guided*, *Model-based* and *Metric-based*. The energy-based segmentation only places boundaries at silence locations, which in general has no direct connection with the acoustic changes in the data. Both the model-based and the metric-based segmentation schemes rely on thresholding of measurements which lack stability and robustness. More importantly, they do not generalize to unseen acoustic conditions.

Energy-based Segmentation

Energy-based approaches have been widely used (e.g. see [Wegmann *et al.*, 1999], [Wactlar *et al.*, 1996]) and are particularly easy to implement. Basically, silence periods in the input signal are detected, and segment boundaries are hypothesized in such silence periods if some additional constraints are satisfied, like minimum length of the silence period.
Energy or decoder-guided segmentation is based on the hypothesis most changes will occur between the speakers silences though there is no clear relationship. Two main examples joins this category. Energy-based systems which use the energy behavior to seek the points most likely to exist a speaker change [Kemp and M. Schmidt, 2000] and decoder-guided systems which by means a recognition system find the points from the detected silence regions [Tranter and Reynolds, 2004]. A constrain in the length of the silences is normally imposed to avoid false alarms and other techniques are in charge of assess the exactly point of change. They are oriented to speech recognition and their use in diarization or tracking is marginal.
It is reported in the literature that model-based and metric-based techniques outperform the simpler energy-based algorithms [Kemp and M. Schmidt, 2000]. While model-based segmenters achieve very high level of segment boundary precision, the metric-based segmenter performs better in terms of segment boundary recall [5].

Model-based Segmentation

In model-based segmentation [Wilcox *et al.*, 1994], [Woodland *et al.*, 1998], a set of models for different acoustic classes is defined and trained prior to segmentation. The incoming audio stream is classified using the models, usually imposing additional minimum class length constraints. Boundaries between the classes are used as segment boundaries. Model-based segmentation assumes knowledge about the type of the audio that is to be segmented thus specific models for a closed set of acoustic classes are trained "a priori". Models are applied to the audio stream to classify by Maximum Likelihood (ML) and to obtain the changing points as the boundaries between classes. Speech, silence, female-male, target speakers and combinations of them are some examples of such prior trained classes [Kubala and et al., 1997], [Gauvain *et al.*, 1998], [Kemp and M. Schmidt, 2000].

[5] The result of a segmentation can contain two possible types of error. Type-I-errors occur whether a true segment boundary has not been detected (deletion). Type-II-errors occur if a found segment boundary does not correspond to a segment boundary in the reference (false alarm, or segment insertion). The information retrieval community uses two closely related numbers, precision (PRC) and recall (RCL).They are defined as $RCL = \frac{\text{number of correctly found boundaries}}{\text{total number of boundaries}}$ and $PRC = \frac{\text{number of correctly found boundaries}}{\text{number of hypothesized boundaries}}$

Model-based segmentation is the common point for algorithms which perform both segmentation and clustering at the same time and without prior knowledge of the classes. The best systems in later NIST RT Evaluations follow this strategy based on a ML decoding of Gaussian Mixture Models (GMM) [Ajmera and Wooters, 2003], [Wooters and Huygbregts, 2008].

Metric-based Segmentation

Metric based segmentation is the most applied technique up to date [Chen and Gopalakrishnan, 1998], [Cettolo and Federico, 2000], [Chen *et al.*, 2002], [Pietquin *et al.*, 2002], [Lu and Zhang, 2002], [Ajmera and Wooters, 2003] and [Perez-Freire and C-Garcia-Mateo, 2004]. The audio stream is segmented at places where the maximum of distances between neighboring windows appear. Therefore it is based on a computation of a distance between two acoustic segments to discern the homogeneity of segments. A change point is detected whenever the distance among two windows is over a threshold. Mainly, there are two different kind of distances: The former is *statistics-based* distances which compares the sufficient statistics from two acoustics sets of data. It relies on the computation of single mean and variance. The latter, called *likelihood-based* distances, are based on the evaluation of the likelihood computed by models representing the data. Nevertheless these distances are costly than the first one, they obtain better results than statistics-based methods. Following, the most popular metrics we can find in the literature are briefly described:

- **Bayesian Information Criterion (BIC)**: Probably is the most extensively used segmentation and clustering metric due to its simplicity and effectiveness. Introduced by Schwarz in [Schwarz, 1973] and [Schwarz, 1978] the BIC value gives a mind about how well the model suits the data. BIC is a maximum-likelihood, asymptotically optimal, Bayesian model selection criterion penalized by the model complexity. Consider a data set Z , and a set of parametric models $\{m_1, m_2, \ldots, m_j\}$ where m_j is a parametric model with parameters trained on the data Z. Model selection aims at finding the model $\hat{m}$ such that

$$\hat{m} = \underset{j}{\operatorname{argmax}}\{\Pr(m_j|Z)\} = \frac{\Pr(m_j)\Pr(Z|m_j)}{\Pr(Z)}, \tag{2.52}$$

  which depends only on the maximization of $\Pr(Z|m_j)$ since $\Pr(Z)$ and $\Pr(m_j)$ are considered constant and uniform prior, respectively. In case of parametric modeling, e.g., HMM/GMM, it is possible to write:

$$\Pr(Z|m_j) = \int \Pr(Z, \boldsymbol{\theta}_j|m_j)d\boldsymbol{\theta}_j. \tag{2.53}$$

  This integral cannot be computed in closed form in the case of complex parametric models with hidden variables (e.g., HMM/GMM). However, several approximations for 2.53 are possible, the most popular one being the Bayesian Information Criterion (BIC) [Schwarz, 1978],

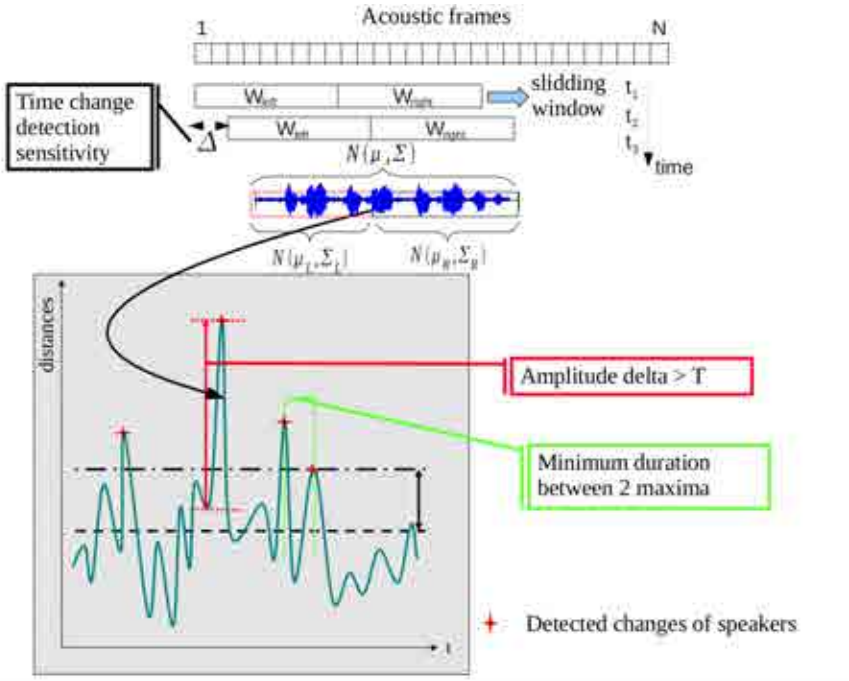


Figure 2.23: *Example of metric-based segmentation. The distance measure between two windows is computed, e.g. by means Bayesian Information Criterion (BIC) and iterated along time by sliding such windows. An algorithm post-processes the distances values and decides whether a speaker change turn occurs.*

$$BIC(m_j) = \log \Pr(Z|\hat{\boldsymbol{\theta}}_j, m_j) - \frac{p_j}{2} \log N, \tag{2.54}$$

where p_j is the number of free parameters in the model m_j, $\hat{\boldsymbol{\theta}}_j$ is the MAP estimate of the model computed from data Z, and N is the number of data samples. Hence models with larger number of parameters will produce high values of log-likelihood but will be penalized by the second term in equation 2.54. BIC is only exact in the asymptotic limit $N \to \infty$ which implies to tune the penalty term according to an heuristic threshold.

For application on speaker turn detection, two different models are employed. Assume that there are two neighboring segments and around time. The problem is to decide whether or and not a speaker change point exists at t_j. Let $Z = X \cap Y$ and N_X, N_Y, N_Z be the numbers of samples in segments X, Y and Z, respectively. Obviously, $N_Z = N_X + N_Y$. The problem is formulated as a two hypothesis testing problem. Under H_0 there is no speaker change point at time t_j. MLE is used to compute the parameters of a Gaussian distribution that models the data samples in Z. Let us denote by $\boldsymbol{\theta}_Z$ the parameters of the Gaussian distribution, i.e., the mean vector μ_z and the full covariance matrix $\boldsymbol{\sigma}_Z$. The log-likelihood under H_0 is

$$L_0 = \sum_{i=1}^{N_x} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Z) + \sum_{i=N_x+1}^{N_z} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Z) \tag{2.55}$$

where $z_i \in \Re^d,\ i = 1, 2, \ldots, N_z$ which are assumed to be independent vector of acoustic features. Under H_1 there is a speaker change point at time t_j. The segments X and Y are modeled by a distinct multivariate Gaussian densities, whose parameters are denoted by θ_X and θ_Y, respectively. The log-likelihood L_1 under H_1 is given by

$$L_1 = \sum_{i=1}^{N_x} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_X) + \sum_{i=N_x+1}^{N_z} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Y) \tag{2.56}$$

The BIC value variation is defined as:

$$\Delta BIC(Z) \quad = \quad L_1 - L_0 - \frac{\lambda}{2}\underbrace{\left(d + \frac{d(d+1)}{2}\right)}_{complexity} \ln N_z \lesseqgtr 0, \tag{2.57}$$

where the term over the bracket is the penalty, which corresponds to the number of free parameters of a multivariate Gaussian process in d dimensions and λ is the weight penalty factor (tuned "a priori" with training data). If $\Delta BIC > 0$, then time t_j is considered to be a speaker change point. Otherwise, there is no speaker change point at time t_j. Such comparison is performed over the data by means a sliding window, as illustrated in figure 2.11, leading to a set of BIC values for further processing, e.g., a second BIC step for refinement [Perez-Freire and C-Garcia-Mateo, 2004]. Detection of acoustic changes clearly depends on λ; as a matter of fact, performance of BIC-based systems is very sensitive to the selection of this parameter. Another parameter that requires special attention is N , i.e. the size of the analysis window, since reliability of Gaussian estimates depends directly on this value.

Δ BIC segmentation was first applied to speaker segmentation task in [Shaobing and Gopalakrishnan, 1998] and [Chen and Gopalakrishnan, 1998] using a single full covariance Gaussian for each neighboring segment model. Most of the implementations make use of a growing window with inner variable length segments to iteratively find the changing points [Shaobing and Gopalakrishnan, 1998], [Lu and Zhang, 2002], [Cettolo and Vescovi, 2003], [Barras and et al., 2004], [Zhu *et al.*, 2008] and [Li and Schultz, 2009]. ΔBIC is computationally demanding compared to other statistics-based metrics but it has been reported to perform better results. Some works propose a two-pass implementations through other faster metrics as GLR and the use of BIC as a second-pass for refinement [Delacourt and Wellekens, 2000], [Vandecatseye and et al., 2004], [Kim and et al., 2005]. Other recent work handles the high computational demanding by computation of ΔBIC just on potential speaker change points estimated from a model of utterance durations [Kotti *et al.*, 2008]. BIC has also been applied as a initial step to

obtain data partitioning for a posterior integrated segmentation/clustering approach as in [Barras *et al.*, 2006]. Another BIC-variant metric, referred to as cross-BIC and introduced in [Anguera and Hernando, 2004], involves the computation of cross-likelihood: the likelihood of a first segment according to a model tuned from the second segment and vice versa.

New approaches have been appeared last years which aim to avoid the dependency of Δ BIC on the overall sample size, i.e. avoid the penalty term estimation. In [Stafylakis *et al.*, 2010] the authors proposed a new variant of BIC, the segmental-BIC where each parameter of the model is penalized only with its effective sample size. Using this approach, the dependency of the BIC formula on the overall sample size is eliminated.

- **Generalize Likelihood Ratio (GLR)**: GLR was first proposed for change detection in [Willsky and Jones, 1976] and [Appel and Brandt, 1982]. Given a set of observations Z and a partition $Z = X \cup Y$, H_0 hypothesis considers that both segments are uttered by the same speaker and, consequently, a single model, θ_Z, trained with the whole Z data segment represents better it. Whilst hypothesis H_1 considers each segment comes from a different speaker, therefore two disjoint models, trained with X and Y respectively, θ_X and θ_Y, represents better it. The test is defined as a likelihood ratio between such two hypothesis, as follows:

$$GLR(X,Y) = \frac{\Pr(H_0|Z)}{\Pr(H_1|X,Y)} = \frac{L(Z|\theta_Z)}{L(X|\theta_X)L(Y|\theta_Y)}. \tag{2.58}$$

The GLR distance is stated as $D(X,Y) = -\log(GLR(X,Y))$ and setting a suitable threshold in order to decide whether the speaker change occurs. GLR differs from the ΔBIC technique in that the probability density functions of the models are unknown and they must to be estimated directly from the segment data. Usually GLR is applied at the first stage of a two-step implementation, over-segmenting the data [Gangadharaiah and et al., 2004], [Delacourt and Wellekens, 2000]. The most representative algorithm of the GLR applied to the speaker segmentation task is DISTBIC [Delacourt and Kryze, 1999], [Delacourt and Wellekens, 2000]. DISTBIC makes use of a two-step segmentation by firstly applying GLR followed by a BIC as refinement boundaries step. Furthermore, a penalized GLR was proposed in [Liu and Kubala, 1999] in order to adapt the criterion taking into account the amount of training data available into the two neighbor segments.

- **Gish distance**: is a likelihood-based metric obtained as a variation to GLR introduced in [Gish and Schmidt, 1994]. GLR is defined as follows,

$$D_{gish}(i,j) = -\frac{N}{2}\log\left(\frac{|S_i|^{\alpha}}{|S_j|^{1-\alpha}/|W|}\right), \tag{2.59}$$

where S_i and S_j represent the sample covariance matrices for each segment , $\alpha = \frac{N_i}{N_i + N_j}$ and W is

their sample weighted average $W = \frac{N_i}{N_i + N_j} S_i + \frac{N_j}{N_i + N_j} S_j$.

The work in [Kemp and M. Schmidt, 2000] reports a comparison of this distance against other metrics in the speaker segmentation task.

- **Kullback-Leibler (KL)**: Introduced by Siegler [Siegler and et al., 1997] is an efficient and fast metric with acceptable results. Given two random distributions X, Y K-L distance between distributions is defined as,

$$KL(X,Y) = E_X(log\frac{P_X}{P_Y}), \tag{2.60}$$

with E_x is the expected value with respect to the pdf of X. Assuming two Gaussian distributions it can be written in function of their covariance and means [Campbell, 1997]:

$$\begin{aligned} KL(X,Y) \quad = \quad & \frac{1}{2}Tr\big[(C_X - C_Y)(C_Y^{-1} - C_X^{-1})\big] + \\ & +Tr\big[(C_Y^{-1} - C_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T\big] \end{aligned} \tag{2.61}$$

A symmetric alternative of KL distance has proved to be more popular in speaker diarization when used to characterize the similarity of two audio segments,

$$KL2(X,Y) = KL(X,Y) + KL(Y,X), \tag{2.62}$$

being X,Y the pdf distribution of the samples into two acoustic segments. A distance between such distribution can be obtained applying 2.62 as in [Delacourt and Wellekens, 2000] and [Zochova and Radova, 2005].

- **Information Change Rate (ICR)**: is a newly introduced distance metric that has shown promise in a speaker diarization task [Vijayasenan *et al.*, 2007], [Han and Narayanan, 2008] and [Vijayasenan *et al.*, 2009]. The Information Change Rate (ICR), or entropy can be used to characterize the similarity of two neighboring speech segments determining the variation in terms of information that would be obtained by merging them. Unlike the measures outlined above, the ICR similarity is not based on a model of each segment but on the distance between segments in a space of relevance variables, with maximum mutual information or minimum entropy. The ICR approach is computationally efficient and, in [Han and Narayanan, 2008], ICR is shown to be more robust to data source variation than a BIC-based distance.

2.3.2 Speaker Clustering

Both speaker diarization and speaker tracking systems refers to those systems that perform a segmentation of the input audio and then a speaker clustering, joining the created segments into homogeneous groups, or in the tracking approach, to the corresponding target speakers. Most of them make use of a same strategy. It defines some sort of distance between segments and iteratively shares out them amongst the clusters minimizing such distance. Therefore speaker clustering can be applied as a second step following the prior segmentation provided by a speaker segmentation algorithm. Such a kind of approaches are known as sequential systems or step-by-step systems.

However, with such an approach to diarization, there is no provision for splitting segments which contain more than a single speaker, and thus performance of diarization algorithms relies on a segmentation of sufficiently high quality. Alternative approaches combine clustering with iterative resegmentation, hence facilitating the introduction of missing speaker turns, correcting early errors, mostly missed speaker turns from the segmentation step. Most state-of-the-art speaker diarization engines unify the segmentation and clustering tasks into one step. In these systems, segmentation and clustering are performed hand-in-hand in one loop by means a Viterbi realignment. During realignment, the audio stream is resegmented based on the current clustering hypothesis before the models are retrained on the new segmentation. Several iterations are usually performed taking into account all data instead of local information as sequential system makes. Most state-of-the-art systems employ some variations on this strategy. Furthermore, the diarization algorithms include mechanisms to estimate the right number of classes (total number of speakers) since no prior knowledge about speakers is given. Most of the clustering distances employed are based on distances presented in previous section 2.3.1. Different segments are usually represented using HMM/GMM models with EM training or MAP adaptation and a Viterbi algorithm is used to reassign all the data into the closest newly-created models. Such processing is sometimes performed several times for the frame assignments to stabilize. Moreover, a minimum assignment duration, according to the estimated minimum length of any given speaker turn, is usually enforced. Such a minimum duration turn avoids an unrealistic assignment of very small consecutive segments to different speaker models.

In a wide view, the most popular speaker clustering technique is agglomerative hierarchical clustering (AHC). It can be categorized into *top-down* or *bottom-up* clustering depending on the initial strategy.

The hierarchical algorithms reach the optimum number of clusters by iterative processing of the different clusters obtained by merging or splitting existing ones. The details of how this strategy works (for the bottom-up alternative) are shown in Algorithm 1. In other words, using given speech segments as initial clusters, AHC recursively merges/splits the closest pair of clusters. The recursive process is stopped when it is decided that extra cluster merging/splitting does not improve clustering performance any more.

- **Bottom-up clustering** Coming from the Pattern Classification field [Duda and Hart, 1973] is by far the mostly used approach for speaker clustering. Such a method was initially proposed by ICSI for a bottom-up system [Ajmera *et al.*, 2002] and [Ajmera and Wooters, 2003] and has subsequently been adopted by many others , [Luque and Hernando, 2008a], [van Leeuwen and Konečný, 2008], [Bozonnet

Algorithm 1 *Agglomerative Hierarchical Clustering (AHC), bottom-up alternative.*

Require: $\{\mathbf{x}_i\},\ i = 1 \ldots, \hat{n}:$ speech segments
$\hat{C}_i,\ i = 1, \ldots, \hat{n}$: initial clusters
Ensure: $C_i,\ i = 1, \ldots, n$: finally remaining clusters
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\},\ i = 1, \ldots, \hat{n}$
2: **repeat**
3: $i, j \leftarrow \text{argmin}\ d(\hat{C}_k, \hat{C}_l),\ k, l = 1, \ldots, \hat{n}, k \neq l$
4: merge $\hat{C}_i$ and $\hat{C}_j$
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** no more extra cluster merging is needed
7: **return** $C_i,\ i = 1, \ldots, n$

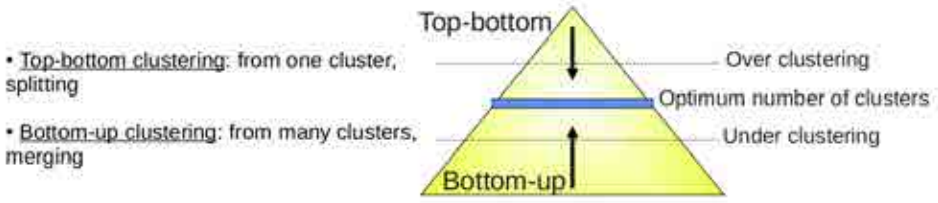


Figure 2.24: *Two main approaches to clustering schemes.*

et al., 2010a], [Friedland *et al.*, 2011]. It usually defines a distance matrix associated with the current clustering and it merges the closest pair iteratively until the stopping criterion is reached. Combinations of the previous distances in the segmentation and merging stages are assessed in the recent literature and newly distances adapted to the multi-Gaussian case have been introduced. For example, [Beigi and et al., 1998] proposed a matrix distance between all Gaussian pairs among two models where Euclidean, Mahalanobis and KL distances are benchmarked. The work in [Rougi and et al., 2006] proposes a distance between two GMM based on KL by defining an average of the minimum KL distances from each Gaussian pairs between such two models,

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{K_1} W_1(i) \min_{j=1}^{K_2} KL(N_1(i), N_2(j)), \tag{2.63}$$

where $N_1(i)$ is the ith Gaussian from the model $\boldsymbol{\theta}_1$, $N_2(j)$ is the ith Gaussian from the second model $\boldsymbol{\theta}_2$; K_1 and K_2 are the number of Gaussian components per each model respectively and $W_1(i)$ is the ith Gaussian weight from model 1. The strategy was oriented to perform on-line speaker clustering by reducing the computation of the distance just taking into account a few number of components. Most of the systems with access to the whole recording, off-line systems, propose some threshold in the distance as stopping criterion as in [Ben and et al., 2004] and [Sankar and et al., 1995]. In [Sankar and et al.,

1995] and [Malegaonkar and et al., 2006] the symmetric relative entropy distance [Juang and Rabiner, 1985] is used for speaker clustering as follow:

$$d(X,Y) = \frac{1}{2}\left[\log \Pr(X|\boldsymbol{\theta}_1) - \log \Pr(X|\boldsymbol{\theta}_2) + x + \log \Pr(Y|\boldsymbol{\theta}_1) - \log \Pr(Y|\boldsymbol{\theta}_2)\right]$$

A particular interest have been focus on the systems which obtain the speakers models by means speaker adaptation. They use a Universal Background Model (UBM) and MAP adaptation to derive speakers models from each cluster. It have been used in [Ben and et al., 2004], [Moraru *et al.*, 2005], [Barras and et al., 2004], [Zhu and et al., 2005] and other work has been defined distances depending of the adaptation models as in [Reynolds, 1998] which defines the distance:

$$d(X,Y) = \log \frac{\Pr(X|\text{UBM})}{\Pr(X|\boldsymbol{\theta}_{2_{\text{UBM}}})} + \log \frac{\Pr(Y|\text{UBM})}{\Pr(Y|\boldsymbol{\theta}_{1_{\text{UBM}}})} \tag{2.64}$$

where the speakers models $\boldsymbol{\theta}_{1_{\text{UBM}}}, \boldsymbol{\theta}_{1_{\text{UBM}}}$ are MAP-adapted through UBM model, see section 2.2.2.

Finally, some other work integrates segmentation with clustering by using model-based schemes and BIC for the stopping criterion as in [Ajmera and Wooters, 2003], [Wooters and et al., 2004] where after an initial segmentation iteratively decode the audio based on ML and adaptative GMM models. They also introduce the modeling of the temporal dynamics of the GMM observations by imposing a minimum duration of the speaker turn length.

- **Top-down clustering** Top-down algorithms were initially proposed by LIA [Meignier and et al., 2001], [Meignier *et al.*, 2006] as used in their latest system [Bozonnet *et al.*, 2010a]. In [Meignier and et al., 2001] and [Anguera and Hernando, 2004] an initial cluster is trained with all acoustic data available. Iterative decoding of MAP adapted models is performed where new clusters are split using a likelihood metric averaged over a window. They start from one cluster and iteratively split it until the stopping criterion is reached in the same way as previous systems. The top-down approach first models all data with a single speaker model and successively adds new models iteratively one-by-one, with interleaved Viterbi realignment and adaptation. Segments attributed to any one of these new models are marked as labeled. Stopping criteria similar to those employed in bottom-up systems may be used to terminate the process or it can continue until no more relevant unlabeled segments with which to train new speaker models remain. The use of this strategy has a low presence in the speaker clustering literature [Johnson and Woodland, 1998], [Tranter and Reynolds, 2004]. In [Johnson and Woodland, 1998] MLLR adaptation and the Arithmetic Harmonic Sphericity (AHS) [Bimbot and Mathan, 1993] are proposed. With AHS as:

$$D(X,Y) = \log\left[Tr(\sigma_Y \sigma_X^{-1}) \cdot Tr(\sigma_Y \sigma_X^{-1})\right] - 2\log(d) \tag{2.65}$$

- **Other clustering approaches**: There exist other systems that do not fit in the previous classification. The list includes a variety of techniques such as Vector Quantization [Mori and Nakagawa, 2001] which proposes the VQ distortion as a distance. Genetic algorithm are proposed in [Tsai and Wang, 2006] and Self-Organizing Maps (SOM) [Kohonen, 1990] are proposed for speaker clustering in [Lapidot, 2003]. Some newly approaches combine eigenvoices and speaker factors for segmentation and clustering in [Castaldo *et al.*, 2008] and spectral clustering [Keshet and Bengio, 2008] has been also applied to speaker diarization in [Iso, 2010] and [Ning *et al.*, 2010].

  A recent alternative approach is the Information Bottleneck (IB) principle. IB is inspired from rate-distortion theory [Cover and Thomas, 1991] and it is based on an information-theoretic framework. Work reported in [Vijayasenan *et al.*, 2007] and [Vijayasenan *et al.*, 2009] introduce two suitable methods: agglomerative information bottleneck (aIB) and sequential information bottleneck (sIB) both of them also bottom-up in nature. IB is completely non parametric and its results have been shown to be comparable to those of state-of-the-art parametric systems, with significant savings in computation. Clustering is based on minimizing the mutual information $I(X,C)$, where X represents the speech segment set at each iteration and C represents the clustering in the classes,i.e. speakers. At the same time, aIB tries to maximize mutual information $I(Y,C)$, which measures the mutual dependence of relevant variables Y and the clustering/partition C. Only a single global GMM is tuned for the full audio stream, in order to represent the relevant variable Y and to compute mutual information $I(Y,C)$ in new space of relevant variables defined by the GMM components. The approach aims at minimizing the mutual information between successive clusterings and the actual segment set X while preserving as much information as possible from the relevant variables Y, i.e. maximizing $I(Y,C)$. This corresponds to the maximization w.r.t the stochastic mapping $\Pr(C|X)$ of the objective function:

$$\mathcal{F} = I(Y,C) - \frac{1}{\beta} I(X,C), \tag{2.66}$$

  where β is a Lagrange multiplier.

  There are other approaches have become popular in speaker diarization by the end of this decade most of them based on Bayesian machine learning [Mackay, 2003]. One of them is Variational Bayes (VB) which refers to a set of methods, the most popular of which being the mean-field VB, that approximate the desired quantities (e.g. marginal likelihoods, posterior probabilities, predictive densities) by bounding the marginal likelihood of the model from below. The use of VB in speaker diarization has been pioneered by [Valente and Wellekens, 2004] and has been refined in [Valente *et al.*, 2010] and applied to i-vectors in [Kenny *et al.*, 2010]. In [Fox *et al.*, 2011] describes a Bayesian non-parametric approach to speaker diarization that builds on the hierarchical Dirichlet process hidden Markov model (HDP-HMM)

[Teh *et al.*, 2004]. VB is a general purpose (approximate) inference method and its use is not limited to finite mixture models, on the contrary, it can be applied to nonparametric models, too (e.g. Dirichlet Process Mixture Models. In [Kenny and Castaldo, 2009] VB is combined successfully with eigenvoice modeling, described in [Kenny, 2008], for the speaker diarization of telephone conversations. However these systems still consider classical Viterbi decoding for the classification. A full review of Bayesian methods applied to diarization can be found in [Stafylakis1 and Katsouros, 2011].

2.3.3 Evaluation Metrics

Regarding diarization and tracking tasks, the main metric used to evaluate either systems or algorithms is the Diarization Error Rate (DER).

2.3.3.1 Diarization Error Rate

The main metric that is used for speaker diarization experiments is the Diarization Error Rate (DER) as described and used by NIST in the RT evaluations (NIST Fall Rich Transcription on meetings 2006 Evaluation Plan 2006). It is measured as the fraction of time that is not attributed correctly to a speaker or to non-speech. To measure it, a script names MD-eval-v12.pl (NIST MD-eval-v21 DER evaluation script 2006), developed by NIST, was used. As per the definition of the task, the system hypothesis diarization output does not need to identify the speakers by name or definite ID, therefore the ID tags assigned to the speakers in both the hypothesis and the reference segmentation do not need to be the same. This is unlike the non-speech tags, which are marked as non labeled gaps between two speaker segments, and therefore do implicitly need to be identified. The evaluation script first does an optimum one-to-one mapping of all speaker label ID between hypothesis and reference files. This allows the scoring of different ID tags between the two files. The Diarization Error Rate score is computed as

$$\text{DER} = \sum_{s=1}^{S} \text{dur}(s) \cdot (\max(N_{\text{ref}}(s), N_{\text{hyp}}(s)) - N_{\text{correct}}(s)) \tag{2.67}$$

where S is the total number of speaker segments where both reference and hypothesis files contain the same speaker/s pair/s. It is obtained by collapsing together the hypothesis and reference speaker turns. The terms $N_{\text{ref}}(s)$ and $N_{\text{sys}}(s)$ indicate the number of speaker speaking in segment s, and $N_{\text{correct}(s)}$ indicates the number of speakers that speak in segment s and have been correctly matched between reference and hypothesis. Segments labeled as non-speech are considered to contain 0 speakers. When all speakers/non-speech in a segment are correctly matched the error for that segment is 0. The DER error can be decomposed into the errors coming from the different sources, which are:

- Speaker error: percentage of scored time that a speaker ID is assigned to the wrong speaker. This type of error does not account for speakers in overlap not detected or any error coming from non-speech frames. It can be written as,

$$\mathrm{E}_{Spkr} = \frac{\sum_{s=1}^{S} \mathrm{dur}(s) \cdot (\min(N_{\mathrm{ref}}(s), N_{\mathrm{hyp}}(s)) - N_{\mathrm{correct}}(s))}{T_{\mathrm{score}}} \tag{2.68}$$

where $T\mathrm{score} = \sum_{s}^{S} \mathrm{dur}(s) \cdot N_{\mathrm{ref}}(s)$, is the total scoring time, in the denominator in equation 2.67.

- False alarm speech: percentage of scored time that a hypothesized speaker is labeled as a non-speech in the reference. It can be formulated as,

$$\mathrm{E}_{\mathrm{FA}} = \frac{\sum_{s=1}^{S} \mathrm{dur}(s) \dot{(} N_{\mathrm{hyp}}(s) - N_{\mathrm{ref}}(s) \dot{)}}{T_{\mathrm{score}}} \quad \forall \big(N_{\mathrm{hyp}}(s) - N_{\mathrm{ref}}(s) \big) > 0 \tag{2.69}$$

  computed only over segments where the reference segment is labeled as non-speech.

- Missed speech: percentage of scored time that a hypothesized non-speech segment corresponds to a reference speaker segment. It can be expressed as,

$$\mathrm{E}_{\mathrm{MISS}} = \frac{\sum_{s=1}^{S} \mathrm{dur}(s) \dot{(} N_{\mathrm{ref}}(s) - N_{\mathrm{hyp}}(s) \dot{)}}{T_{\mathrm{score}}} \quad \forall \big(N_{\mathrm{ref}}(s) - N_{\mathrm{hyp}}(s) \big) > 0 \tag{2.70}$$

  computed only over segments where the hypothesis segment is labeled as non-speech.

- Overlap speaker: percentage of scored time that some of the multiple speakers in a segment do not get assigned to any speaker. This errors usually fuses either into the $\mathrm{E}_{\mathrm{MISS}}$ or E_{FA}, depending on whether it is the reference or the hypothesis containing non assigned speakers. If multiple speakers appear in both the reference and the hypothesis the error produced belongs to E_{Spkr}.

Given all possible errors one can rewrite equation 2.67 as

$$\mathrm{DER} = \mathrm{E}_{\mathrm{Spkr}} + \mathrm{E}_{\mathrm{MISS}} + \mathrm{E}_{\mathrm{FA}} + \mathrm{E}_{\mathrm{overlap}} \tag{2.71}$$

When evaluating performance, a collar around every reference speaker turn can be defined which accounts for inexactitudes in the labeling of the data. It was estimated by NIST that a 250ms collar could account for all these differences. When there is people overlapping each other in the recording it is stated so in the reference file, with as many as 5 speaker turns being assigned to the same time instant. As pointed out in the denominator in equation 2.67, the total evaluated time includes the overlaps. Errors produced when the system does not detect any or some of the multiple speakers in overlap count as missed speaker errors. Once

the performance is obtained for each individual meeting excerpt, the time weighted average is done among all meetings in a given set to obtain an overall average score. The scored time is the one used for such weighting, as it indicates the total (overlapped speaker included) time that has been evaluated in each excerpt.

2.4 Speaker Information Fusion

All those works that address the speaker recognition issue by fusing several cues of information, such as multi-modality or multi-sources approaches has a common goal: search for getting all the useful information from the environment and, of course, the fusion of such information in an intelligent manner. It is the inevitable direction driven by the fact that the more knowledge the recognition system has, the more the performance it obtains.

2.4.1 Multi-microphone approaches

Speaker diarization and tracking using multi channel information have been addressed in recent work and it have been proved to be useful [Pardo *et al.*, 2007; Pardo *et al.*, 2012] and [Anguera *et al.*, 2007a]. In a smart-room environment multiple microphones are usually available for processing and their use can aid the global identification system. As example, it is clear that the position of the speaker across the time is a feature that can help to discriminate between different speakers since same spatial position can not be shared for different speakers. Estimates of inter-channel delay may be used not only for delay-and-sum beamforming of multiple microphone channels, but also for speaker localization. If we assume that speakers do not move, or there exists a tracking algorithm of their positions, then estimates of speaker location may thus be used as alternative features, which have nowadays become extremely popular. In this line, the computation of the time delay of arrival (TDOA) or the direction of arrival (DOA) between channels are the techniques most cited to estimate speaker location. The combination of the MFCC features and the TDOA have reported promising results in [Ajmera and et al., 2004], [Pardo *et al.*, 2007] and [Barra-Chicote *et al.*, 2011].
The Weighted-Delay-and-Sum (W-D&S) technique [Flanagan *et al.*, 1985] is one of the simplest beamforming techniques but still gives a very good performance. It is based on the fact that applying different phase weights to the input channels the main lobe of the directivity pattern can be steered to a desired location, where the acoustic input comes from. It differs from the simpler D&S beamformer in that an independent weight is applied to each of the channels before summing them. The principle of operation of W-D&S can be seen in figure 2.25. If we assume the distance between the speech source and the microphones is far enough we can hypothesize that the speech wave arriving to each microphone is flat. Therefore, the difference between the input signals, only taking into account the wave path and without taking care about channel distortion, is a time delay of arrival due the different positions of the microphones with regard to the source. So if we estimate the time τ, see figure 2.25, we could synchronize two different input signal in order to enhance the speaker information and reduce the additive white noise.

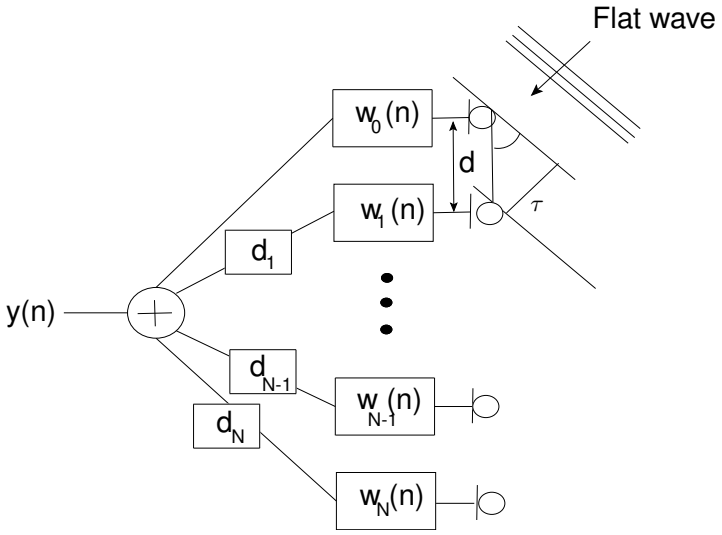


Figure 2.25: *Weighted-Delay-and-Sum algorithm block diagram*

Given the signals delivered by N microphones, $x_i[n]$ with $i = 0 \dots N-1$ (where n indicates time steps) if we know their individual relative delays $d(0,i)$ (called Time Delay of Arrival, TDOA) with respect to a common reference microphone x_0 , we can obtain the enhanced signal as follows,

$$y(n) = x_0[n] + \sum_{i=1}^{N-1} W_i x_i[n - d(0,i)]. \tag{2.72}$$

By adding together the aligned signals the usable speech adds together and the ambient noise (assuming it is random and has a similar probability function) will be reduced. Using D&S, according to [Flanagan *et al.*, 1985], we can obtain up to a 3dB SNR improvement each time that we double the number of microphones. In order to deal with acoustic conditions into a meeting room, like as reverberation or noise due far-field microphone conditions, the generalized cross correlation with phase transform (GCC-PHAT) method [Knapp and Carter, 1976]; [T. Gustafsson and B. Rao and M. Trivedi, 2003] has reported robust performance. Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as follows,

$$\hat{G}_{PHAT_{ij}}(f) = \frac{X_i(f)\left[X_j(f)\right]^*}{\left|X_i(f)\left[X_j(f)\right]^*\right|}, \tag{2.73}$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals and $[]^*$ denotes the complex conjugate. The TDOA for two microphones is estimated as:

$$\hat{d}_{PHAT_{ij}} = \arg\max_d \hat{R}_{PHAT}(d_{ij}) \tag{2.74}$$

where $\hat{R}_{PHAT_{ij}}(d)$ is the inverse Fourier transform of $\hat{G}_{PHAT_{ij}}(f)$, the Fourier Transform of the estimated cross correlation phase. The end-best maximum value of $\hat{R}_{PHAT_{ij}}(d)$ are computed for each frame and the best sequence across all the meeting is finally selected for the beamforming according to a continuity maximization algorithm.

TDOA features, without relying on the room setup i.e. without knowledge about the microphone locations, are usually mixed, in the context of NIST evaluation, with MFCC features at weighted log-likelihood level. Several evaluation systems make use of those two feature streams as in [Pardo *et al.*, 2007; Wooters and Huygbregts, 2008; van Leeuwen and Konečný, 2008; Friedland *et al.*, 2011; Barra-Chicote *et al.*, 2011] where TDOA-MFCC combination showed to be very effective in reducing the diarization error. Recent work in [Vijayasenan *et al.*, 2011] investigates the integration of the Modulation Spectrum and the frequency domain linear prediction (FDLP) features together with MFCC and TDOA in the space of relevance variables of the agglomerative information bottleneck framework.

For the segmentation task a speaker tracking approach is proposed in [Lathoud and et al., 2002] using only between channel differences. [Koh and et al., 2008] submitted a novel diarization system to NIST RT 2007 and 2009 Evaluation with excellent results. It was based on a first segmentation and clustering by means DOA features and followed a spectral stage to ensure purity in the obtained segments. Also, the RT Evaluations have shown the common use of speech enhancement techniques. In the most recent NIST RT evaluation, in 2009, most of participants used estimates of inter-channel delay both for beamforming and as features. The success of these systems in NIST RT evaluations would seem to support their use.

Other implementations are focused on the source separation handling with the overlapping speech issue, an inherent characteristic of conversational speech. In [Anliker and et al., 2006] a two steps strategy was presented for a two-microphone system in a mobile environment. The proposed system makes use of the feedback between a blind source separation algorithm based on the degenerate unmixing estimation technique (DUET) [Yilmaz and Rickard, 2004] and a speaker tracking algorithm based on various distances.

Application to Overlap Detection

Speaker overlap is a commonly occurring event in human conversation. Shriberg *et al.* in [Shriberg *et al.*, 2001] reported this phenomenon is frequently observable not only in meetings including several people, but also in telephone dialogues. Overlaps can stem from various situations. For instance, in meeting domain, listeners sometimes try to interrupt the speaker in order to grab floor or encourage his talk with backchannel sounds or words, e.g some interjections such as ”uh-huh” and ”mhm” (meaning yes) and ”uh-uh” (meaning no), the word ”uhhuh,” which serves both to signal agreement and a back-channel ”I’m listening,’. Some overlaps, obviously, originate accidentally. Their amount is related with the spontaneity and formal nature of the discourse.

Overlapping speech has been identified as one of the main challenges for automatic human language technologies [Shriberg, 2005], speaker diarization being no exception. The regions where more than one speaker is active, missed speech errors will be incurred and, given the high performance of some state-of-the-art systems,

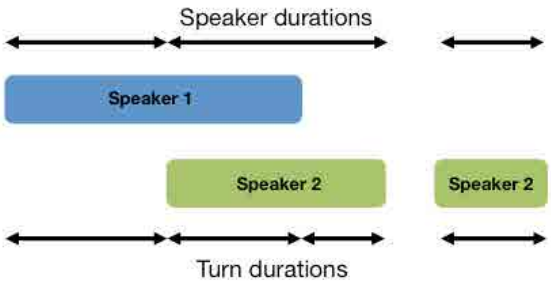


Figure 2.26: *Examples of turn and speaker durations in the presence of overlapped speech and silences.*

this can be a substantial fraction of the overall diarization error. Such a common drawback of conventional diarization systems is mainly due most systems are only able to assign one speaker label per segment. Hence in cases when a segment contains simultaneous speech, this implicitly leads to missed speech errors. In addition, the effect of overlapped speech in diarization degrades speaker clustering and modeling [Otterson and Ostendorf, 2007] since speaker models could be corrupted if simultaneous speech is included into their training data, also known as cluster purity.

Several algorithms have been published in the literature that aims to detect overlap speech as a result of multi-speaker speech activity detection on personal close-talking microphones [Pfau *et al.*, 2001; Laskowski *et al.*, 2004; Wrigley *et al.*, 2005; Laskowski and Schultz, 2006; Nwe *et al.*, 2008]. In order to deal with crosstalk,[Pfau *et al.*, 2001] applied cross-correlation analysis to detect speaker overlaps, and the system presented in [Laskowski *et al.*, 2004] used joint maximum cross-correlation exclusively. [Wrigley *et al.*, 2005] proposes a speech overlap model included in segmentation step into an ergodic hidden Markov model (HMM) framework and discussed a large set of suitable features. Among the most discriminating features were kurtosis and cross-correlation metrics.

The speaker diarization approach discussed in [van Leeuwen and Huijbregts, 2006] employed an initial diarization output to select training data for building a combined Gaussian mixture model (GMM) for every pair of previously detected speakers. These new speaker-pair GMMs were then integrated within the original single-speaker models into a new HMM and the meeting data was resegmented again. Even though some overlap was detected in this way, it did not lead to a reduction of diarization error. On the other hand, assuming that speaker overlap is likely to occur around speaker-turn points, the training of an ”ad hoc” overlap model in one diarization pass and its application for the second pass improved results on the NIST RT data [Huijbregts *et al.*, 2009]. The HMM-based system presented in [Boakye *et al.*, 2008a] and [Boakye *et al.*, 2008c] utilized various spectral features (cepstrum, entropy, modulation spectrogram, etc.) for detecting overlaps on single distant channel. Detected overlapping speech was subsequently processed in speaker diarization.

A few algorithms for speaker overlap detection make use of time delays between microphone-pairs. [G.

Lathoud and I.A. McCowan, 2003] suggested to segment the audio according to speakers using microphone-pair time delays and showed the possibility to detect two simultaneous talkers by modeling short-term turns for each speaker combination. In [Luque *et al.*, 2008b], TDOAs features are employed to segment speakers aiming to obtain an initial segmentation with highly stable speakers positions along time. Such a segmentation provides the starting point for a classical ergodic-HMM clustering. In the same way, they suggests the ability of TDOAs to detect overlapped speech.
In [Zelenák *et al.*, 2011], the authors propose the use of more spatial information in the overlap detection stage previous to the speaker diarization system itself. Spatial features are extracted from the Generalized Cross-Correlation with Phase Transform weighting (GCC-PHAT) [T. Gustafsson and B. Rao and M. Trivedi, 2003] from all channel pairs, and consist of the main peak magnitude of the cross-correlation, the rate of change of the TDOA and a dispersion ratio that measures the energy dispersed in the neighborhood of the main peak in the GCC-PHAT [Zelenák *et al.*, 2010]. Features obtained per each microphone-pair are projected by means a PCA and compared with a multilayer perceptron (MLP) fusion, whose output classification score is used as an extra spatial feature.

2.4.2 AudioVisual Diarization

There exists a variety of work which describes systems performing localization and identification fusing several cues such as audio and video sources. In [Bernardin and Stiefelhagen, 2007] an audio-visual algorithm is described. It fuses the information coming asynchronously from the sources at the higher level. The speaker ID are accumulated across the time and they are associated with the speaker localization creating a dynamic histogram of the confidence of the cues. In this manner they get increased the system's global confidence in the set of target identities. Busso et al. [Busso and et al., 2005] presented a smart meeting room application by which the location of the participants is extracted from audio and video recordings. They are fused to an overall location estimation. The microphone array is steered towards the estimated position by beamforming techniques and the speaking identity are obtained from the steered audio signal. In the speaker identification and verification tasks a great variety of work related the audio-video approach can also be found. They fuses several classes of features as lips features, cross-modal features, voice spectral, etc. at different level and they use different techniques for this purpose such as DBN, Evolving Connectionist Systems (ECOS) [Kasabov, 1973], Neural Networks (NN), Particle and Kalman Filters. Some examples in [Zhang and et al., 2004], [Whu and et al., 2005], [Chetty and Wagner, 2007]
Focusing on speaker diarization and tracking, the work in [Noulas and Krose, 2007] address the on-line diarization problem by means a system based on audio and video modalities. The Dynamic Bayesian Networks (DBN), an extension of factorial hidden Markov models (fHMM) [Ghahramani and Jordan, 1997], a generalization of the HMM models, and mutual information (MI) [Peng *et al.*, 2005] between the audio-visual cues are applied, creating evolved complexity models as more data become available. Another use of DBN is proposed in [Noulas *et al.*, 2011] as an audiovisual framework. The factorial HMM arises by forming a dynamic Bayesian belief network composed of several layers. Each of the layers has independent dynamics

but the final observation vector depends upon the state in each of the layers.
In [Salah *et al.*, 2008a] the authors implemented a person tracking system within a smart-room environment at UPC equipped with various cameras and microphone sensors and in which the participants are able to move around freely in the room. Several technologies are employed. For visual-based identification, face recognition, based on the work in [Luque *et al.*, 2006b], is used to identify persons initially as they enter into the smart room and a visual tracking method based on probabilistic occupancy maps [Fleuret *et al.*, 2008] is extended and adapted to the experimental setting at the room. For audio-based localization and identification, algorithms based on GCC-PHAT and AHC with HMM [Luque and Hernando, 2008a] are employed. The information of audio and video sources is also effectively combined, when acoustic information is available, employing Particle Filtering (PF) [Gordon *et al.*, 1993] [Carpenter *et al.*, 1997] strategies for active speaker tracking [Nickel *et al.*, 2005b] or audiovisual multi-person tracking [Gatica-Perez *et al.*, 2007].
Anyway, few articles discus joint audiovisual diarization. The work in [Friedland *et al.*, 2009a] for multiple-camera and [Friedland *et al.*, 2009b] using only a single, low-resolution overview camera has been first attempts to deal with joint audiovisual diarization. The algorithm relies on very few assumptions and is able to cope with an arbitrary amount of cameras and subframes. Most importantly, as a result of training a combined audiovisual model, the authors found that speaker diarization algorithms can result in speaker localization as side information. This way joint audiovisual speaker diarization can answer the question who spoken when and from where.

2.4.3 Multi-decision approaches

Like in other pattern classification tasks, combining information from multiple sources of evidence, a technique called fusion, has been widely applied in speaker recognition. System or component combination is often reported in the literature as an effective means for improving performance in many speech processing applications [Chen *et al.*, 1997; Slomka *et al.*, 1998; Rodrguez-Liares *et al.*, 2003; Farrús *et al.*, 2006; Hernando *et al.*, 2006]
In speaker identification and verification, typically, a number of different feature sets are first extracted from the speech signal; then an individual classifier is used for each feature set, note that several models per speaker are stored for such purpose; following that the sub-scores or decisions are combined. It is also possible to obtain fusion through modeling the same features using different classifier architectures, feature normalizations, or training sets [Brummer *et al.*, 2007]. A general belief is that successful fusion system should combine as independent features as possible low-level spectral features, prosodic features and high-level features. But improvement can also be obtained by fusion of different low-level spectral features (e.g. MFCCs and LPCCs) and different classifiers for them [Brummer *et al.*, 2007; Campbell *et al.*, 2006b]. Fusing dependent (correlated) classifiers can enhance the robustness of the score due to variance reduction [Poh and Bengio, 2004].
Recently, some improvements to fusion methodology have been achieved by integrating auxiliary side information, also known as quality measures, into the fusion process [Ferrer *et al.*, 2008; Kryszczuk *et al.*,

2007]. Unlike the traditional methods where the fusion system is trained on development data and kept fixed during run-time, the idea in side-information fusion is to adapt the fusion on each test case. Signal-to-noise ratio (SNR) [Ferrer *et al.*, 2008; Kryszczuk *et al.*, 2007] and nonnativeness score of the test segment [Ferrer *et al.*, 2008] have been used as the auxiliary side information, for instance.

A theoretically elegant technique for optimizing the fusion weights based on logistic regression has been proposed in [Brümmer and Preez, 2006; Brummer *et al.*, 2007]. An implementation of the method is available in the Fusion and Calibration [Brummer, 2005]. By considering outputs from the different classifiers as another random variable, score vector, a "backend" classifier can be built on top of the individual classifiers [Abad *et al.*, 2011; Abad *et al.*, 2010].

In addition to this, in the speaker diarization and tracking tasks, very few studies related to speaker diarization have been reported in recent years. Some of the combination strategies proposed consist of applying different algorithms/components sequentially, based on the segmentation outputs of the previous steps in order to refine boundaries (referred to as hybridization or piped systems in [Meignier *et al.*, 2006]. In [Vijayasenan *et al.*, 2008] the authors combine two different algorithms, aIB and sIB based on the Information Bottleneck framework. In [Bozonnet *et al.*, 2010b], the best components of two different speaker diarization systems implemented by two different French laboratories (LIUM and IRIT) are merged and/or used sequentially, which leads to a performance gain compared to results from individual systems. An original approach is proposed in [Gupta *et al.*, 2007], based on a real system combination. Here, a couple of systems uniquely differentiated by their input features (parametrization based on Gaussianized against non-Gaussianized MFCCs) are combined for the speaker diarization of phone calls conversations.

Part II

Speaker Identification and Verification

Chapter 3

Speaker Identification in Meetings: The CHIL Project and CLEAR Evaluations

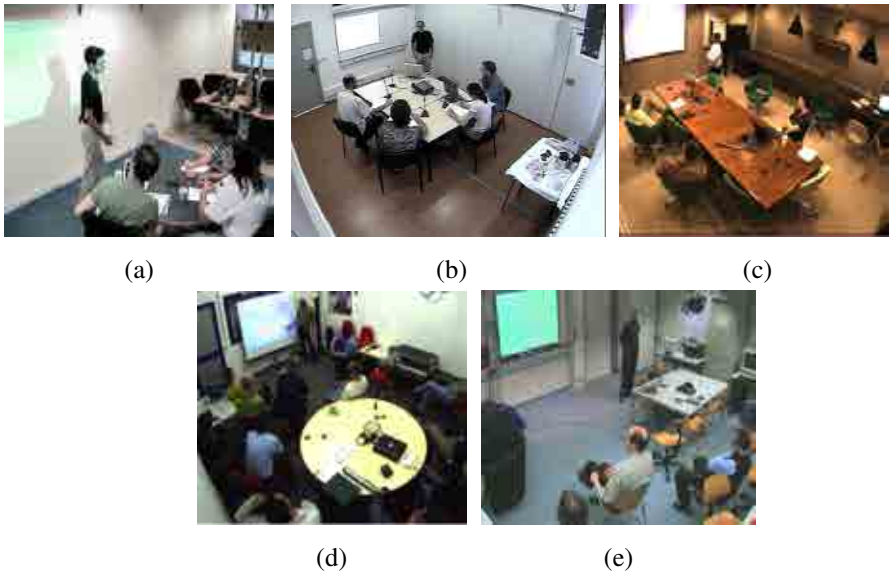
This chapter covers the description of the UPC person identification systems submitted to the CLEAR 2006 and 2007 evaluations [Luque *et al.*, 2006b; Luque and Hernando, 2008a]. In addition, it also presents the work developed on technology demonstrations at the UPC meeting room concerned to speaker recognition, inside the framework of the CHIL (Computers in the Human Interaction Loop) project. The CHIL project [Casas and Stiefelhagen, 2005] is an Integrated Project (IP 506909) funded by the European Union under its 6th framework program. The project started on January 1st, 2004 and had a duration of four years. Rather than requiring user attention to operate machines, CHIL services attempt to understand human activities and interactions to provide helpful services, aiming to radically change the way we use computers. Instead than expecting a human attending to technology, CHIL attempts to develop computer assistants that attend to human activities and interactions or even assess human's intentions. To achieve this goal, machines must understand the human context and activities likewise human being does. Computers must adapt to and learn from the humans' interests, activities, goals and aspirations. This requires machines to better perceive and understand all the human communication signals including speech, facial expressions, attention, emotion, gestures, and many more. The team sets out to study the technical, social and ethical questions that will enable this next generation of computing in a responsible manner. The CHIL results were disseminated and made available to a wide community of interested researchers.

CLEAR was meant to bring together projects and researchers working on related technologies in order to establish a common international evaluation framework in this field. CLEAR evaluations were supported by the European Integrated project CHIL and the US National Institute of Standards and Technology (NIST). Spring 2006 and 2007 CLEAR evaluations and workshops were an international effort to evaluate systems designed

from the team members of the CHIL consortium to recognize events, activities, and their relationships in interaction scenarios. The CLEAR evaluations consist on a set of audiovisual recordings which were collected between years 2004 to 2006. The database is composed of several speakers in various smart rooms from partners of the CHIL consortium. The tentative tasks to be addressed in CLEAR included the following:

- Person Tracking (2D and 3D, audio-only, video-only, multimodal)
- Face Tracking
- Vehicle Tracking
- Person Identification (audio-only, video-only, multimodal)
- Head Pose Estimation (2D, 3D)
- Acoustic Event Detection and Classification

Specifically, CLEAR PID (Person Identification) evaluations campaign were designed to study issues causing important degradations in highly interactive scenarios like occurs in meetings or conferences. One of the main focus along evaluations was the identification errors due the amount of speaker data available for training speaker models as a function of available testing data. In most of the real situations, there not exists enough data available to compute an accurate estimation of the person model. In such a situation the performance degradation is a common feature of most of the systems. Robustness against that issue has become a "tour de force" in the person identification community. For instance, if available data for testing reduces from 5 seconds to 1 second, the systems show a big drop in correct identification rates whilst a human being could recognize voices with high accuracy just by half of a second .
Far-field sensor conditions of CLEAR evaluations also bring to researchers the chance of study another of the common problems of performance degradation. CLEAR database is composed of a set of audiovisual recordings coming from a wide variety of sensors: head-mounted and wall-mounted microphones, microphone arrays, table-top microphones, wall-mounted cameras and pan-tilt-zoom cameras are some examples.
On one hand, such a quantity of sensors give to researchers an opportunity to study a great variety of issues. For instance, the systems performance due to signal degradation by acquisition in multi-path propagation in audio modality; and due occlusion or tiny images in image modality, respectively. On the other hand, studying the benefit of the redundant information from multiple input sources brings to them a wide field to explore more complexes and ambitious approaches. In such a situation, the system implementation could be improved by means of multi-microphone processing techniques to deal with channel and noise distortion or, through multimodal approaches, fusing video and audio modalities.
Furthermore, CLEAR data was collected by different CHIL partners from several smart-rooms. It also brings an opportunity to study system performance degradation imputable to different room conditions or, taking into account the international nature of the evaluation, due the English speaker accent variability among sites.



(a) (b) (c)

(d) (e)

Figure 3.1: *Sample camera recordings from the different smart-room of the partners of the CHIL project. Seminars (a) AIT, Greece, (b) UPC, Spain and (c) IBM, USA. All of them corresponds to interactive seminars. (d) AIT-irst, Italy and (e) UKA, Germany which corresponds to non interactive seminars where the lecture monopolizes the session.*

The UPC person identification systems presented to CLEAR evaluations were composed of different approaches. They include a single microphone identification system, the use of combined microphone inputs by means signal enhancement or multi-decision approaches and, finally, the use of audiovisual information through the integration of audio and video data in a multimodal system.

3.1 The CHIL project: CLEAR Evaluations in Speaker Identification

The project CHIL [chil, 2006] is an Integrated Project (IP 506909) funded by the European Union under its 6th framework program. The project started on January 1st, 2004 and had a planned duration of four years. The CHIL team is a consortium of internationally renowned research labs in Europe and the US, who collaborate to bring friendlier and more helpful computing services to society. The CHIL team is a consortium of internationally renowned research labs in Europe and the US, who collaborate to bring friendlier and more helpful computing services to society. The research consortium includes 15 leading research laboratories from 9 countries representing today's state of the art in multimodal and perceptual user interface technologies in European Union and the US.

Requiring user attention to operate machines is the usual way we actually interact with computers. Nonetheless, CHIL services attempt to understand human activities and interactions to provide helpful services implicitly and unobtrusively. Considerable human attention is expended in operating and attending to computers, and humans are forced to spend precious time on fighting technological artifacts, rather than on human interaction

and communication. CHIL aims to radically change the way we use computers. Rather than expecting a human to attend to technology, CHIL attempts to develop computer assistants that attend to human activities, interactions, and intentions. Instead of reacting only to explicit user requests, such assistants pro actively provide services by observing the implicit human request or need, much like a personal butler would. To achieve this goal, machines must understand the human context and activities better; they must adapt to and learn from the humans' interests, activities, goals and aspirations. This requires machines to better perceive and understand all the human communication signals including speech, facial expressions, attention, emotion, gestures, and many more.

Inside the framework of CHIL project, the UPC has built the smart-room, a room equipped with multiple cameras and microphones, with the purpose of investigating the video and audio perception of the computer systems. In a smart room, the typical situation is to have one or more cameras and several microphones. UPC's smart-room is an intelligent space designed as a meeting-room with a table in the center and chairs around it. The configuration of the UPC's smart-room is depicted in figure A.1 and a more accurate description is given in appendix A. Among others, there are several audio-visual sensors (cameras and microphones), synchronization and acquisition equipments, working computers, and a video projector. The smart-room is the indispensable installation for the UPC research groups that work on multimodal interfaces. The acquired audio-visual signals allow both developing the technologies of audio and video analysis, and making demonstration of the technology that can offer specific services in the configuration of meeting rooms or teaching rooms.

Perceptually aware interfaces can gather relevant information to recognize, model and interpret human activity, behavior, actions and intentions. The main goal is to give to the computer systems awareness of the activity that is going on in the room and to interact with human beings as reaction to a request or interactions among them. in conclusion, if computers know the environment they can interact with us in the same manner we interact with each other. The speech-related technologies like speech and speaker recognition are part of the fundamentals of the analysis of the human activity in the smart-rooms. At present, robust speech recognition systems, that use a signal from a far-field microphone, are investigated in order to avoid bothering people wear cables or close-talk microphones. On the other hand, the video technologies analyze the presence, localization and movements of the peoples, face recognition, gesture detection, postures and attention tracking, in order to classify the events, activities, and relationships. The detection technologies, classification and recognition based on multiple sensors, like audio and visual localization, person identification based on speech and face, activity detection based on acoustics or images, can increase the robustness of existing systems.

In order to be in accordance with ambitious goals of CHIL, a series of international technology evaluations were supported by the CHIL project. Computer perception and awareness is achieved by means perceptual components based on state-of-the-art technologies. Aiming to evaluate performance of such technologies, several evaluation campaigns were conducted during CHIL project. Among them, person identification evaluations are of our interest inside the scope of this PhD thesis.

CLEAR data evaluation is composed of a set of audiovisual recordings of seminars and of highly-interactive small working-group. These recordings were collected by the CHIL consortium for the CLEAR 2006 and

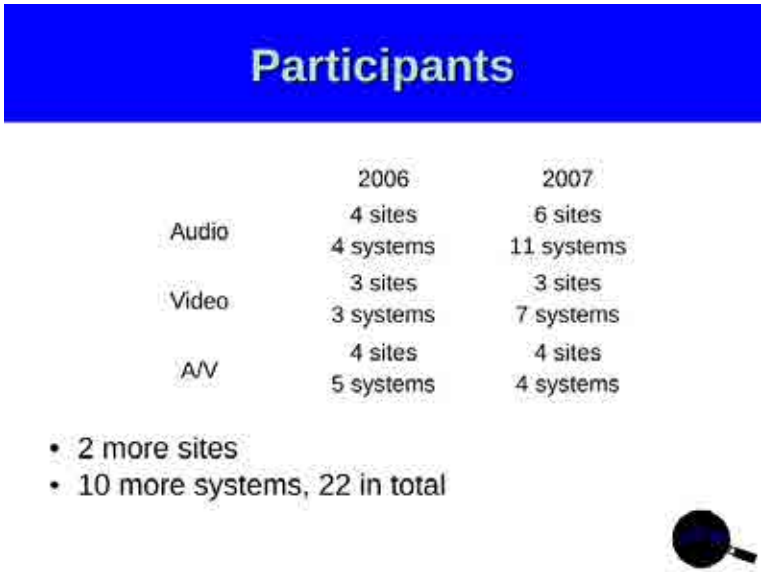


Figure 3.2: *CLEAR participants in both 2006 and 2007 person identification evaluation campaigns.*

2007 evaluations. The recordings were done according to the "CHIL Room Setup" specification [Casas and Stiefelhagen, 2005] matching features as far-field conditions. A complete recordings description can be found in [Moreau *et al.*, 2008; Mostefa and et al., 2006; Mostefa and et al., 2007].

In order to evaluate how the duration of the training data has effect upon the performance of the system, two training conditions have been considered: 15 and 30 seconds of training segment duration. Furthermore, test segments of different durations: $1, 2, 5, 10$ and 20 seconds, were employed during algorithm development and testing stage. For CLEAR 2006 evaluation, a total of 26 personal identities were used to compile the recognition experiments. In CLEAR 2007 database, such a number reached a total of 28 personal identities and 108 identification trials per speaker (of different durations) were evaluated. The person identification (PID) task consists in determining the identity of a person by means speech segments and/or video segments. It was assumed that all the possible speakers are known, i.e. a classical "closed-set" person identification task. Far-field conditions have been considered for both modalities, i.e. corner cameras for video and Mark III microphone array for audio. Audio recordings are composed of several hammerfall channels in WAV or RAW format sampled at 44kHz with 24 or 16 resolution bits and 64 channels from MarkIII array in Smartflow format [smartflow, 2002], sampled at 44kHz with 24 resolution bits. From each seminar, belonging to the CLEAR database, just one audio signal from microphone number 4 of the Mark III array is selected to benchmark systems' implementation based on single channel whereas rest of channels are allowed to be used for development purposes. Each audio signal is manually split into excerpts which theoretically contain information about just one speaker. These excerpts are merged to form the final testing segments of 1, 5, 10 and 20 seconds (see 3.2) and training segments of 15 and 30 seconds.

For video modality, video is composed of 4 or 5 video sequences. It is recorded in compressed JPEG format,

	Number of segments			
	CLEAR PID 2006		CLEAR PID 2007	
Segment Duration	Development	Evaluation	Development	Evaluation
1 sec	390	613	560	2240
5 sec	182	0	112	448
10 sec	78	411	56	224
20 sec	26	178	28	112
Total	676	1202	756	3024

Table 3.1: *Number of segments composing the data employed in algorithm development as well as for each test condition in CLEAR 2006 and 2007 PID evaluation, respectively.*

with frame-rates of 15, 25 and 30 frames per second and at different resolutions depending on the recording. In video task, four fixed position cameras are continuously monitoring the scene. All frames in the 1/5/10/20 seconds segments and all synchronous camera views can be used jointly. Furthermore, the various cameras' information can be fused with other modalities to find out the identity of the concerned person. In order to search faces to be identified, a set of labels is provided along with the position of the bounding box per each person's face in the scene. These labels are supplied each 1 second. The face bounding boxes are linearly interpolated to estimate their position in intermediate frames. To help this process, an extra set of labels is provided, giving the position of both eyes of each individual each 200 ms.

3.2 The UPC Speaker Identification System

Different approaches are described in this section. Audio, video and multimodal person identification techniques are described and the official results obtained in CLEAR 2006 and 2007 evaluation campaigns are also reported. The CLEAR person identification evaluation is a closed-set task, that is, all the possible speakers are known "a priori". In addition, matched training and testing conditions and far-field data acquisition are assumed, as well as no "a priori" knowledge about room environment.

For audio person identification, the mono-microphone algorithm is based on a short-term estimation of the speech spectrum using Frequency Filtering (FF) parameters over the filter bank energies, described in [Nadeu *et al.*, 1997]. Such parameters are modeled by Gaussian Mixture Models (GMM) [Reynolds, 1995] with diagonal covariance matrix. This approach is used as a baseline system and we will refer it to as: Single Distant Microphone (SDM) approach.

In addition, two multi-microphone approaches are studied that try to take advantage of the information diversity. The former applies a Delay and Sum [Flanagan *et al.*, 1985] algorithm with the purpose to obtain an enhanced and noise filtered version of the speech wave. The latter profits the multi channel diversity fusing, through a voting scheme, three identical SDM classifiers.

In the case of visual identification, an appearance-based technique is used as response to the low quality of images. Face images of the same individual are gathered into groups. Frontal images within a group are jointly compared to the models for identification. These models are composed of several images representative of the individual. The joint recognition enhances the performance of a face recognition algorithm applied on single images. Individual decisions are based on a PCA [Kirby and L.Sirovic, 1990] approach given that the variability of the users' appearance is assumed to be low and so are the lighting variations.
Multimodal recognition involves the combination of two or more human traits like voice, face, fingerprints, iris, hand geometry, etc. to achieve better performance than using monomodal recognition [Bolle and et al., 2004], [R.Brunelli and Falavigna, 1995]. In the CLEAR approach submitted, a multimodal score fusion technique, Matcher Weighting with equalized scores, was applied improving the correct identification rate in most of the evaluation conditions.

3.2.1 Audio Person Identification

Below we describe the main features of the UPC acoustic speaker identification system. The three audio approaches submitted to CLEAR 2006 and 2007 evaluations have in common the same characteristics about parametrization and speaker modeling, but they differ in the way how they take benefit of the multi-microphone information. Firstly, the single distant microphone system (SDM) approach is summarized. In next subsection, the signal enhanced approach is described in which beamforming technique is applied on a set of audio channels to obtain an improved version of the signal to apply in the identification process. Finally, we describe a fusion scheme composed of three SDM classifiers and a simple fusion decision rule in order to judge the person identity.
Following speech wave processing is shared among the three approaches: The audio data provided is decimated from 44.1KHz to 16KHz sampling rate. The audio is analyzed in frames of 30 milliseconds at a rate of 10 milliseconds. Then, each frame window is processed subtracting the mean amplitude (Cepstral Mean Subtraction) and no pre-emphasis filter is applied. A Hamming window is applied to each frame and a short-time frequency estimation based on FFT is computed. Finally, the FFT amplitudes are averaged in a 30 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. All test are conducted in matching conditions.

Baseline: Single Distant Microphone System

The SDM approach is based on a short-term estimation of the spectrum energy in several sub-bands. The scheme we present follow the classical procedure applied to obtain the Mel-Frequency Cepstral Coefficients (MFCC), see section 2.2.1, however in this approach instead of using the Discrete Cosine Transform, such as usual in the MFCC procedure [Davis and Mermelstein, 1980], the log filter-bank energies are filtered by a linear and second order filter. This technique is known as Frequency Filtering (FF) [Nadeu *et al.*, 1997]. The filter selected to this implementation has the following transform frequency response:

$$H(z) = z - z^{-1}, \tag{3.1}$$

and it is applied over the log of the filter-bank energies. The shape of this filter allow a best classification due it emphasizes regions of the spectrum with high speaker information yielding more discriminative information. A vector of 30 FF coefficients are extracted from the speech signal every 10ms. The choice of this kind of parameters is based on the fact that the use of the FF instead of the classic MFCC has shown promising results in both speech and speaker recognition tasks [Nadeu *et al.*, 2001]. This features have exhibit both computational efficiency and robustness against noise than the MFCC. In addition, as it can be seen as a filter in the frequency domain, they have frequency meaning. By means this notion, FF features allows the use of frequency techniques as masking, noise subtraction and so forth. Furthermore, other interesting characteristics can be found: Such as they are uncorrelated, computationally simpler than MFCCs and it does not decrease clean speech recognition results [Macho and Nadeu, 1999]. Summarizing, the FF filter technique must be seen as a liftering operation performed in the spectral domain equalizing the variance of the cepstral coefficients.
Aiming to capture the temporal evolution of FF parameters, the first and second time derivatives of the features, so called Δ and Δ-Δ coefficients [Furui, 1986], are appended to a basic static feature vector yielding to a vector of dimension 90. Note that the first and the last coefficients of the FF output of each frame contain absolute energy [Nadeu *et al.*, 1995], so despite of they may carry much noise, they are also employed to compute model estimation as well as its velocity and acceleration parameters.
Finally, in order to compute the likelihood between the training and the testing speech, for each speaker that the system has to recognize, a Gaussian Mixture Model (GMM) [Reynolds, 1995] of the probability density function of the parameter vectors is estimated. A weighted sum of size 64 Gaussians is applied in this approach. Given the collection of training vectors for one speaker, maximum likelihood (ML) model parameters are estimated through the iterative Expectation-Maximization (EM) algorithm. It is well known, the sensitive dependence of the number of EM-iterations in the conditions of few amount of training data. Hence, to avoid over-training of the models, 10 iterations are considered enough for parameter convergence in both training and testing conditions.
In the testing stage of the speaker identification system, a set of FF parameters $\mathbf{X} = \{\mathbf{x}_i\}$ is computed from testing speech signal, in the same way as explained above. The likelihood that each client model performs over the vector $\mathbf{X}$ is computed and the speaker exhibiting the largest likelihood is chosen,

$$s = \underset{j}{\operatorname{argmax}} \{ L(\mathbf{X}|\lambda_j) \}, \tag{3.2}$$

where s is the recognized speaker and L is the likelihood function from a linear combination of M unimodal Gaussian of dimension D. $L(\mathbf{X}|\lambda_j)$ thereby is the likelihood that the vector $\mathbf{X}$ has generated by the speaker with the model λ_j.

Speech Beamforming

The Delay-and-Sum beamforming technique [Flanagan *et al.*, 1985] is a simple yet effective way to enhance an input signal when it has been recorded on more than one microphone. It does not assume any information about the position of the microphones or their placement.

We can hypothesize that the speech wave arriving to each microphone is flat whether we assume the distance between the speech source and the microphones is enough far. Therefore the difference between the input signals, only taking into account the wave path and without take care about channel distortion, is the delay of arrival due the different positions of the microphones with regard to the source. So if we estimate the delay between two microphones we could synchronize two different input signal with the aim of enhancing speaker information and reduce additive white noise.

Hence, given the signals captured by N microphones, $x_i[n]$ with $i = 0 \dots N-1$ (where n indicates time steps) and their individual relative delays $d(0,i)$ (TDOA) with respect to a common reference microphone x_0 , we can obtain the enhanced signal by adding together the aligned signals by means equation:

$$y(n) = x_0[n] + \sum_{i=1}^{N-1} W_i x_i[n - d(0,i)]. \tag{3.3}$$

In order to estimate the TDOA between two segments from two microphones we have used the generalized cross correlation with phase transform (GCC-PHAT) method [Knapp and Carter, 1976] as explained in previous chapter in section 2.4.1.

The complete sentence is employed for TDOA estimation. TDOA values are computed as the maximum value of $\hat{R}_{PHAT_{ij}}(d)$ in both testing and enrollment stages. This estimation is obtained through different window size depending upon duration of the testing speech (1s/5s/10s/20s). During training stage, the same scheme is applied and TDOA values are computed from the training sets of 15 and 30 seconds, respectively. It is worth to mention that differences in the window size in each TDOA estimation are due to use of all speech data available to compute it. The weighting factor W_i, which is applied to each microphone to compute the beamformed signal, is fixed to the inverse of the number of channels taking into account. It relies on the assumption that each microphone has the same frequency response. A total of 20 microphones are mixed to perform the acoustic delay-and-sum, selecting 1 out 3 channels from the MarkIII array.

Multi-microphone Decision Fusion

In this approach a multiple distant microphone (MDM) system is implemented by fusing three SDM classifiers, described at the beginning of this section, working on three different microphones. The microphones number: $4, 34$ and 60, chosen from the total of 64 mics of the MarkIII array, are employed. The three systems estimate independently the speaker identification and, by means a simple voting rule, a final ID decision is taken. The figure 3.3 depicts the approach implementation.

Although the system identification methodology applied to each microphone is essentially the same as in the SDM case, in some cases the three classifiers do not agree about the detected speaker. That is due to the

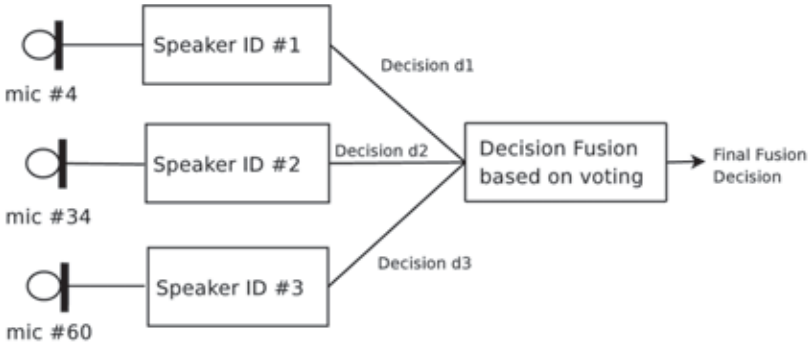


Figure 3.3: *Once the individual decisions of each classifier are performed, a simple voting rule is in charge to estimate the correct ID among the three different identifications provided by each individual SDM system*

fact that the input speech wave processed per each of the systems are not the same. Speech input on each microphone depends upon the microphone position with respect to the speaker. Thus different mic channels can suffer different degradation caused by reverberation effects into the room or by multi-path propagation. In order to decide a sole ID in function of each classifier output, a simple fusion of decisions is applied based on the following voting rule,

$$\begin{cases} \text{if} \quad D_i \neq D_j \quad \vee i, j \neq i & \text{select the central microphone ID} \\ \text{if} \quad D_i = D_j \quad \text{for some } i \neq j & \text{select } D_i \end{cases} \tag{3.4}$$

where D_i corresponds to the decision of the i-th SDM classifier. An ID is decided whether two or more of the individual systems agree about it. Otherwise, the central microphone decision is chosen, e.g., in the case all three classifier decide different ID outputs. Such a selection of the central microphone decision is motivated by its better single performance during the SID development experiments.

3.2.2 Video Person Identification

Recognition is stand-alone, taking detection and tracking for granted, i.e., the system is semi-automatic. A specific technique has been developed for face recognition in smart environments. The technique takes advantage of the continuous monitoring of the scenario and combines the information of several images to perform the recognition. Appearance based face recognition techniques are used given that the scenario does not ensure high quality images. As the visual identification evaluation is a close-set identification task, models for all individuals in the database are created off-line using two sets of video segments: the former consists on one segment of 15 seconds per each individual in the database, whilst the latter consists on one segment of 30 seconds per individual.

The image based system works with groups of face images of the same individual. For each test segment, face images of the same individual are gathered into a group. Then, for each group, the system compares such images with the model of the person[1].

Briefly, let $x_i = \vec{x}_1, \vec{x}_2, ..., \vec{x}_M$ be a group of M probe images of the same person, and let $C_j = C_1, C_2, ..., C_S$ be the different models or classes stored in the (local or global) model database where S is the number of individual models. Each model C_j contains Nj images, $y_n^j = \vec{y}_1^j, \vec{y}_2^j, \ldots \vec{y}_{N_j}^j$ where N_j may be different for every class. The group was assigned to the class resulting in a highest likelihood value. Likelihood was computed based upon a PCA approach as in [Kirby and L.Sirovic, 1990]. This way, the decision function, which computes similarity among the probe images and the model images, is the Euclidean distance between the projections of $\vec{x}_i$ and $\vec{y}_n^j$ on the subspace spanned by the first eigenvectors of the training data covariance matrix:

$$d(\vec{x}_i, \vec{y}_n^j) = \|W^T \vec{x}_i - W^T \vec{y}_n^j\|, \tag{3.5}$$

where W^T is the projection matrix. The XM2VTS database [Messer and et al., 1999] was used as training data for estimating the projection matrix and the first 400 eigenvectors are preserved. Due to the images being recorded continuously using the corner cameras, face images can not be ensured to be all frontal. Mixing frontal and non-frontal faces in the same models can be quite a problem for face recognition systems. To avoid this situation, eye coordinates are used to determine the face pose for each image. Only frontal faces were used for identification. Note that models per each person were automatically generated, without human intervention. All images for a given individual in the training intervals are candidates to form part of the model. Candidate face bounding boxes were projected on the subspace spanned by the first eigenvectors of the training data covariance matrix W^T. The resulting vector was added to the model only if it was different enough from the vectors already present in the model.

3.2.3 AudioVisual Person Identification

The modality integration is addressed on the example of a smart room environment aiming to perform person identification by combining acoustic features and 2D face images. Results from various sensory modalities, speech and faces, are performed both individually and jointly with the purpose to compare the different approaches.

In a multimodal biometric system that uses several characteristics, fusion is possible at three different levels: feature extraction level, matching score level or decision level. Fusion at the feature extraction level combines different biometric features in the recognition process, while decision level fusion performs logical operations upon the monomodal system decisions to reach a final resolution. Score level fusion matches the individual scores of different recognition systems to obtain a single multimodal score. Fusion at the matching score level is usually preferred by most of the systems [Hernando *et al.*, 2006; Farrús *et al.*, 2006].

[1] For further details about the image person identification in CLEAR evaluations approach see the works in [Luque *et al.*, 2006b; Luque *et al.*, 2006a].

Matching score level fusion is a two-step process: normalization and fusion itself [Fox and et al., 2003], [Indovina and et al., 2003], [Lucey and Chen, 2003], [Yuan *et al.*, 2004]. Since monomodal scores are usually non-homogeneous, the normalization process transforms the different scores of each monomodal system into a comparable range of values. One conventional affine normalization technique is z-score, that transforms the scores into a distribution with zero mean and unitary variance [Lucey and Chen, 2003],[Yuan *et al.*, 2004]. After normalization, the converted scores are combined in the fusion process in order to obtain a single multimodal score. Product and sum are the most straightforward fusion methods. Other fusion methods are min-score and max-score that choose the minimum and the maximum of the monomodal scores as the multimodal score.

Normalization and Fusion Techniques

Scores must be normalized before being fused. One of the most conventional normalization methods is z-score (ZS), which normalizes the global mean and variance of the scores of a monomodal biometric. Denoting a raw matching score as a from the set $S = \{s_1, s_2, \ldots, s_n\}$ of all the original monomodal biometric scores, the z-score normalized biometric x^n_{ZS} is calculated according to following equation,

$$x^n_{ZS} = \frac{s_n - \mu_S}{\sigma_S}, \tag{3.6}$$

where μ_S is the statistical mean of set S and σ_S is its standard deviation.
Once score normalized are obtained, histogram equalization (HE) is applied aiming to equalize the variances of two monomodal biometrics, looking for reducing the non linear effects typically introduced by speech systems [de la Torre *et al.*, 2005; Farrús *et al.*, 2007]. Histogram equalization (HE) is a general non parametric method to match the cumulative distribution function (CDF) of some given data to a reference distribution. HE is a widely used non linear method designed originally for the enhancement of images. HE employs a monotonic, non linear mapping which reassigns values from input in order to control the shape of the output values in order to match a desired distribution. Hence, the objective of HE is to find a non linear transformation to reduce the mismatch of the statistics of two signals. For instance, in [Pelecanos and Sridharan, 2001; Skosan and Mashao, 2006] this concept was applied to the acoustic features, instead of scores, to improve the robustness of a speaker verification system.
In this case, the HE technique matches the histogram obtained from the speaker verification scores and the histogram obtained from the face identification scores, both evaluated over the training data. The designed equalization takes as a reference the histogram of the scores with the best accuracy, which can be expected to have lower separate variances, in order to obtain a bigger variance reduction.
The figure 3.4 gives a visual explanation of HE technique. N intervals with the same probability are assigned in the distributions of both signals. Each interval in the reference distribution, $x \in [q_i, q_{i+1}[$, is represented by $(x_i, F(x_i))$. Where x_i is the average of the scores and $F(x_i)$ is the maximum cumulative distribution value:

$$x_i = \frac{\sum_{j=1}^{k_i} x_{ij}}{k_i}, \qquad F(x_i) = \frac{K_i}{M}, \tag{3.7}$$

where x_{ij} are the scores in the interval, k_i is the number of scores in the interval, K_i is the number of data in the interval $[q_0, q_{i+1}[$, and M is the total amount of data. All the scores in each interval of the source distributions are assigned to the corresponding interval in the reference distribution. $F(x_i)$ sets the boundaries $[q_i^*, q_{i+1}^*[$ of the intervals in the distribution to be equalized. These boundaries limit the interval of values that fulfills the following condition: $F(q_i) \leq F(y) < F(q_{i+1})$, and all the values of the source signal lying in the interval $[q_i^*, q_{i+1}^*[$ will be transformed to their corresponding x_i value.

In Matcher Weighting (MW) fusion of each monomodal score is weighted by a factor proportional to the recognition rate, so that the weights for more accurate classifiers are higher than those of less accurate matchers. When using the Identification Error Rates (IER) the weighting factor for every biometric is proportional to the inverse of its IER [Indovina and et al., 2003]. Denoting w^m and e^m the weighting factor and the IER for the m-th biometric x^m and M the number of biometrics, the fused score u is expressed as

$$u = \sum_{m=1}^{M} w^m x^m, \tag{3.8}$$

where,

$$w^m = \frac{\frac{1}{e^m}}{\sum_{m=1}^{M} \frac{1}{e^m}}. \tag{3.9}$$

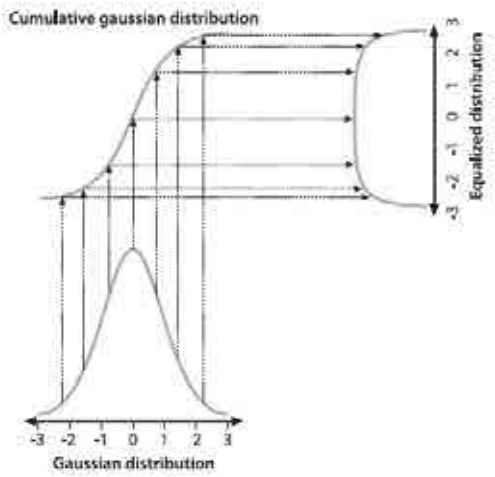


Figure 3.4: *Example of histogram equalization.*

Before carrying out the fusion process, HE is applied over all the previously obtained monomodal scores. Since the best recognition results have been achieved in the acoustic recognition experiments, the histogram of the voice scores has been taken as a reference in the histogram equalization. After the equalization process, the weighting factors for both acoustic and face scores are calculated by means corresponding Identification Error Rates, as in equation 3.9. Z-score normalization is also applied, and final fused scores are obtained through equation 3.8.

3.3 Experiments

In this section we summarize the official results for the UPC person identification systems for both CLEAR 2006 and 2007 evaluations. We examine the differences between the two evaluations as well as between the single and multiple microphone approaches and the audiovisual approach. The metric selected to benchmark the performance of the algorithms is the percentage of correctly recognized people from test segments.
Table 3.2 shows the correct identification rate for both audio and video modalities and the fusion identification rate obtained depending on the time duration of the test files. Related to acoustic identification task, it can be seen that the results, in general, are better as the segments length increases. For different test segment duration, the recognition rate increases as more data is used to match against speaker models. Overall, using the 30 seconds training segments, an improvement of up to 8% in the recognition rate is obtained with respect to the case where 15 seconds segments are used. For the face identification evaluation, in general, these results show a low performance of the system. Results for the training set B (using a segment of 30s to generate the models) show only a slight increase of performance with respect to training set A. It can also be seen that the results improve slowly as the segments length increases.
The reasons for this low performance in video approach are manifold: First of all, the system uses only frontal faces to generate the models and for recognition. However, most of the face views found in the recordings are non frontal. Another reason for the low percentage is the low quality of the images. The need to cover all the space in the room with four cameras results in small images, where the person's faces are tiny. In the worst cases, face sizes are only 13×13 pixels. In addition, poor illumination conditions in some recordings causes cameras to work at large diaphragm apertures. As a result, the depth of field is very shallow and several images are out of focus. Other recordings present interlacing errors. The figure 3.5 shows several examples of all these problems. Another problem is that, due to the fact that face bounding boxes are interpolated from the 1 second labels, our system is, in many cases, considering as "frontal" faces that are not really frontal ones. The figures (a), (b), (c) and (e) are examples of this situation.
This leads us to conclude that, under these conditions, a more elaborated video technique should be used. For instance, non-frontal face views should be taken into account, as most of the views found in the recordings are non-frontal. Even in this case, person identification using face detection alone is probably not going to give good results in these conditions. Identification should be performed combining more features other than those obtained from face bounding-boxes.

CLEAR 2006		% ID rates in TRAIN A			% ID rates in TRAIN B		
Duration (sec)	No. Segments	Speech	Video	A/V Fusion	Speech	Video	A/V Fusion
1	613	75.0	20.2	**77.3**	84.0	19.6	**87.8**
5	411	89.3	21.4	**92.0**	97.1	22.9	**97.3**
10	289	88.2	22.5	**93.4**	97.6	25.6	**98.6**
20	178	92.1	23.6	**97.7**	98.8	27.0	**100**

Table 3.2: *Percentage of correct identification for both audio and video unimodal modalities and multimodal fusion in CLEAR 2006 evaluation data. First column shows the duration of test segments in seconds. Second one shows the number of tested segments. Train A and B corresponds to training sets of* 15 *seconds and* 30*, respectively*

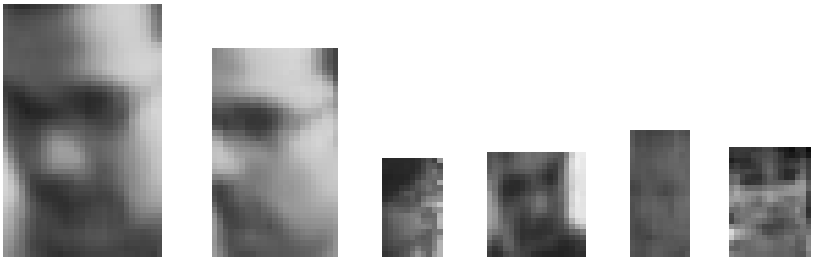


Figure 3.5: *Examples of face bounding boxes taken from several recordings and its relative sizes. Smallest image (c) is* 13×13 *pixels and larger image (a) is* 29×47 *pixels. Images are taken from the training segments of the AIT, IBM, ITC, UPC and UKA recordings. Images in the test sequences are similar.*

CLEAR 2007	% ID rates in TRAIN A				% ID rates in TRAIN B			
Duration (sec)	SDM'06	SDM'07	Fusion	D&S	SDM'06	SDM'07	Fusion	D&S
1	75.0	78.6	**79.6**	65.8	84.01	83.3	**85.6**	72.2
5	89.3	**92.9**	92.2	85.7	97.08	95.3	**96.2**	89.5
10	88.2	**96.0**	95.1	83.9	96.19	**98.7**	97.8	87.5
20	92.1	**98.2**	97.3	91.1	97.19	**99.1**	**99.1**	92.9

Table 3.3: *Percentage of correct identification in both TRAIN A and TRAIN B conditions just for audio systems. The table shows the rates obtained for the single microphone (SDM'07), Decision Fusion and Beamforming (D&S) systems. In addition, results from the single channel system from previous evaluation (SDM'06) are also provided.*

For the audiovisual approach, determination of the weighting factors applied to multimodal fusion has been estimated by using the training signals of 30 seconds as a development set. First 15 seconds have been used for training and the other 15 seconds for testing. The recognition results obtained in the evaluation for multimodal identification can also be seen in the table 3.2. Fusion results of both systems are also shown for the different

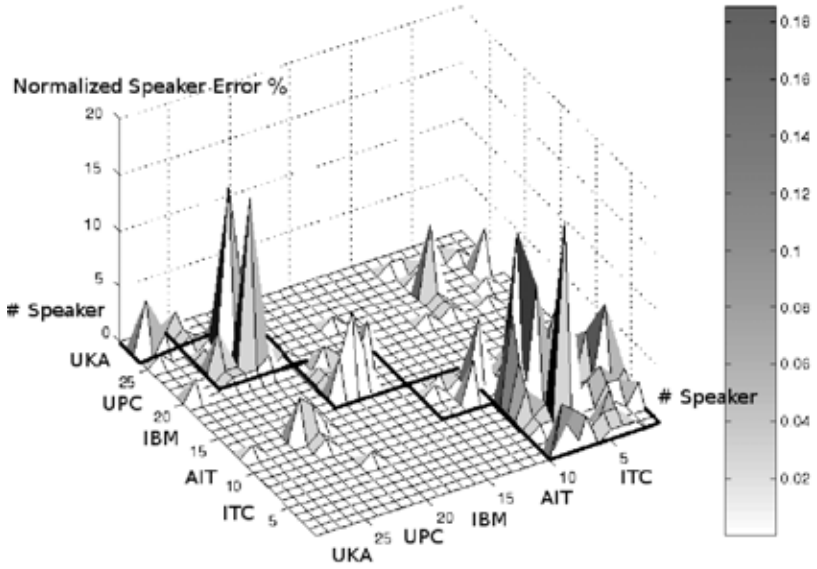


Figure 3.6: *Normalized Speaker Error from SDM in all test conditions. We can see the error mostly appears between the speakers of the same recording conditions.*

test durations. Overall, fusion correct identification rates are higher than the monomodal rates, outperforming those obtained with both monomodal systems.

Note that microphone number 4 from the MarkIII array was selected for testing in the SDM algorithm with the purpose of comparing among CLEAR evaluations. Table 3.3 shows the correct identification rate for both 2006 and 2007 CLEAR Speaker Identification Evaluation using the single microphone approach. In addition, we also can see the identification rates obtained by the two multi-microphone implementations described in the previous section.

Some improvements have been performed on the system since the CLEAR'06 Evaluation, leading to better results than the ones presented in that. It can be seen that the results, in general, are better as the segments length increases. The tables show that for the different test segment duration the recognition rate increases when more data is used to test the speaker models. Overall, using the 30 seconds training segments, an improvement of up to 6% in the recognition rate is obtained with respect to the case where 15 seconds segments are used.

On one hand, as we can see in the tables 3.3 the delay-and-sum system is not well adapted to the task. The low performance of this implementation may be due to a not accurate estimation of the TDOA values, nonetheless all systems presented in the evaluation based on any kind of signal beamforming neither did not show good results. By contrast, the same technique was applied in the Rich Transcription Evaluation-07 [Luque *et al.*, 2006b] obtaining good results in the diarization task. Other possibility to this low performance could be the background noise and the reverberation effects from each room setup. The recordings was collected from 5

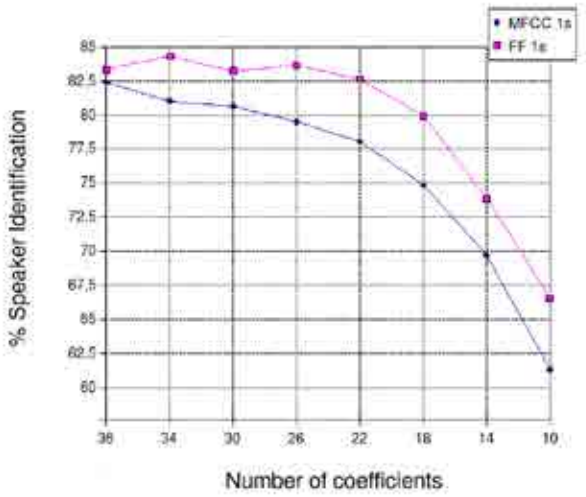


Figure 3.7: *Percentage of correct identification of the SDM approach in function of the number of short-term coefficients and the kind of parametrization. The* $30s$ *training set and the* $1s$ *test condition from the Evaluation data 07 were employed to draw the figure*

different sites, which could aid the MFCC+GMM system to discriminate between the recorded speakers from different room environments.

On the other hand, the decision fusion system seems, even with a very simple voting rule, to exploit the redundant information from the multi channel system. This technique achieves the best results in the tests of $1s$ using any enrollment set and, in general, in all the conditions of the training set of $30s$.

The figure 3.6 depicts the error behavior between speakers from the SDM implementation, a total of 348 over 3024 ID experiments. The boxes around the main diagonal enclose speakers from the same site, i.e., recordings with the same room conditions. As we had commented above, we can see the distribution of speakers errors is focused over the main diagonal. The picture shows that the system in general confuses the speakers from the same site. This kind of behavior could be motivated not only by the room conditions, such as room setup and geometry or the microphone response for the data acquisition, but also by accent and dialect of the speakers. Therefore, the speech parameters seem also modeling the room environment as well as the speaker features.

Some post-evaluation experiments focusing on the signal parametrization, that is, a comparison between frequency filtering and cepstrum-based parameters, are also provided. The figure 3.7 draws correct identification rates as function of number of parameters extracted from speech wave. The figure 3.7 reports the results in the TRAIN A and 1s test condition as function of number of these parameters. We can see that the selection of 30 dimensional static parameters vector employed in the evaluation system, value which was tuned through CLEAR development data, is so close to the optimum one, 26. Furthermore, the figure 3.7 also shows the

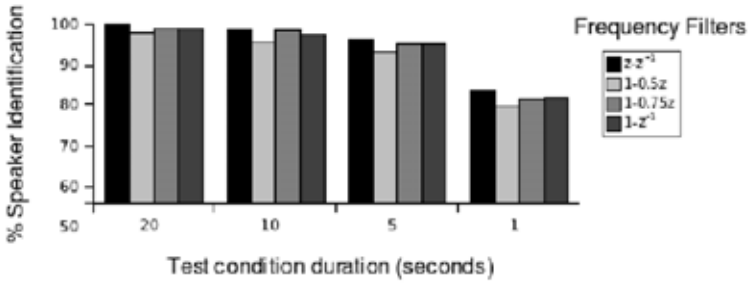


Figure 3.8: *Percentage of correct identification of the SDM approach using 4 frequency filters. The former is the same employed approach than in the SDM submitted system: $H(z) = z - z^{-1}$. The others are $H(z) = 1 - \alpha z^{-1}$ as function of α value.*

performance achieved by MFCC coefficients. The performance of the MFCC parametrization is always below the FF result curve independently of number of parameters fixed for each couple of FF and MFCC compared. Moreover, a comparison between several frequency filters is also provided in the figure 3.8. The filter applied to outputs of the filter-bank energies in the evaluation, $z - z^{-1}$, is compared to first-order filter $1 - \alpha z^{-1}$ for different values of alpha, that is, we are weighing differently in the quefrency dimension. As the development experiments had showed, the best behavior is obtained by the second-order filter.

3.4 Conclusions

In this chapter several techniques for acoustic and visual person identification into smart room environments has been described. The approaches were submitted to the international CLEAR evaluations.

Several audio approaches based on Gaussian mixture models and frequency filtering coefficients has been employed to perform speaker recognition, as well as two multi-microphone systems based on acoustic beamforming and a fusion rule at decision level. For video, an approach based on joint identification over groups of images of a same individual using a PCA approach has been followed.

For the acoustic identification task, the results show that the presented approach is well adapted to the conditions of the evaluation. For the visual identification task, the low quality of the images results in a low performance of the system. In this case, results suggest that identification should be performed combining more features other than frontal face bounding-boxes.

To improve the obtained results, a multimodal score fusion technique was suggested and implemented. Matcher weighting join histogram equalized scores are applied to monomodal scores of both audio and video systems. The multimodal results show that this technique leads to improvement of the recognition rate in all train/test conditions compared to individual modalities rates.

It is worth to mention that UPC acoustic evaluation [Luque and Hernando, 2008a] results in CLEAR evaluation 2007 ranked the best among the different approaches submitted by other participants [Stiefelhagen *et al.*,

2007]. The UPC audio result shows that GMM technique based on FF parameter flawlessly suits the evaluation requirements reaching identification rates over $90\ \%$ in most of the conditions. Anyway, controlled condition (test excerpts are manually extracted from recordings) of the evaluation that ensures a unique speaker per test segment, absence of long silences and overlap speech or few amount of speaker to be identified, helps to such a easy recognition system to obtain so high recognition rate results. In next chapters such issues will be addressed within the framework of a recognition system equipped to deal with them in a really multi-party environment, that is, a speaker diarization system.

Chapter 4

Speaker Verification in Conversational Telephone Speech: NIST Speaker Recognition Evaluations

Research on speaker recognition began in the 1960's when scientists attempted to use the speech spectrogram as a tool for speaker recognition [Kersta, 1962]. Nevertheless, computer technology was, at that time, not advanced enough to complement the manual work of phoneticians interpreting the spectrograms. According to [Reynolds, 2002], there are two main factors that make human voice a compelling characteristic to recognise people: in the first place, speech is a natural signal to produce that is not considered threatening by users to provide. Second, the telephone system provides currently a ubiquitous, familiar network of sensors for obtaining and delivering the speech signal. One can find in the literature an enormous amount of different approaches to the problem of speaker recognition (SR). Although short-term cepstral based systems are still the core of some of the most successful systems, current state-of-the-art systems typically employ a combination of different features and classification approaches permitting a better characterization of speakers and consequently an improved performance. NIST SRE evaluations has become the best place to assess and discuss approaches, given a common framework of development and discussion to researchers in speaker recognition community.

NIST (The National Institute of Standards and Technology) has coordinated evaluations in various relevant speech processing topics. Over the past sixteen years, NIST organized evaluations of text independent speaker recognition using conversational telephone speech [Doddington *et al.*, 2000; Przybocki *et al.*, 2004; Przybocki *et al.*, 2006]. These evaluations aim to explore promising new ideas in speaker recognition, developing advanced technology by incorporating these ideas and measuring its performance. By providing explicit evaluation plans, common test sets, standard measurements of error, and a forum for participants to openly discuss algorithms, the NIST series of Speaker Recognition Evaluations (SRE's) has provided a means

for recording the progress of text-independent speaker recognition performance.
This chapter presents the speaker recognition system jointly developed by the INESC-ID's Spoken Language Systems Laboratory (L^2F) and the TALP Research Center from the Technical University of Catalonia (UPC) for the SRE'10 campaign [Abad *et al.*, 2010]. Systems' implementation were carry out in the international L^2F speech laboratory during a four month stage by the author of this PhD. dissertation. A great variety of systems were implemented and submitted to SRE'10 evaluation, but most of them with similar characteristics and the same objective: robustness against a huge variety of speech conditions and a crowd speaker database. Among all of them, we will refer as the primary system, the one composed by the fusion of five individual SR sub-systems of very different characteristics. Two of the sub-systems are based on Joint Factor Analysis (JFA) with two different sets of speech features, two additional sub-systems are based on Gaussian Supervectors (GSV) and relying on different approaches for the combination of GMM and SVM techniques. Finally, an original system based on adaptation features from an ASR is also compared:

- (I) *JFA-spectral* based on Perceptual Linear Prediction (PLP) features with log-RelAtive SpecTrAl (log-RASTA) processing.
- (II) *JFA-prosodic* which mimics previous system's implementation but relies on prosodic features.
- (III) *GSV-SVM* which is the standard supervector approach, combining Gaussian mixture models (GMM) with Support Vector Machines (SVM), using also PLP features with log-RASTA processing.
- (IV) *GSV-GMM* which is the pushing-back version of the supervector approach and makes use of same set of features as previous system.
- (V) *Transformation Network* features with SVM modeling (TN-SVM) system is a new approach modeling features obtained from the adaptation transforms applied to the Multi-Layer Perceptrons (MLP) that form a connectionist speech recognizer.

The TN-SVM sub-system is the only one that makes use of the automatic transcripts provided by NIST. In addition to the primary system, two alternative systems consisting of different system combinations were submitted. The first alternative submission consists of the fusion of the two JFA sub-systems, that is JFA-spectral + JFA-prosodic (I+II). The second one is the combination all the sub-systems that do not depend on the automatic transcriptions provided by NIST, that is JFA-spectral + JFA-prosodic + GSV-SVM + GSV-GMM (I+II+III+IV).

4.1 Speaker Verification in Conversational Telephone Speech

NIST Speaker Recognition Evaluation (SRE) main task consists of determining whether a specified speaker is speaking during a given segment of speech uttered by an unknown speaker, that is, a speaker verification task. During the sixteen years of NIST Speaker Recognition evaluations the tasks have evolved focusing on

different topics, i.e., tracking and segmentation [Martin and Przybocki, 2001]. But the basic task of speaker detection (determining whether or not a given speaker is speaking) has remained the primary focus of all the NIST Speaker Recognition Evaluations. The evaluations have all included the basic one-speaker detection task consisting of a series of trials. Each trial presents the system with a target speaker, defined by some speech by the speaker (duration are condition dependent), and with a test segment spoken by a single or various unknown speakers. For each trial, the system must decide whether or not the unknown speaker is the target, producing both a yes-or-no hard decision and a likelihood score. Each SRE evaluation have its own characteristics, train and test conditions and corpora, e.g., but mainly them are focused on the speaker detection task in conversational telephone speech. The factors of interest include most particularly variations in the telephone handsets used and the types of transmission channels involved, and the match or mismatch of these between the training and test speech data. As example of such a variety of tasks and conditions, in last SRE'10 [Martin, 2010] evaluation the gender of speakers is known "a priori" in both train and test segments, the vocal effort is also annotated as well as the speech transcription, there exits a wide variety of microphone types and handsets (more than 20), different speaker languages (but mainly English) are used, also a two-speaker detection task and tracking using summed two-channel telephone is included, different train and test trial durations, session and conversation variability, a huge amount of trials (just the core condition last SRE'10 included more than half a million of them) and so on. All of them are some examples can be reeled off from a long list. Detailed information on the SRE'10 campaign can be found in the specific evaluation plan document [Martin, 2010] and, in general, in the NIST SRE website [National Institute of Standards and Technology, NIST, 1995]. The particular advantage offered by voice as a biometric is that it is transmissible over telephone channels. Telephone handsets, landline or cellular, are ubiquitous in modern society. The variability of telephone handsets and telephone channels makes the recognition task far more difficult and degrades the quality of performance. Nevertheless this has been the area of greatest application interest, and thus of greatest interest for evaluation.

It is well known that recognizing speakers in conversational telephone speech is a significantly more challenging task than speaker detection in broadcast news (BN) or meeting data. BN audio or meeting data are for the most part wide band, with telephone speech data accounting for only a very minor portion of the data in the case of BN. Hence spectral estimation is performed on the 0-3.8KHz band as opposed to the 8KHz bandwidth used for BN or meeting data with the consequent loss of speaker information.

One of the main factors of interest, in addition to those related to the voices of the speakers themselves, include most particularly variations in the telephone handsets used and the types of transmission channels involved. Performance of SR systems is usually enhanced when matching conditions occurs between training and testing conditions, e.g., when speakers use same handsets. This is not surprising since different speakers essentially always use different handsets, so success may be attained by identifying handsets rather than voices. Requiring that training and test handsets always be different is therefore a desirable evaluation objective, becoming one of the milestones of SRE evaluations in spite of being a challenging goal.

NIST evaluations have also shown how the two common handset microphone types (carbon button and

electret) of landline phones affect performance [Przybocki *et al.*, 2004; Przybocki *et al.*, 2006]. Performance is generally enhanced both by the use of electret microphones and by the use of matched type between training and test. Carbon button handsets now are becoming uncommon. Recent NIST evaluations have also shown that cellular transmission generally produces performance inferior to that with landline transmission. This is perhaps not surprising, but further investigation of related issues is needed. The previous NIST evaluations have made clear the need to investigate the effects of different handset and telephone transmission types on performance. The use of cellular and cordless phones has become pervasive in the past decade, and the use of specialized handsets such as speaker phones and headsets has increased. There has also been renewed interest in the effect on performance of speakers of different languages, particularly if some speakers should use multiple languages. For forensic applications there is interest on the interaction of collection channels that may include different types of microphones as well as telephone data.
Following a list of some of the main factor of interest in processing CTS data:

- **Handset type** In addition to the microphone type, telephone handsets may differ in how speakers use them for speaking and listening. They may involve speaker phones, headsets, ear-buds, or just ordinary hand-held devices. It is of interest to learn how these options, in different training and test combinations, may affect speaker recognition performance.

- **Transmission type** The effects of different types of cellular transmission are also worthy of examination in CTS data. Land-line, cellular, and cordless transmission are all widely used today. While older SRE evaluations had focused on either land-line or cellular calls, a careful examination of the alternatives, with training and test data always involving different handsets and sometimes involving different transmission modes, is attempted in recent SRE evaluation.

- **Language** The effect of language differences on recognition performance has been a subject of great interest, but one that has received limited study, due perhaps to a lack of comparable data involving multiple languages, and especially a lack of data involving bilingual speakers. It is generally believed that speaker recognition performance should not vary greatly with language, as long as the speech data used is entirely in one language, but this has not been verified in a formal evaluation. It is less clear what may be the effect on performance of having speech, for some speakers, in more than one language. The use of ”higher level” types of features such as word n-grams, in conjunction with traditional acoustic type features, to achieve improved greater performance levels, as pioneered in recent NIST evaluations, could make cross-language recognition performance more problematic. But test data from bilingual speakers is needed to investigate this.

- **Microphones** The primary application interests for speaker recognition, especially text-independent speaker recognition, have involved voice transmission over telephone lines. This is the area of advantage that voice possesses over other biometrics. But there is some interest, particularly for forensic applications, in recognizing voices recorded over various types of microphone channels. Of particular concern is the impact on performance of training and test data being recorded over different channel types,

perhaps telephone in one case and microphone in the other. This cross channel speaker recognition problem was investigated to a limited extent in the 2002 NIST evaluation and extended in last SRE evaluations 2008 and 2010.

4.2 NIST 2010 Speaker Recognition Evaluation Data

Appropriate data is essential for research in speaker recognition, and large quantifies of appropriate data are needed for statistically significant results. NIST has benefited from the ongoing collections of conversational telephone speech by the Linguistic Data Consortium [Linguistic Data Consortium, LDC, 2002]. Several collections of Switchboard style corpora [Godfrey *et al.*, 1992], each of which included hundreds of speakers and thousands of conversations, were used extensively in the detection tasks of the NIST Speaker Recognition Evaluations from 1996 to 2003. The 2004, 2005, 2006, 2008 and 2010 evaluations all used conversational speech data of the recently collected Mixer Corpora of the LDC [Campbell *et al.*, 2004; Przybocki *et al.*, 2007]. These corpora are based on a platform utilizing an automaton that can initiate contacts via phone to find pairings of registered participants to engage in recorded conversations on assigned topics. As with the previously used Switchboard platform, the participants can also initiate calls, and have the platform find them a conversational partner. The objective is to secure from a large number of target speakers a significant number (eight or more for the recent evaluations) of conversation sides from a single handset (telephone number) that may be used for training, and some number of conversations from other handsets, which may be used for test segment data.

SRE 2010 evaluation presents some new features compared to previous evaluations, e.g., by including in the training and test conditions for the core (required) test not only conversational telephone speech recorded over ordinary telephone channels, but also such speech recorded over a room microphone channel, and conversational speech from an interview scenario recorded over a room microphone channel. But unlike in SRE 2008 and prior evaluations, some of the data involving conversational telephone style speech has been collected in a manner to produce particularly high, or particularly low, vocal effort on the part of the speaker of interest.

SRE 2010 evaluation includes 9 different speaker detection tests defined by the duration and type of the training and test data. The data used comes from the Mixer telephone speech corpus collected by the Linguistic Data Consortium (LDC) [Campbell *et al.*, 2004; Przybocki *et al.*, 2007], as part of the various phases of its Mixer project or of its earlier conversational telephone collection projects, which consists of thousands of telephone conversations between hundreds of speakers within the US. The speakers cover a distribution of age, gender, location, and native languages. The participants in a telephone call are given general topics to discuss, but the conversations are unscripted and about five minutes in duration. Participants were encouraged to make many calls to the system over several weeks and to use varied telephone instruments and locations to provide large session and handset variability in the data.

The core test interview segments are of varying duration, ranging from three to fifteen minutes. Systems know whether each segment comes from a telephone or a microphone channel, and whether it involves the interview scenario or an ordinary telephone conversation, but it is required to process trials involving all segments of each type. Systems not know ”a priori” information about level of vocal effort in the conversational telephone style speech. Submitted results will be scored after the fact to determine performance levels for telephone data, for microphone data of different conversational styles and microphone types, for conversational telephone style data of different levels of vocal effort, and for differing combinations of training and test data.
The training segments in the 2010 evaluation are continuous conversational excerpts. As in recent evaluations, there will be no prior removal of intervals of silence. Also, except for summed channel telephone conversations as described below, two separate conversation channels will be provided (to aid systems in echo cancellation, dialog analysis, etc.). For all such two-channel segments, the primary channel containing the target speaker to be recognized will be identified. Word transcripts (always in English), produced using an automatic speech recognition (ASR) system, are also provided for all training and testing segments of each condition. These transcripts may be wrong, with English word error rates typically in the range of 15-30%.

SRE 2010 training conditions

The four training conditions to be included involve target speakers defined by the following training data:

- **10-sec**: A two-channel excerpt from a telephone conversation estimated to contain approximately 10 seconds of speech of the target on its designated side. (An energy-based automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.)

- **core**: One two-channel telephone conversational excerpt, of approximately five minutes total duration, with the target speaker channel designated or a microphone recorded conversational segment of three to fifteen minutes total duration involving the interviewee (target speaker) and an interviewer. In the former case the designated channel may either be a telephone channel or a room microphone channel; the other channel will always be a telephone one. In the latter case the designated microphone channel will be the A channel, and most of the speech will generally be spoken by the interviewee, while the B channel will be that of the interviewer’s head mounted close-talking microphone, with some level of speech spectrum noise added to mask any residual speech of the target speaker in it.

- **8conv**: Eight two-channel telephone conversation excerpts involving the target speaker on their designated sides.

- **8summed**: Eight summed-channel excerpts from telephone conversations of approximately five minutes total duration formed by sample-by-sample summing of their two sides. Each of these conversations will include both the target speaker and another speaker. These eight non-target speakers will all be distinct. speakers will all be distinct.

SRE 2010 Testing Conditions

For the interview segments, the provision of the interviewer's head-mounted close-talking microphone signal in a time aligned second channel, with speech spectrum noise added to mask any residual speech of the interviewee, is intended to assist systems in doing speaker separation, such as by using a speech detector to determine and remove from processing the time intervals where the interviewer is speaking. The test segments in the 2010 evaluation will be continuous conversational excerpts. As in recent evaluations, there will be no prior removal of intervals of silence. Also, except for summed channel telephone conversations as described below, two separate conversation channels will be provided (to aid systems in echo cancellation, dialog analysis, etc.). For all such two-channel segments, the primary channel containing the putative target speaker to be recognized will be identified.
The three test segment conditions to be included are the following:

- **10-sec**: A two-channel excerpt from a telephone conversation estimated to contain approximately 10 seconds of speech of the putative target speaker on its designated side (An energy-based automatic speech detector will be used to estimate the duration of actual speech in the chosen excerpts.)
- **core**: One two-channel telephone conversational excerpt, of approximately five minutes total duration, with the target speaker channel designated or a microphone recorded conversational segment of three to fifteen minutes total duration involving the interviewee (speaker of interest) and an interviewer. In the former case the designated channel may either be a telephone channel or a room microphone channel; the other channel will always be a telephone one. In the latter case the designated microphone channel will be the A channel, and most of the speech will generally be spoken by the interviewee, while the B channel will be that of the interviewer's head mounted close-talking microphone, with some level of speech spectrum noise added to mask any residual speech of the target speaker in it.
- **summed**: A summed-channel telephone conversation of approximately five minutes total duration formed by sample-by-sample summing of its two sides.

For the interview segments, the provision of the interviewer's head mounted close-talking microphone signal in a time aligned second channel, with speech spectrum noise added to mask any residual speech of the interviewee, is intended to assist systems in doing speaker separation, such as by using a speech detector to determine and remove from processing the time intervals where the interviewer is speaking.
The results presented in this chapter are focused on SRE 2010 core condition (core train vs core test) which is composed of 5 and 8 minutes audio excerpts from a conversational two sides telephonic (conv) or from an interview (int) each recorded using several telephonic channels or microphones placed at the room. Within the core test there are 9 Common Conditions. They are summarized as follows:

- *CC1*: Interview speech trials with matched microphones for train and test.
- *CC2*: Interview speech trials with unmatched misc for train and test.

Core condition	Styles train-test	Vocal effort	tel/mic	Number of speakers	Number of trials
CC1	*int-int*	-	*same mic*	2,159	62,864
CC2	*int-int*	-	*dif. mic*	2,159	219,842
CC3	*int-conv*	*NVE*	*tel*	1,609	58,043
CC4	*int-conv*	*NVE*	*mic*	1,520	85,902
CC5	*conv-conv*	-	*tel*	580	30,373
CC6	*conv-conv*	*NVE-HVE*	*tel*	365	28,672
CC7	*conv-conv*	*NVE-HVE*	*mic*	360	28,356
CC8	*conv-conv*	*NVE-LVE*	*tel*	300	286,04
CC9	*conv-conv*	*NVE-LVE*	*mic*	293	27,520
Total				9,345	570,176

Table 4.1: *Core (required) conditions in SRE 2010 evaluation. Each of the columns corresponds to: Speech style in train and test, conversation or interview; Vocal effort: low (LVE), normal (NVE), high (HVE); number of speakers both male and female; and number total of trials.*

- *CC3*: Trials involving interview training speech and normal vocal effort conversational telephone test speech.
- *CC4*: Trials involving interview training speech and normal vocal effort conversational telephone test speech recorded over a room microphone channel.
- *CC5*: Different number trials involving normal vocal effort conversational telephone speech in training and test
- *CC6*: Telephone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test.
- *CC7*: Room microphone channel trials involving normal vocal effort conversational telephone speech in training and high vocal effort conversational telephone speech in test.
- *CC8*:Telephone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test.
- *CC9*: Room microphone channel trials involving normal vocal effort conversational telephone speech in training and low vocal effort conversational telephone speech in test.

Tables 4.1 and 4.2 summarize the conditions in train-test with different styles as Normal Vocal effort (NVE), High Vocal Effort (HVE) or Low Vocal Effort (NVE) summarize the core-core condition in the evaluation. A completed description of the evaluation data in SRE 2010 evaluation can be found in the evaluation plan [Martin, 2010].

Common Condition	Trials Target (non-target)	Extended Trials Target (Non-target)
CC1	2,152 (60,712)	4,304 (795,995)
CC2	7535 (212307)	15,084 (2,789,534)
CC3	1,633 (56,410)	3,989 (637,850)
CC4	2,366 (83,536)	3,637 (756,775)
CC5	708 (29,665)	7,169 (408,950)
CC6	361 (28,311)	4,137 (461,438)
CC7	359 (27,997)	359 (82,551)
CC8	298 (28,306)	3,821 (404,848)
CC9	290 (27230)	290 (70,500)

Table 4.2: *Core (required) conditions in SRE 2010 evaluation. Number total of trials both target an non-target for several common conditions in core test and for extended trials.*

4.3 The L^{2}F-UPC Speaker Verification System

In this section some common characteristics shared by various sub-systems of the L^{2}F-UPC submission are described. The UPC-L^{2}F system is composed of a fusion of five different SR sub-systems. Figure 4.1 gives a brief scheme of the SR systems and techniques employed. In general, most of the systems are based on features estimated from a classical UBM-GMM system which are enhanced with robust techniques to deal with speaker and session variability like as JFA analysis and NAP. Furthermore, systems based on SVM, GMM push-back are also evaluated and contrasted. Prosodic characteristics are also applied in addition to spectral features as well as a newly set of features estimated from a connectionist transformation network in which ASR transcriptions are used.

4.3.1 Common characteristics

Most of the systems have in common several characteristics and strategies. That is the case of speech parametrization, universal background models (UBMs) or speaker and trials sub-sets applied to normalize scores. Following such a common features are illustrated.

Development and Training Corpora

Previous NIST evaluations' data is used for algorithm development and training of different sub-systems. Specifically, NIST SRE 2004, 2005 and 2006 telephone data sets were used for the training of systems and the SRE 2008 core test condition for development and algorithm assessment. Different subsets were selected for the various training stages:

- Gender dependent Universal Background Models (UBM).

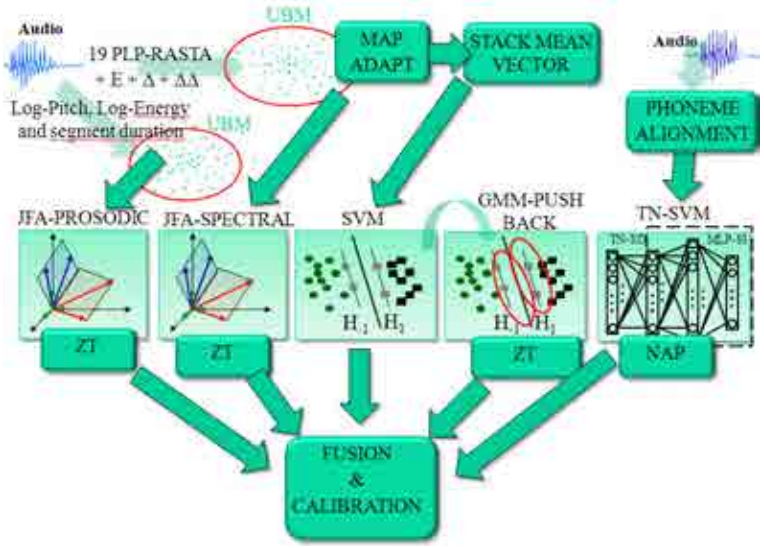


Figure 4.1: *Schematic diagram of the UPC-L^{2}F system submitted to SRE'10 evaluation.*

- Background impostor set in SVM based sub-systems (for sub-systems III and V).
- ZT-score normalization.
- Speaker/channel variability compensation techniques as modeling of JFA channel and speaker space and for NAP.

The performance of the individual sub-systems and several other tested SR approaches was assessed in the NIST SRE 2008 *telephonic-telephonic* test sub-set. The SRE 2008 core test condition, the so called *short2-short3* task condition, with around one hundred thousand trials was also used for system calibration and fusion of the final submission.
Notice that some of the tools used by the SR system and developed at the L^{2}F during the last years have been trained with additional data. For instance, the MLP speech-non-speech detector was trained mainly with down sampled broadcast news (BN) data, augmented with music and sound effects data. The MLP acoustic models of the hybrid speech recognizer, used in phone-alignment for the system (V), were trained on 140 hours of manually transcribed HUB-4 data.
The total amount of speech training material is around 400 hours, composed of 150 hours from males and 250 hours from females roughly, from more than 2000 male speakers and 3000 female.

Speech/Non-speech Segmentation

In order to detect low-energy and highly likely non-speech frames, a MLP-based speech-non-speech detector was trained with down sampled broadcast news (BN) data. Its output is combined with the alignment generated

by a simple bi-Gaussian model of the log energy distribution computed for each speech segment to obtain speech/non-speech segmentation for training data and trials.
The speech/non-speech detector is used in most of sub-systems. JFA-spectral, JFA-prosodic, GSV-SVM and GSV-UBM (sub-systems I through IV). All of them make use of silences detection in some manner, particularly, in the feature extraction process, by discarding non-speech frames and computing feature mean and variance normalization on speech frames. Notice that the segmentation available in the automatic transcriptions provided by NIST is only used by sub-system TN-SVM (sub-system V).
Segmentation of the interview segments was additionally post-processed. In order to obtain a better target speaker segmentation, the speech/non-speech segmentation of the interviewer (non-target) channel was obtained. Then, regions with simultaneous speech activity in the interviewee and the interviewer channels were removed from the target speaker segmentation.

Speech Features

The spectral features used in sub-systems *JFA-spectral* (I), *GSV-SVM* (III) and *GSV-GMM* (IV), consist of 19 PLP features with log-RASTA processing [Hermansky and Morgan, 1994] and the frame energy, from a sliding window of 20 ms with a step size of 10 ms. First and second derivatives are concatenated to form 60 element feature vectors. Low-energy and highly likely non-speech frames are removed according to the speech segmentation previously described. Finally, mean and variance feature normalization is applied with mean and variance being computed independently for every speech utterance.
The prosodic features used in sub-system *JFA-prosodic* (II) are aimed at modeling the prosodic contours (both energy and pitch) of syllable-like regions [Ferrer *et al.*, 2010]. We use the Snack toolkit [Sjlander, 1997] to extract the log-pitch and the log-energy of the voiced speech regions of every utterance. Log-energy is normalized on an utterance basis. The prosodic contours are segmented into regions by splitting the voiced regions wherever the energy signal reaches a local minimum (the minimum length of the regions is 60 ms). For each region, the log-energy and log-pitch contours are approximated with a Legendre polynomial of order 5, resulting in 6 coefficients for each contour. The final feature vector is formed by the two contour coefficients and the length of the syllable-like region, which results in a total of 13 elements.
The third set of features applied to sub-system *TN-SVM* (V) are obtained from a connectionist transformation network and are explained in detail in the system description section 4.3.3.

Universal Background Model

Gender-dependent Universal Background Models (UBMs) were trained on NIST SRE 2004, 2005 and 2006 telephone data. The Audimus software package [Meinedo *et al.*, 2003] and its utilities developed at the L^{2}F Laboratory were used for GMM modeling. A total of 72 hours from 870 male speakers and 100 hours from 1200 female speakers were used. The two gender-dependent UBMs were incrementally trained up to 1024 Gaussians, doubling the number of Gaussians at each iteration up to 25 iterations of the EM algorithm.

For each gender, two sets of UBMs were also trained depending upon speech features extracted from speech wave, described above. The former was trained using spectral features and used for building the sub-systems (I), (III), and (IV); and the latter was trained with the prosodic features used for the development of sub-system *JFA-prosodic* (II).

Score Normalization: ZT-norm

As has been reported in the literature [Auckenthaler *et al.*, 2000; Wan and Renals, 2005; Kenny *et al.*, 2008b], score normalization is a key point in SR systems. Specially, in order to enhance performance in JFA-based and GMM-based systems. Thus, raw scores are ZT-normalized [Zheng *et al.*, 2005] in sub-systems *JFA-spectral* (I), *JFA-prosodic* (II) and *GSV-GMM* (IV). In the case of sub-systems *GSV-SVM* (III) and *TN-SVM* (V), which are based on SVM classifiers, a significant impact of score normalization strategies was not observed and hence these strategies were not applied in the submitted version.
Gender-dependent sets were defined for score normalization. We used 400 speech segments (200 male and 200 female) for modeling the impostor set of speakers (set T of impostor speakers, corresponding to T-normalization) and a total of 400 speech segments (200 male and 200 female) for modeling the impostor score distribution per each target speaker (set Z of impostor trials, corresponding to Z-normalization). Both sets were randomly selected from the SRE2004 and SRE2005 data and no care was taken to avoid overlapping with the data used for UBM training, that is, no speaker overlap between UBM-GMM data and ZT-norm data was taken into account.
In the case of sub-systems *GSV-SVM* (III) and *TN-SVM* (V), which are based on SVM classifiers, a significant impact of score normalization strategies was not observed in development data and hence these strategies were not applied in the submitted version.

4.3.2 Speaker Verification using an Automatic Speech Recognizer

Sub-system *TN-SVM* (V) uses a set of novel features extracted from adaptation techniques applied to the Multi Layer Perceptrons that form a connectionist speech recognizer [Abad and Luque, 2010; Abad *et al.*, 2011]. Such features are extracted from speaker-phoneme adaptation thus phonetic alignment is required and a speech recognizer transcription was obtained for such purpose.

The Audimus Hybrid Speech Recognizer

The Audimus [Meinedo *et al.*, 2003] ASR module uses MLP networks that act as phoneme classifiers for estimating the posterior probabilities of a single state Markov chain mono phone model. The baseline system combines three MLP outputs trained with PLP features (13 static + first derivative), log-RASTA features (13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). When applied to narrow band recordings, the advanced Font-End from ETSI features (13 static + first and second derivatives) are also used. The number of context input frames is 13 for the PLP, RASTA and ETSI networks and 15 for the MSG

network. The system adopted in this work models only mono phone units, resulting in MLP networks of 40 soft-max outputs for English. The decoder of the recognizer is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

Narrow-band Acoustic Models

The lack of conversational telephone speech (CTS) orthographically labeled data prevented us from developing an ASR system matched to the characteristics of the NIST Speaker Recognition Evaluation data sets. Consequently, a simple narrow-band speech recognizer with acoustic models trained with down-sampled BN data was used for this evaluations. The MLP acoustic models were trained on the same 140 hours of manually transcribed HUB-4 speech used for our American English BN transcription system [Pellegrini and Trancoso, 2009].

Generation of Phonetic Alignments

Word-level automatic transcriptions provided by NIST were forced aligned using the narrow-band acoustic networks to obtain phonetic alignments. Then, the alignments were used for training the speaker dependent transformation networks. Whenever the NIST transcriptions were not available, the narrow-band speech recognizer with the BN language model was used to generate a (weak) automatic transcription.

4.3.3 The L^{2}F-UPC SR Sub-systems

The complete L^{2}F-UPC speaker recognition system is the result of the fusion of five speaker verification scores generated by 5 individual sub-systems. We will briefly describe particularities of the sub-systems.

(I) The JFA-spectral Sub-system

JFA approach has become one of the most successful compensation techniques for speaker verification as has been reported in [Kenny and Dumouchel, 2004], [Kenny *et al.*, 2007] and [Kenny *et al.*, 2008b]. L^{2}F-UPC's JFA system closely follows the description of "Large Factor Analysis model" in paper [Kenny, 2005] and described in the section 2.2.3 in which speaker model is represented by mean supervectors (stacked GMM-UBM means):

$$\mathbf{M} = \mathbf{m} + \mathbf{Vy} + \mathbf{Dz} + \mathbf{Ux}, \tag{4.1}$$

where m is the speaker independent mean supervector, V is a subspace with high speaker variability whose columns are referred to as eigenvoices, U is a subspace with high intersession/channel variability whose columns are referred to as eigenchannels, and D is a diagonal matrix describing remaining speaker variability not covered by V. Speaker factors y, z and channel factors x are assumed to be normally distributed random

variables. It is worth to mention that this representation constrains all supervectors m to lie in an affine subspace which is spanned by the columns of V.

Our JFA based submission consists of a Universal Background Model generation and JFA itself. UBMs are those described in the section 4.3.1. For JFA, the cookbook developed by Ondrej Glembek at Brno University of Technology [Glembek, 2009] has been used. Finally, JFA scores are ZT-normalized as described in previous section 4.3.1.

The UBMs were used to collect zero and first order statistics for training two gender-dependent JFA systems. The mean $\mathbf{m}$ and the variances of Gaussian components are set to the UBM mean and UBM variance respectively thus their estimation are not computed during the training of JFA parameters. WE briefly review the data and techniques employed for such parameter estimation:

- For JFA speaker modeling, 300 eigenvoices are trained on the NIST SRE 2004 and 2005 sets using speakers with at least 8 recordings or sessions (a total of 372 male and 519 female speakers). MAP point estimates of speaker factors are obtained and they are fixed for the following training of eigenchannels.
- A set of 80 eigenchannels were trained on NIST SRE 2004 and 2005 telephone data (1806 recordings from 184 different male speakers and 2301 segments from 245 different female speakers).
- The diagonal matrix $\mathbf{D}$ in the JFA equation was estimated on all eigenvoices and eigenchannels.
- A set of NIST SRE 2006 speakers composed of 2384 and 3215 recordings of 298 male and 402 female speakers respectively is used for this purpose. MAP estimates of speaker and channel factors are fixed for estimating this diagonal matrix which stands for a speaker variability not represented in matrix $\mathbf{D}$.
- The speaker factor $\mathbf{y}$ was jointly estimated with the channel factor $\mathbf{x}$ from the enrollment data. The common factor $\mathbf{z}$ was also estimated from training data.

In the testing stage, zero and first order statistics are extracted from the trial data. The channel's shift from UBM, i.e. the channel factor $\mathbf{x}$, is estimated from trail segment, fixing it for all the speaker models, following the UBM point estimate assumption in [Glembek *et al.*, 2009]. A linear scoring was performed to obtain the scores. Finally, factor analysis likelihood ratios were ZT-normalized, as described in section 4.3.1.

(II) The JFA-prosodic Sub-system

The JFA-prosodic (II) system have in common the same architecture of previous JFA-spectral system, as explained above, but relies on a complete different set of features for speech representation. Instead of the classical spectral coefficients, the prosodic features described in section 4.3.1 are modeled in this system. The data sets for UBM modeling, estimation of speaker parameters in equation 4.1 and for score normalization remain the same as in the *JFA-spectral* (I) sub-system. Likewise, ZT score normalization normalization is also applied by applying same sets as described in the section 4.3.1.

(III) The GSV-SVM Sub-system

Combining Gaussian mixture models with Support Vector Machines [Campbell *et al.*, 2006a], the so-called Gaussian supervector approach, is known to be a high performance speaker recognition approach.
For this evaluation, a GSV system based on mean supervectors was developed. First Gaussian mixture models for each target speaker are obtained with MAP adaptation of the Gaussian means of the UBM based on spectral features. UBM means are adapted with 20 MAP iterations with a relevance factor of 16 to finally obtain GMM-UBM speaker models.
The Gaussian Super Vector (GSV) system concatenates the mixture means of the MAP adapted Gaussian speaker models to obtain super vectors of every speech segment. The linear SVM kernel of [Campbell *et al.*, 2006b] is used for training speaker models by means libSVM [Chang and Lin, 2001] software package. The background set used as negative examples for SVM training is formed by 874 male, and 1204 female speech segments extracted from the SRE2004, SRE2005 and SRE2006 1 side training corpora. Finally, enrolled SVM speaker models are used for scoring supervectors obtained from the trials segments.
Due to time constraints, we did not implement Nuisance Attribute Projection (NAP) for this sub-system, which is known to provide additional benefits [Campbell *et al.*, 2006a].

(IV) The GSV-GMM Sub-system

The GSV-GMM sub-system is based on the GSV-SVM speaker recognition system explained in the previous section, but uses the alternative scoring approach of [Campbell, 2008].
In contrast to the conventional GSV, each speaker SVM model is *pushed back* to a *positive* and a *negative* speaker GMM model, which are used in testing to calculate log-likelihood ratio scores. In certain situations, especially on short utterances, this approach provides improved performances. In this sub-system score normalization is also applied. However, at the time of the submission all the necessary trials for performing the complete ZT-norm with 200 files per normalization and per gender were still not available. For that reason, Z-norm with only 100 Z-segments per gender was applied to the scores generated by the GSV-GMM subsystem.

(V) The TN-SVM-NAP Sub-system

The Transformation Network features with SVM modeling system is a novel approach [Abad and Luque, 2010; Abad *et al.*, 2011] that makes use of adaptation transforms employed in speech recognition as features for speaker recognition. However, in contrast to [Stolcke *et al.*, 2005], the automatic speech recognizer that we used relies on for computing the "differences" between the speaker independent and the speaker dependent model is the connectionist hybrid artificial neural network/hidden Markov model (ANN/HMM) system described in 4.3.2. Sub-system approach makes use of a method known as Transformation Network [Abrash *et al.*, 1995] to train a linear input network that maps the speaker-dependent input vectors to the speaker independent system, while keeping all the other parameters of the neural network fixed. The necessary

phonetic alignments for network adaptation are obtained as described in section 4.3.2. For each MLP network that composes the acoustic models, described in 4.3.2, TN adaptation method is applied and a set of adaptation weights is obtained. A single TN feature vector of total size 3895 is formed with the linear transformation weights of the four MLP networks, and with the mean and variance statistics of the features data.
Additionally, nuisance attribute projection (NAP) is applied to the TN features as described in section 2.2.3. Gender-dependent NAP projections were trained with the multisession conversational telephone speech sets from SRE2004, SRE2005 and SRE2006. A total of 7195 recordings from 921 different female speakers and 5226 recordings from 670 male speakers were applied for such purpose. A nuisance space of dimension 32 was employed in this work.
The resulting TN features with NAP compensation are used for training SVM speaker models. Gender-dependent negative examples for SVM training are obtained from the 1 side conversation training corpus of SRE2004, SRE2005 and SRE2006 like as in previous SVM-based approaches. A total of 867 and 1201 male and female segments are used for modeling of the background SVM samples. In this case, score normalization was not applied to the TN-SVM-NAP system since no enhanced results was noted during development. Additional implementation details can be found in [Abad and Luque, 2010; Abad and Trancoso, 2010; Abad *et al.*, 2011].

4.3.4 Calibration and Sub-systems Fusion

The SRE2008 *short2-short3* evaluation condition data set was used for adjusting calibration and fusion of the sub-systems which compose the L^2F-UPC submission. Unfortunately, this set is known to be small and not adequate to the particularities of the new cost function considered in SRE2010 evaluation. We are quite confident that the quality of calibration and fusion stage may be improved through a larger number of trials.

Linear Logistic Regression with FoCal

Linear logistic regression tools provided by the FoCal Toolkit [Brummer, 2005] were used for both calibration and fusion. In a first stage, sub-system was independently calibrated and in the case of gender-dependent system two different calibration were performed for male and female speakers, respectively. In some cases, some of the sub-systems were not able to produce a score for a concrete trial, due to the fact that some data excerpts are malformed. In that case, a score of 0 was given to the trial for that sub-system after the first calibration stage. Then, with all the scores from the five sub-systems, a second linear logistic regression was trained to compute the final fusion score. The decision threshold was set in accordance to the new SRE2010 cost function.

Configurations

Three different calibration and fusion configurations were trained depending on the characteristics of the training/testing segments involved in a given trial. The channel type: microphone (mic) or telephone (tel) and

Configuration	SRE'08	SRE'10
MIC-TEL	*int-conv/tel*	*int/conv-tel*
TEL-TEL	*conv-conv/tel*	*conv-tel/conv-tel*
MIC-MIC	*int/int*	Rest of trials

Table 4.3: *Calibratrion training sets applied to SRE'10 test data.*

the speech style: interview (int) or phonecall (conv). The following table summarizes such configurations.
The "mic-mic" configuration in SRE'10 was trained with the *interview-interview* subset of the *short2-short3* data set. The "mic-tel" configuration was obtained with the *interview-phonecall/telephone* trials. The *phonecall-phonecall/telephone* trials were used for estimating the calibration and fusion weights of "tel-tel" configuration.
Finally, the "mic-mic" configuration is used for the rest of the test trials: trials with *interview* data segments in both training and testing (independently of their length), trials with models trained with *interview* data and tested with *phonecall/microphone*, and trials with both *phonecall/microphone* data in train and test.
In testing stage, the "tel-tel" configuration was used for the trials with both the training segment and the test segment identified as *phonecall telephone* data segments. The "mic-tel" calibration and fusion is used for trials that involve speaker models trained with *interview* data (both *3min* and *8min*) and test segments with *phonecall/telephone* data.

4.4 Experiments

In this section, a brief guide of the strategy followed for developing of previous reviewed algorithms is given. Previous NIST SR evaluations were applied for such purpose. In addition, official evaluation results for the different systems submitted to the last NIST SR 2010 evaluation are also presented.

Algorithm Development and Training

The SRE2008 *short2-short3* evaluation data set were used for adjusting calibration and fusion of the subsystems, algorithm development and tuning of parameters.

- *Short2* training condition: A two-channel telephone conversational excerpt, of approximately five minutes total duration, with the target speaker channel designated (telephone telephone) A microphone recorded conversational segment of approximately three minutes total duration involving the target speaker and an interviewer (interview microphone)

- *Short3* testing condition: A two-channel telephone conversational excerpt, of approximately five minutes total duration, with the putative target speaker channel designated (telephone telephone) A similar

such telephone conversation but with the putative target channel being a (simultaneously recorded) microphone channel (telephone microphone) A microphone recorded conversational segment of approximately three minutes total duration involving the putative target speaker and an interviewer (interview microphone)

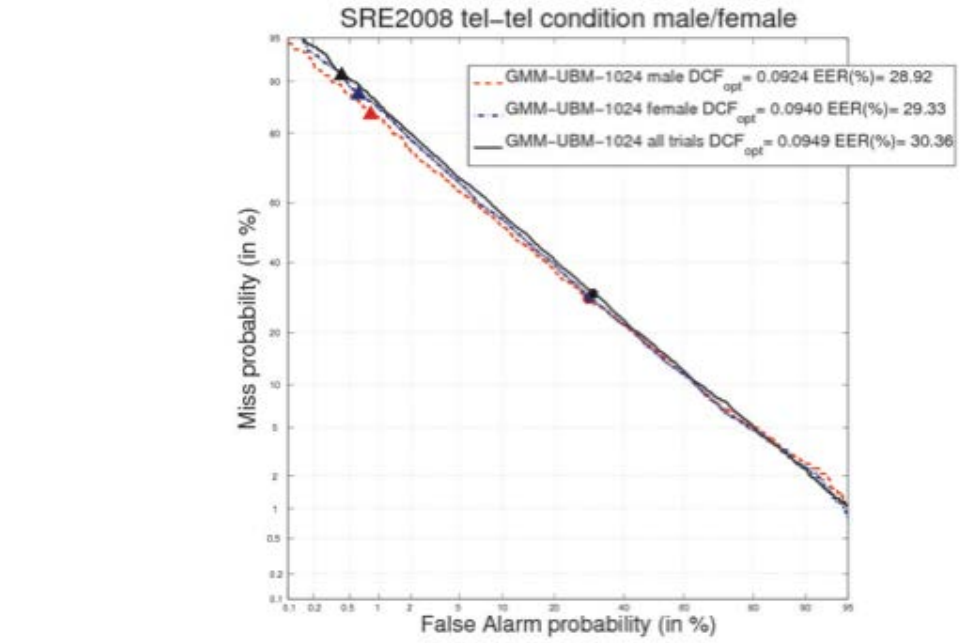
The *short2-short3* condition is composed of $3,623$ speakers models ($1,270$ male and $1,993$ female) and a total of $6,377$ different files. Combination of test segment and claimed identities results in a total of $98,776$ different trials. In order to speed-up the development experiments a sub-set from *short2-short3* was selected which includes all trials from tel-tel condition, known as condition 6 of the SRE'08 core condition. This set is composed of $1,789$ speakers models (648 male and $1,141$ female) and a total of $2,573$ different files resulting in 37.069 trials.

The figures in next pages present the development results for the implementation of the different algorithms described in previous sections. Firstly, the performance results of the UBM employed for most of the systems are depicted in figure 4.2. The figure (a) draws the DET curves in tel-tel SRE 2008 data for both male and female trials. The optimum detection cost function (DCF) point is drawn as a triangle point while the equal error rate (EER) as a circle dot, see section 2.2.4. In the figure (b) the comparison with different normalization techniques as well as with the calibrated scores are depicted. As we can see at the DET curves the normalization plays and important role in UBM-system performance. ZT-norm shows the best results on SRE 2008 tel-tel data reaching a 16% of EER. Anyway, results presented are far from those obtained by best UBM-based systems presented in previous evaluations [Martin and et al., 2008a; Martin and et al., 2008b]. Such a deviation is due to the fact that not enough data is considered for UBM estimation as well as for the normalization sets Z and T. For instance, most of the best systems presented in SRE'08 make use of 3-4 times more data for that purpose.

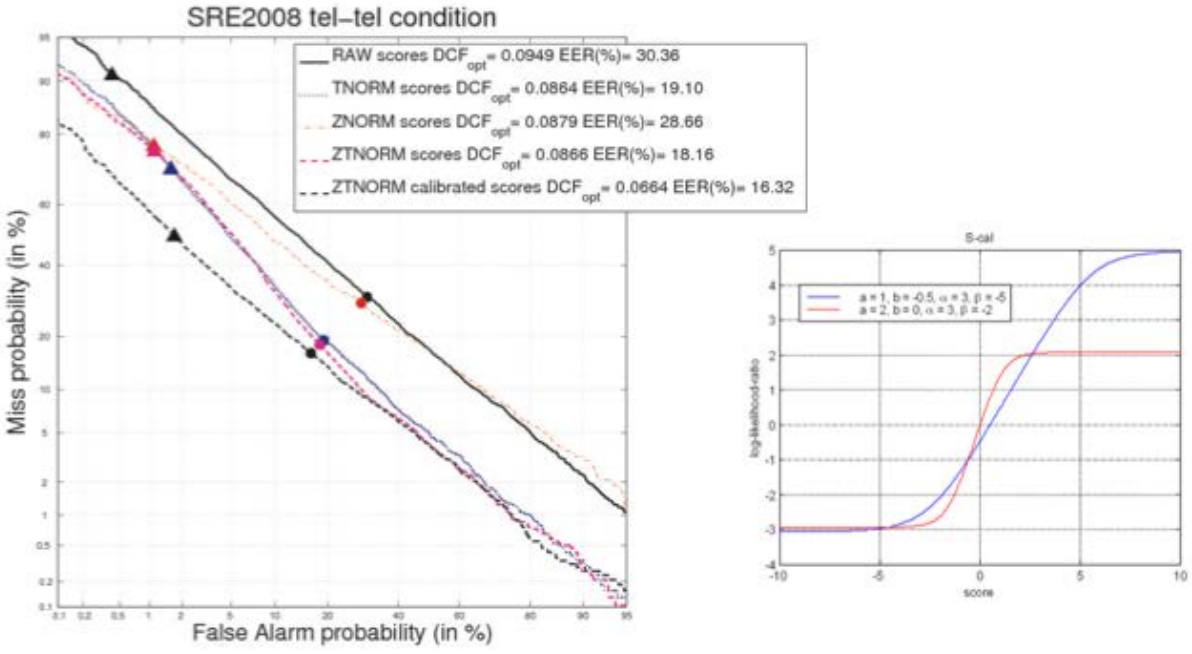
The figure 4.3 (a) depicts the results obtained by the Gaussian Super Vector with SVM classifier and by the push-back GMM version compared to the ZT-normalized GMM-UBM system. Scores from male and female speakers are polled together in this picture and for all the following. Furthermore, all the systems are calibrated and normalized with same Z and T sets described in section 4.3.1. The figure 4.3 (b) also reports the results

Task		#Models	#Tests	#trials
short2-short3	SRE'08	3,263	6,377	98,776
short2-short3	SRE'08-male	1,270	2,528	39,433
short2-short3	SRE'08-female	1,993	3,849	59,343
tel-tel	SRE'08	1,788	2,573	57,050
tel-tel	SRE'08-male	648	895	12,922
tel-tel	SRE'08-female	1,140	1,678	24,128

Table 4.4: *Number of different target models, different test signals and total number of trials in the developments data sets.*

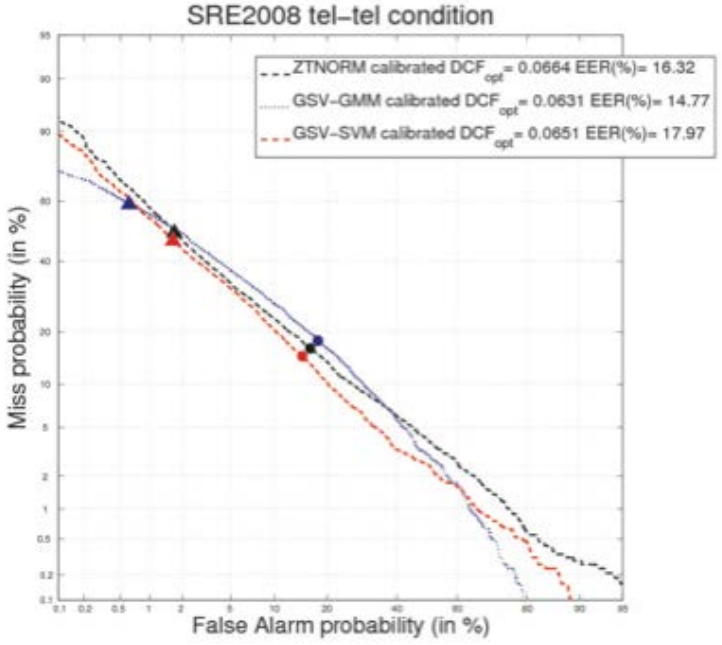


(a)

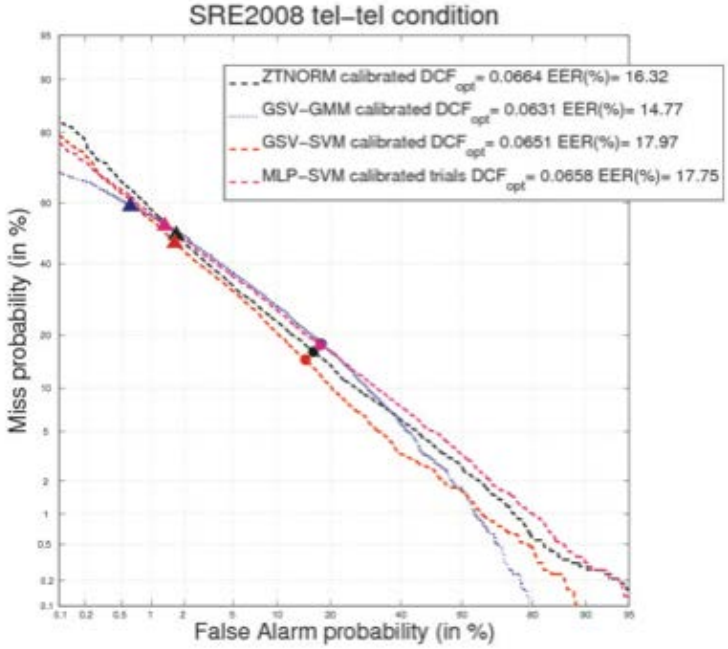


(b)

Figure 4.2: *DET curves in SRE 2008 short2-short3 in tel-tel condition data. Equal Error Rate (EER) (circle dot) and optimum Detection Cost Function (DCF) point is also depicted (triangle point).* ***(a) Top****: Detection curves for both male and female trials (by using a gender dependent GMM-UBM in each case) and the resulting DET curve for arranging in stack the gender-dependent-scores for all trials.* ***(b) Bottom****: Detection curves for GMM-UBM system with score normalization: Z-norm, T-norm and ZT-norm normalization techniques as well as the ZT-norm calibrated version.*

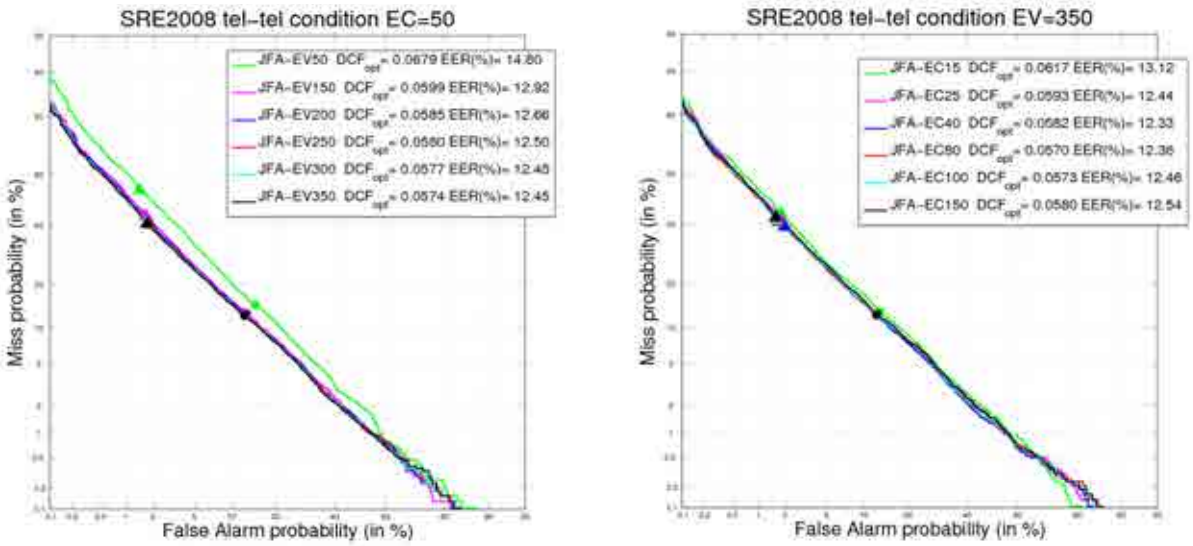


(a)

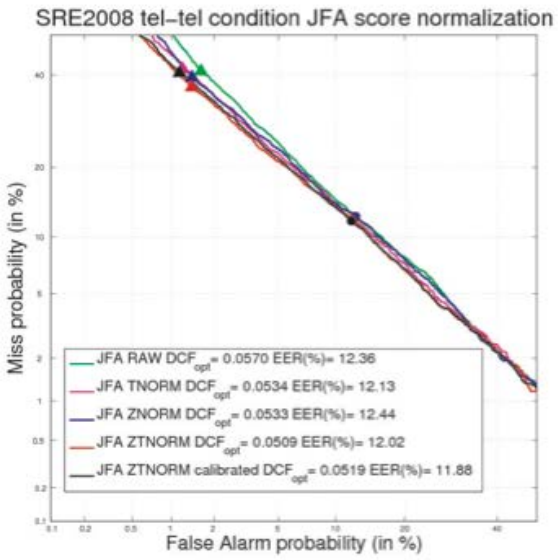


(b)

Figure 4.3: *DET curves in SRE 2008 short2-short3 in tel-tel condition data. Equal Error Rate (EER) (circle dot) and optimum Detection Cost Function (DCF) point is also depicted (triangle point).* ***(a) Top****: Detection curves for both Gaussian Super Vector (GSV) SVM and push-back GMM versions compared to GMM-UBM system.* ***(b) Bottom****: Comparison for previous systems to MLP-SVM system based on connectionist transformation network features.*



(a) (b)



(c)

Figure 4.4: *Development of JFA spectral sub-systems. DET curves in SRE 2008 short2-short3 in tel-tel condition data. Equal Error Rate (EER) (circle dot) and optimum Detection Cost Function (DCF) point is also depicted (triangle point).* ***Left****: figure draws JFA sub-system's DET curves for different sub-space dimensions of eigenchannels and fixing eigenvoice sub-space to dimension* 350. ***Right****: figure draws JFA sub-system's DET curves for different sub-space dimensions of eigenvoices and fixing eigenchannel sub-space to dimension* 50. ***Bottom****: figure draws DET curves for JFA sub-systems joining different normalization techniques as Z-norm, T-norm, ZT-norm and ZT-norm augmented with calibration. Raw scores DET curve is also shown for comparison purposes.*

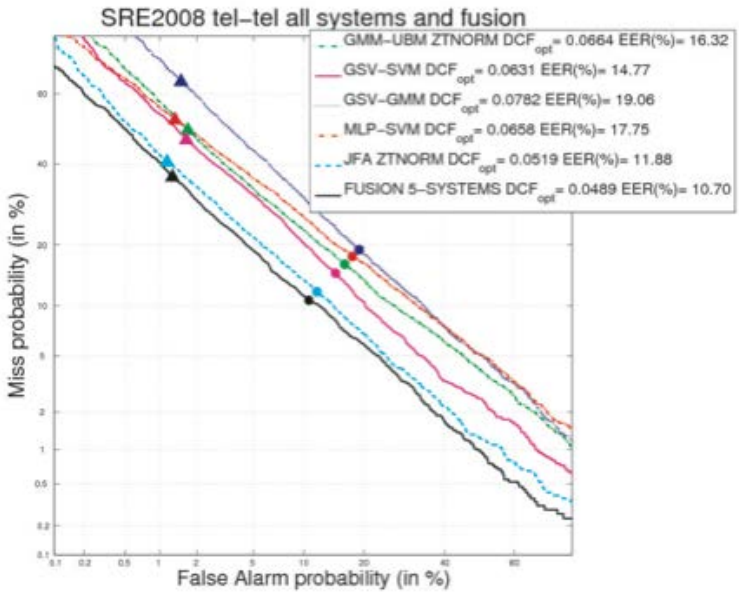


Figure 4.5: *DET curves in SRE 2008 short2-short3 in tel-tel condition data. Equal Error Rate (EER) (circle dot) and optimum Detection Cost Function (DCF) point is also depicted (triangle point). Development DET curves for the five sub-systems and the fusion calibrated system submitted to SRE 2010.*

for the original MLP-SVM system based on ASR transcriptions with same normalization and calibration procedure as previous systems. The results reported are consistent with state-of-the-art performance: The GSV-SVM based system outperforms the results obtained by other implementations. It is worth to mention the promising performance obtained by the MLP-SVM system despite of the use of not well adapted acoustics models and transcriptions for this task.

The development results of the JFA spectral sub-systems are also reported in figures 4.4 (a)-(c). The figures (a) and (b) in 4.4 study the effect of eigenvoice and eigenchannel dimensions on the performance of the system. As we can observe, the best results on SRE'08 data were obtained by setting the sub-space dimension of eigenvoices to 350 and the sub-space dimension for eigenchannels to 40. The figure 4.4 (c) depicts the DET curve comparison for the JFA spectral based-system for various normalization techniques. The enhancement of results by using score normalization techniques is obvious but it is not enough compared to those results reported in the literature [Burget *et al.*, 2008; Kenny *et al.*, 2007]. Furthermore, Z normalization does not seem to work properly since not significant improvement is perceived in the results. However, in the case of ZT-norm and ZT-norm joining together with logistic calibration, their results outperform those from scores without normalization. Note that the normalization sets for Z-norm and ZT-norm remain the same in both cases.

Finally, the figure 4.5 reports the DET curve, the EER and the DCT points for the five sub-systems on the SRE'08 tel-tel data and the comparison to the final fused system. The last was submitted to the SRE 2010

evaluation. Following we outline the results.

Evaluations Results in SRE 2010

The table 4.5 summarizes the official results per each "common condition" (CC) in core test condition, that is, the single required condition. The JFA based on spectral features (I) obtained the best individual results whereas the fusion of all individuals sub-systems outperforms it.
Overall, the results in terms of EER(%) are as poor as we expected since our development was mainly focused on SRE'08 short2-short3 tel-tel and rounding 9% EER which is in agreement with CC5. Anyway, our results are far away from state of the art results or best-team results in SRE'10 evaluation [Martin and et al., 2008b], due to the fact of a not well adapted training corpus for UBM estimation as well as for the score normalization sets.
Another issue affecting most of the participants was the poor calibration showed by the systems. Maybe as consequence of the reduced set of trials for adapting the calibration to each condition in the SRE'10 data and to the new cost function [Martin, 2010]. The new operation point for SRE'10 evaluation makes systems work around very low False Alarm rates due the target prior of 0.001. In our case, the actual DCF of our systems showed such behavior in most of the cases resulting in normalized DCF greater than 1 as can be observed in the table 4.5.
The fusion of individual sub-systems outperforms the previous showing as each individual system can provide useful information. That is the case of the FUSION-1 system where the JFA-prosodic system provides an additional improvement to the spectral one. Anyway the fusion of the 4 sub-systems also provides a slight enhancement.

System	*CC1*	*CC2*	*CC3*	*CC4*	*CC5*	*CC6*	*CC7*	*CC8*	*CC9*
(I) JFA-spectral	13.72 0.98467 1	21.69 0.9996 1	16.47 0.99694 0.9939	17.87 0.98648 1	15.23 0.97175 0.98305	15.62 1 1	18.69 0.99721 1	7.112 0.98322 1	12.38 0.94014 1
(II) JFA-prosodic	28.96 1 1	36.56 1 1	34.20 1 1	28.82 1 1	27.54 1 1	32.09 1 1	35.46 1 1	25.25 1 1	28.11 1 1
(III) GSV-SVM	16.36 0.9698 6.0497	26.96 0.99602 1.2819	20.16 0.99694 1.1874	18.87 0.97802 2.0057	14.42 0.93785 1.3177	21.41 0.99723 1.2304	21.95 0.94933 1.2069	9.39 0.97987 1.0421	11.58 0.86552 0.96495
(IV) GSV-GMM	21.41 0.97649 0.99954	32.36 0.99589 1	30.24 0.99633 0.99878	25.76 0.98998 1	19.90 0.94774 0.96045	26.64 1 1.0353	25.46 0.97214 1	13.19 0.95973 0.98993	15.31 0.89876 1
(V) TN-SVM-NAP [1]	38.43 1 1	46.28 1 1	40.45 1 1	43.89 1 1	14.98 0.92328 1	16.27 1 1	43.57 1 1	18.03 0.9966 1	44.44 1 1
(I + II) FUSION 1	12.48 0.99164 1	21.81 1 1	15.82 0.99571 0.9988	15.72 0.99577 1	15.07 0.97034 0.9845	15.06 1 1	20.10 0.99443 1	6.75 0.97651 1	11.56 0.96897 1
FUSION 2 (I + II + III + IV)	12.37 0.97072 2.4069	22.46 0.99456 1.0277	**15.38** 0.99449 1.0091	**14.62** 0.97971 1.0644	13.75 0.9548 0.96045	14.75 1 1	**17.95** 0.9354 1.0078	**6.55** 0.95476 0.98993	9.66 0.8069 0.87241
ALL FUSION	**11.31** 0.97305 2.3737	**20.96** 0.99522 1.0406	16.19 0.99388 1.2955	14.94 0.98436 1.186	**13.23** 0.91243 0.96328	**11.50** 1 1.1384	18.05 0.89919 0.96657	8.54 0.99664 0.99664	**8.14** 0.7931 0.86897

Table 4.5: *Summary of results in 2010 data. For each system, the first line contains EER[%], the second line is the optimum DCF and the third one is the actual DCF both normalized by the target prior 0.001*

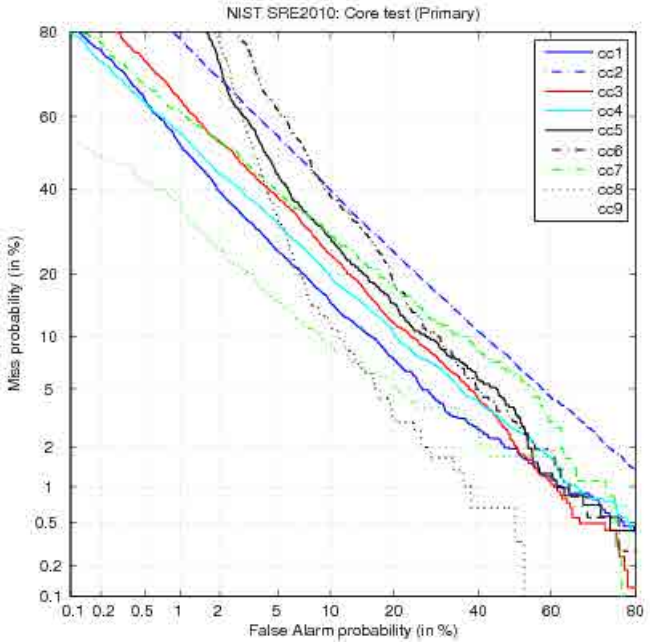


Figure 4.6: *DER official results on SRE'10 evaluation for the primary system. DET curves in SRE 2010 core condition data for each common condition.*

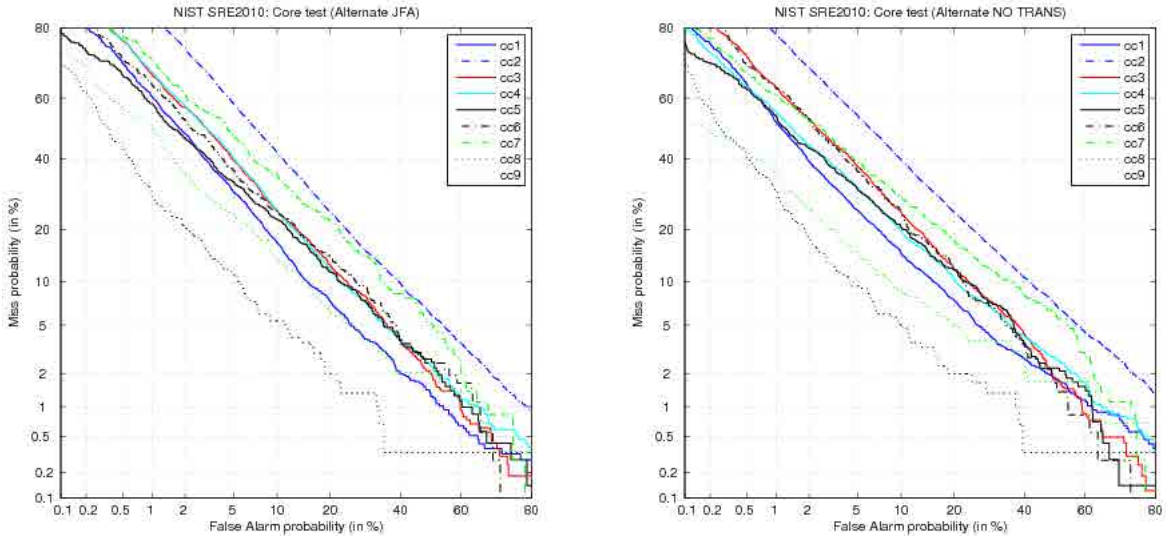


Figure 4.7: *Official results on SRE'10 evaluation for the two alternative systems. DET curves in SRE 2010 core condition data for each common condition.*

4.5 Conclusions

The speaker recognition teams of L^2F (from Lisbon, Portugal) and UPC (from Barcelona, Spain) presented a joint primary submission at the core condition of the NIST SRE 2010 campaign, consisting of the fusion of five different sub-systems. Additionally, two different combinations of the sub-systems that form the primary system have been presented as alternative contrastive systems. Time constraints made it impossible for us to submit results for the other evaluation conditions. We expect to evaluate our primary system in some of the alternative conditions as part of our post-evaluation work.

Our main objective in participating in this evaluation was to introduce ourselves to the speaker recognition community, to explore the recently proposed methods and to learn as much as possible. In this sense, independently of the final results, our participation was already quite successful. Additionally, the collaboration between two research groups from different countries was a nice achievement and we hope that can produce future fruitful collaborations.

Since it was the first participation for both L^2F and UPC at NIST SRE, most of our work prior to the evaluation was focused on the development and assessment of SR algorithms and methods. As a consequence, we could not devote enough attention to the new challenges proposed in the 2010 campaign. For instance, no special attention was given to "low vocal effort" challenge or to the problems introduced as a consequence of the new cost function.

One important limitation of the submitted system is that cross-channel problems have been little or not studied during the development. Most of the data used for development is telephonic (background, UBM, eigenchannels, eigenvoices, NAP...). In fact, in most cases SR experiments during the development of the sub-systems were performed only in the *tel-tel* condition of SRE 2008. Thus, we can expect a considerable better performance in the *tel-tel* condition compared to the other evaluation conditions.

Additionally, two different combinations of the sub-systems that form the primary system have been presented as alternative contrastive systems. Time constraints made it impossible for us to submit results for the other evaluation conditions. We also believe that significant improvements could be potentially obtained just by selecting carefully the sets for calibration and fusion, for normalization or compensation techniques. Some sub-systems could have been significantly improved. In fact, some modules were removed at the very last minute due to time problems and implementation difficulties. For instance, NAP was not applied to sub-system III, although it was in our initial plans. Neither we were able to submit ZT-norm scores of sub-system IV, having just applied z-norm. We also believe that significant improvements could be potentially obtained just by selecting a better calibration and fusion development set.

Part III

Speaker Diarization and Tracking

Chapter 5

Speaker Diarization in Meeting Domain

Speaker and speech recognition use similar speech signal processing techniques. However, speech recognition (if it is to be speaker independent), focuses on those aspects of the speech signal carrying more linguistic information, whereas speaker recognition is based on those idiosyncratic speech features that characterise an individual. Indeed, speaker recognition in continuous audio streams is refereed as speaker tracking, whether the speaker identity tracked is known, otherwise diarization task. The two approaches involve several processing stages, they are really close to each other and generally share some main components. Among them, there are various speech technologies such as audio and speaker segmentation, speaker clustering, speaker identification and speaker verification. The techniques from all of these areas are usually applied together to obtain the desired outcome at each time: *"Who is speaking?"*.

The main purpose of this chapter will be to develop a research on speaker diarization in meeting environment using as starting point the techniques and implementations briefly described in chapter 2. The research will focus on the interaction between the related topics looking for the global improving in real situations together with the on-line constrain. Therefore this chapter is devoted to explain the baseline diarization system, the development strategy and benchmark applied to assess the algorithm performance and some novelties applied to the system.

The remaining sections are organized as follows. In the section 5.1, a description of the main parts of the baseline system developed during this PhD thesis is highlighted. The diarization engine was benchmarked within the framework of NIST Rich Transcription evaluations. In order to focus on conference and meeting data, several databases from NIST evaluations have been used to perform experiments. In the section 5.2, the baseline system based on a single audio channel is augmented by the inclusion of multiple audio inputs. It allows to improve previous results by exploiting speaker source position as well as dealing with diarization overlap issues. In section 5.3, a newly published diarization approach is implemented based on spectral theory aiming to compare recent clustering algorithm with the classical agglomerative approach. Finally, the tuning of the diarization algorithm as well as the results of new techniques are reported in the section 5.4. The results are presented using data delivered for the last three Rich Transcription evaluations campaigns, aiming to reach

statistical significance by agglomerating around 24 different recordings from several sites and speakers.

5.1 AHC Single Channel Diarization

Speaker diarization refers to the systems performing speaker segmentation of the input signal and then speaker clustering of the created segments into a homogeneous groups, all within the same input stream. The diarization task assumes no prior knowledge about the speakers or how many people participate in the meeting. In order to get acquainted with the problem, the data and the evaluation methodology, we have taken as a baseline a simplified version of the International Computer Science Institute (ICSI) RT'06 system as presented in [Anguera *et al.*, 2006c].

The system currently used at ICSI, which has been used as a base for our diarization system, was originally created by Jitendra Ajmera around 2003 to perform speaker diarization in broadcast news (BN) data. The system was built while he was a PhD student at EPFL (Lausanne, Switzerland) and IDIAP (Martigny, Switzerland) and implemented it at ICSI while visiting for 6 months. ICSI participated in several NIST evaluations on BN including NIST 2003 Rich transcription of broadcast news spring evaluation and the RT'04 evaluation (Wooters et al. 2004). Afterwards, it was improved and adapted to meeting domain by Xavier Anguera during his stay at ICSI as a PhD student from UPC. The system has been maintained and improved during the elaboration of this PhD. It has also been submitted on the following NIST evaluations focused on meeting data, specifically in conference domain:

- The Rich Transcription Spring 2007 Evaluation (RT'07) focused on the English Meeting Domain speech. The cross site evaluation corpora included conference room meetings and lecture room meetings.
- The Rich Transcription Spring 2009 Evaluation (RT'09) focused on the English Meeting speech. Including just conference domain data.

The system is a bottom-up agglomerative clustering approach that uses a modified version of the BIC distance [Ajmera and Wooters, 2003] in order to iteratively merge the closest clusters until the same BIC distance determines the system to stop merging. Speaker segmentation of the data is not explicitly done before the clustering part, as step-by-step approaches do, but it is done via Viterbi decoding of the data given the current speaker models at every iteration. For a thorough description of the system refer to [Ajmera and Wooters, 2003; Anguera, 2006]. The philosophy behind the system and all research that has been done towards its improvement is based on the same key concepts as previous developers:

- Make the system as robust as possible to data within the same domain.
- Allow for a fast adaptation of the system to use it in new domains (i.e. broadcast news, meetings, telephone speech, and others).

These key concepts were put into practice by imposing the following guidelines:

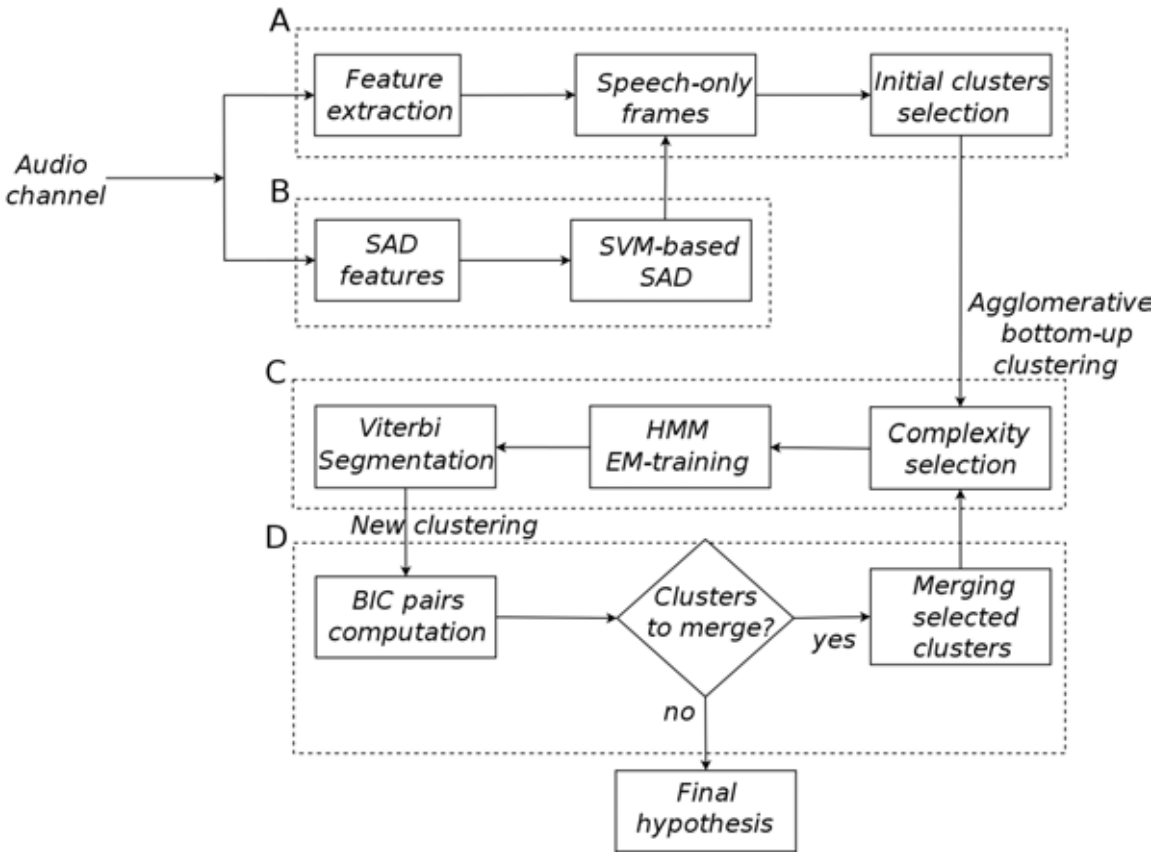


Figure 5.1: *Brief scheme of the implementation of the single channel diarization system.*

- Use as few training data as possible so the system can be easily adapted to new domains and it is not over-tuned to the data it is trained on.
- Avoid the use of thresholds and tuning parameters as much as possible. If not, try to define parameters that once tuned can achieve good performance in different kinds of data.

The figure 5.1 depicts the main blocks constituting the diarization core system that will be used as a baseline in the experiments section. In the following sections a detailed description of the different blocks is given. It differs from the previous diarization system by Ajmera and Anguera in several points: First, the inclusion of a speech/non-speech detector developed at UPC based on SVM classifier in order to filter out the non-speech segments prior to doing any further processing to the data; second, the automatic initial selection of clusters which aims to ensure statistical significance in the GMM model parameter estimation as in [Imseng and Friedland, 2010]. Furthermore, in order to speed-up the agglomerative clustering process, we have adopted several strategies: A merging rule based on a threshold in the standard deviation of BIC values which allows the system to merge more than one cluster at each iteration, a look-up table for the logarithm function based on [Vinyals and Friedland, 2008] and a multi-threading version based on MPI [Snir *et al.*, 1998] for parallel

computations. In the multiple channel approach (known as MDM condition) for meetings, the system still uses the multi channel capabilities developed by Anguera et al., as the use of a beamformed channel [Anguera *et al.*, 2007a] and the estimated position feature of speaker by means TDOA as in [Pardo *et al.*, 2007]. Moreover, some novelty contributions are included as the initialization of the clustering based upon estimated speaker positions as in [Luque *et al.*, 2008b].

5.1.1 Front-end Processing

The speech parameterization is based on a short-term estimation of the spectrum energy in several sub-bands. The speech channel is analyzed in frames of 30 milliseconds at intervals of 10 milliseconds and 16 kHz of sampling frequency. Each frame window is processed subtracting the mean amplitude from each sample. A Hamming window was applied to each frame and a FFT computed. The FFT amplitudes were then averaged through overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale.

Two main sets of features have been studied, one of them based on Mel-Frequency cepstral coefficients (MFCC) and the other based on the Frequency Filtering approach. The frequency filtering squeme we apply follows the classical procedure used to obtain the Mel-Frequency Cepstral Coefficients (MFCC), however in this approach, instead of using Discrete Cosine Transform, such as in the MFCC procedure [Davis and Mermelstein, 1980] log filter-bank energies are filtered by a linear and second order filter. The filter $H(z) = z - z^{-1}$ has been applied for some experiments over the log, of the filter-bank energies. The shape of this filter allows a best classification due it emphasizes regions of the spectrum with high speaker information yielding more discriminative information. These parameters have shown a good results in the last CLEAR Evaluation Campaign in the acoustic person identification task [Luque and Hernando, 2008a] and its choice is based on the fact that the use of the FF instead of the classic MFCC has shown the best results in both speech and speaker recognition [Nadeu *et al.*, 2001].

The use of complementary coefficients Δ and $\Delta\Delta$ parameters have been suppressed from the formation of

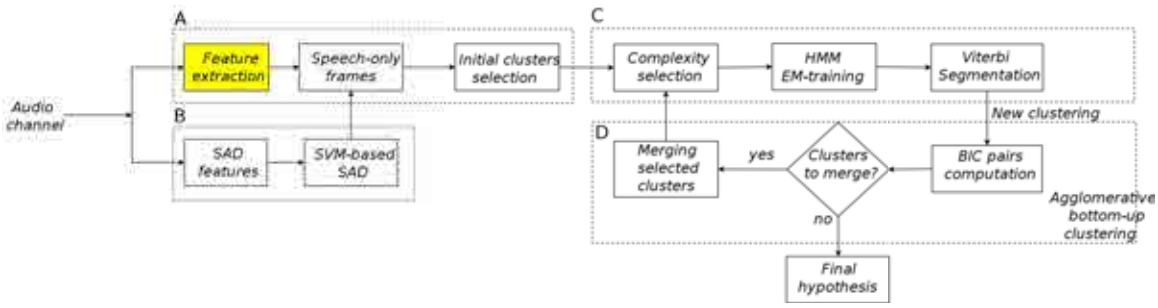


Figure 5.2: *Brief scheme of the implementation of the single channel diarization system. The highlighted box corresponds to the feature extraction module. It computes a set of parameter, usually based on spectral information as MFCC, in order to extract useful and compact information from the audio signal.*

the feature vector applied in speaker diarization and tracking experiments. Their application did not report significant improvements, therefore, a set of static spectral coefficients are usually applied to perform most of the experiments.

SVM-based Speech Activity Detection

The Speech Activity Detector (SAD) applied in this work is based on the SVM classifier [Schlkopf and Smola, 2002]. The system has been developed at UPC and it has shown a good performance in the last RT SAD Evaluations [Temko *et al.*, 2007]. The algorithm is based on Proximal SVM (PSVM) [Fung and Mangasarian, 2001] and on a fast training technique which allows the training of huge amounts of data. Additionally, the SAD algorithm makes use of a cross-validation technique to select those frames which show higher speech/non-speech detection accuracy. Finally, such frames were used to train a classical SVM model which we applied to obtain the speech/non-speech segmentation.
A set of several hundred of thousand of examples is a usual amount of data for classical audio and speech processing techniques that involves GMM. Nevertheless, it is an enormous number of feature vectors to be used for a usual SVM training process and it makes challenging such training feasible in practice. Alternative methods should be effectively applied to reduce the amount of data.
Proximal Support Vector Machine (PSVM) has been recently introduced in [Fung and Mangasarian, 2001] as a result of the substitution of the inequality constraint of a classical SVM $y_i(wx_i + b) \geq 1$ by the equality constraint $y_i(wx_i + b) = 1$, where y_i stands for the label of a vector x_i, w is the norm of the separating hyperplane H_0, and b is the scalar bias of the hyperplane H_0. This simple modification significantly changes the nature of the optimization problem. Unlike conventional SVM, PSVM solves a single square system of linear equations and thus it is very fast to train. As a consequence, it turns out that it is possible to obtain an explicit exact solution to the optimization problem [Fung and Mangasarian, 2001].
The proposed algorithm of dataset reduction consists of the following steps:

- Step 1. Divide all the data into chunks of 1000 samples per chunk

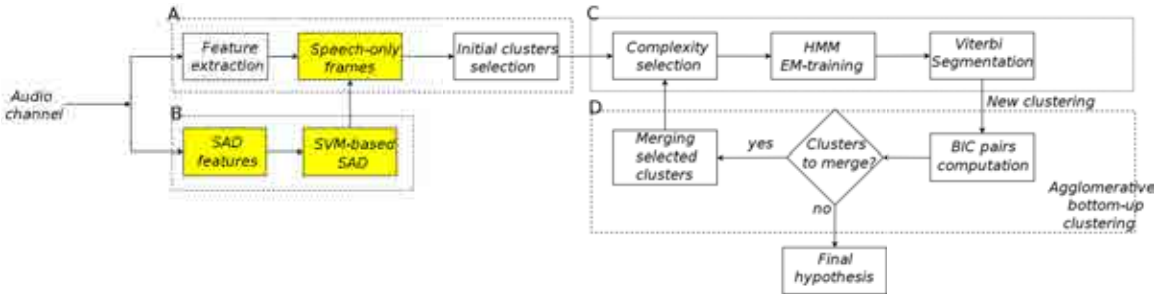


Figure 5.3: *Brief scheme of the implementation of the single channel diarization system. The highlighted boxes correspond to SVM-based speech activity detector which is in charge of detecting speech frames.*

- Step 2. Train a PSVM on each chunk performing 5-fold cross-validation (CV) to obtain the optimal kernel parameter and the C parameter that controls the training error
- Step 3. Apply an appropriate threshold to select a pre-defined number of chunks with the highest CV accuracy
- Step 4. Train a classical SVM on the amount of data selected in Step 3

The proposed approach is in fact similar to Vector Quantization (VQ) used for dataset reduction for SVM in [Lebrun *et al.*, 2004]. With Step 2 some kind of clustering is performed, and Step 3 chooses the data that corresponds to the most separable clusters. However, unlike VQ, SVs which are obtained with the proposed algorithm in Step 4 are taken from the initial data. Besides, additional homogeneity is achieved because the PSVM data clustering is performed in the transformed feature space, through the transformation functions that correspond to the Gaussian kernel. Finally, the same kernel type is applied to the chosen data in Step 4.
The feature set used is mainly based on Frequency Filtering (FF) parameters. Its computation is divided into two parts. The first part, extracts information about the spectral shape of the acoustic signal in a frame. It is based on Linear Discriminant Analysis (LDA) of FF parameters. The size of the FF representation ($16\text{FF}+16\Delta\text{FF}+16\Delta\Delta\text{FF}+\Delta\text{E}= 49$) is reduced to a single scalar measure by applying LDA. The second part of the feature set focuses on the dynamics of the signal along the time, observing low- and high-frequency spectral components. The contextual information is involved in several ways. First, before applying the LDA transform, the current delta and delta-delta features involve an interval of 50 and 70 ms, respectively, in their calculation. Next, for the representation of the current frame, eight LDA measures are selected from a time window spanning the interval of 310 ms around the current frame. Finally, low and high frequency dynamics involve a smoothed derivative calculation that uses 130 ms interval. The first and the second part of the feature set form a vector of 10 components. Additionally, a cross-frequency energy dynamic feature, which is obtained as a combination of low and high frequency dynamics, is added to the final feature vector. A more accurate description of the features employed is given in [Macho *et al.*, 2006].
The SVM based SAD system applied in this PhD. proposal was trained with the RT 2005, 2006 and 2007 conference data, the CHIL 2007 meeting data and the Speecon (far-field microphone) data. It yielded to more than 25 hours of training material.

5.1.2 Clustering Initialization

At the beginning of the clustering algorithm, a uniform initialization is performed so the system starts with an homogeneous splitting of the whole data among the initial number of clusters (see figure 5.4 block A). The number of initial clusters is determined automatically depending on the meeting length with minimal and maximal value constraints. In this PhD. proposal, the total amount of clusters was constrained to a minimum and a maximum of 35 and 85 clusters respectively, aiming to avoid overclustering and to reduce the computational cost of the iterative approach. The automatic selection of the number of clusters (K_{init}) is defined as,

$$K_{\text{init}} = \frac{N}{G_{\text{init}}\, R_{\text{CC}}} \tag{5.1}$$

The previous expression takes into account the total amount of data available per speaker cluster (N), the number of Gaussian mixtures initially assigned to each speaker cluster (G_{init}) and the cluster complexity ratio (R_{CC}). The R_{CC} is a constant value across all meetings that defines the number of frames per Gaussian. It was fixed to 7 seconds of speech per Gaussian whereas the initial number of Gaussians per model (G_{init}) was set to 5. In addition, the total amount of clusters was constrained to a minimum and a maximum values respectively, aiming to avoid overclustering and to reduce the computational cost of the iterative approach. Moreover, a method to reduce manual tuning of these values [Imseng and Friedland, 2010] is also implemented. The solution reduces the sensitivity of the initialization values and therefore reduces the need for manual tuning significantly while at the same time increases the accuracy of the system.

The figure 5.5 presents the results plotted as the number of seconds per Gaussian vs the Speaker Error for different durations of segments. By tuning the seconds per Gaussian parameter yields low speaker error even in short meetings. It also can be observed that the optimal amount of speech per Gaussian used for the training procedure seems to roughly follow a curve that has a global minimum. The estimation method balances the relationship between the optimal value of the seconds of speech data per Gaussian and the duration of the speech data. In [Imseng and Friedland, 2010] the authors use a linear regression to estimate R_{CC},

$$R_{\text{CC}} = 0.01 \cdot \text{speech in seconds} + 2.6 \tag{5.2}$$

fixing $G_{\text{init}} = 4$ and then using the linear regression to estimate K_{init} using 5.1, leads to improved results. Such a strategy produces relative improvements of up to 50% for very short meeting segments (100 seconds) while maintaining the performance of the system for long recordings (600-700 seconds).

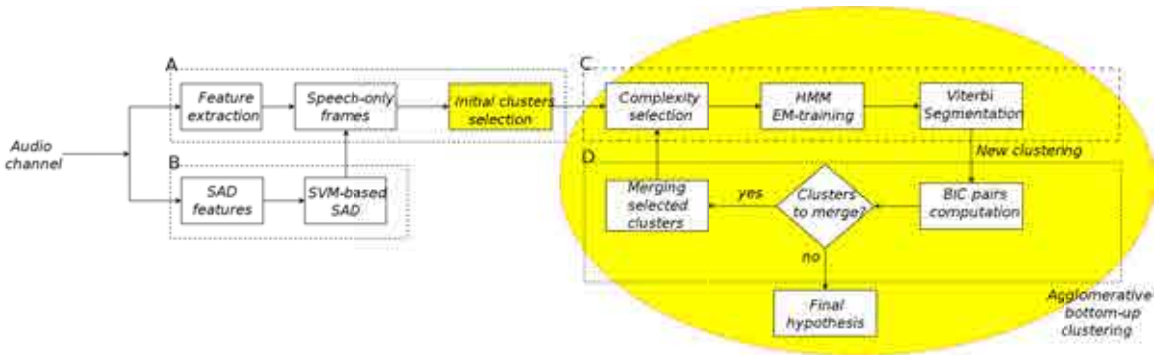


Figure 5.4: *Speaker diarization scheme. The left highlighted box is in charge of selecting the initial number of clusters (classes) to feed the agglomerative approach. The right highlighted side corresponds to the agglomerative hierarchical clustering stage by itself.*

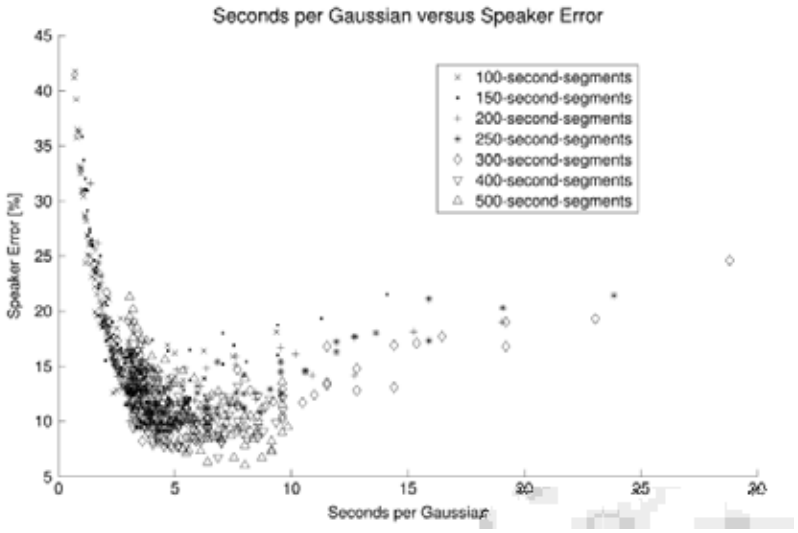


Figure 5.5: *Speaker Error versus seconds per Gaussian. Each data point corresponds to the average speaker error of 12 meetings (2.05 hours of data) for one particular configuration. Configurations for all tested segment durations are depicted in the same plot. One can recognize a combination of curves, the minimum seems to be similar for different recording durations. Image from [Imseng and Friedland, 2010].*

5.1.3 HMM-based Agglomerative Hierarchical Clustering

Our speaker diarization system follows the commonly used agglomerative hierarchical clustering (AHC) approach as explained in chapter 2 in section 2.3.2. Firstly, speech is broken into short uniform segments. Such clusters are modeled by mixture of Gaussians and the successive clustering stage groups acoustically similar segments, assigning them to speaker clusters based upon a Bayesian information criterion (BIC) metric among Gaussian distributions. The figure 5.4 depicts an overall scheme of the diarization system submitted to Rich Transcription (RT) 2007 and 2009s evaluations [Luque *et al.*, 2008a]. The main stages of the diarization can be condensed in the following points:

- Feature extraction and removal of non-speech frames. At this stage, a clustering initialization is also performed based on an homogeneous partition of the data (see figure 5.4 block A).

- Model complexity selection based on the amount of data per cluster and the cluster complexity ratio (CCR), which fixes the amount of speech (seconds) per Gaussian. A HMM/GMM training and cluster realignment by Viterbi decoding based on maximum likelihood (see figure 5.4 block B).

- Agglomerative clustering based on the Bayesian information criterion (BIC) metric among clusters. The stopping criterion, also based on the BIC, drives the ending point of the algorithm (see figure 5.4 block C).

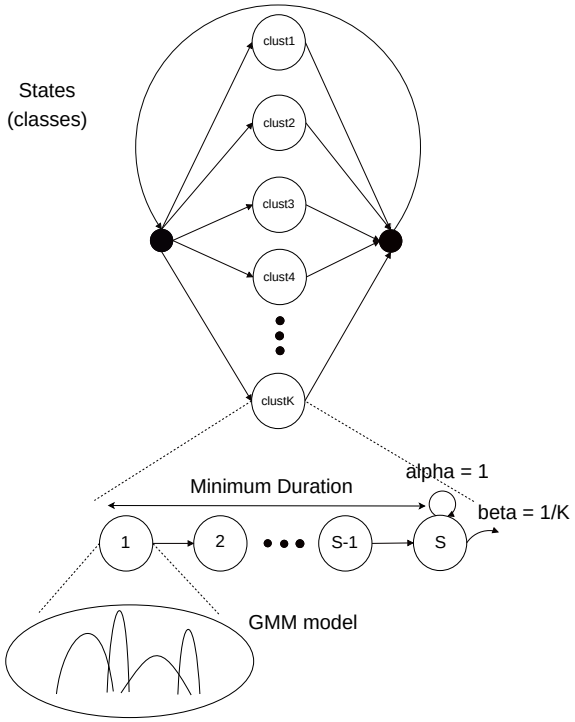


Figure 5.6: *The clusters in the diarization algorithm are modeled by means of an ergodic HMM/GMM with a minimum duration constrain which ensures a minimum length in the speaker turn duration.*

Once the initial segmentation is performed, each cluster is modeled by one mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm (see figure 5.4 block B).

It follows an iterative bottom-up strategy driven by a loop of BIC estimations and HMM alignments (see figure 5.4 block C). In this step the segments which belong to the same speaker are combined in a new model at each iteration. A time constraint as in [Ajmera and Wooters, 2003] is also imposed on the duration of the speaker segments through a hierarchical modeling of each state, see figure 5.6. In that sense, Viterbi decoding decisions are taken based on the estimation of the observation probabilities by accumulating the likelihoods per cluster/state in a 3 seconds window.

This procedure is iterated until the stopping criterion is reached. It is met whenever all the remaining set of BIC cluster-pairs show negative values, meaning that no suitable candidates are found to merge and consequently the algorithm ends. Finally, at the last iteration and once the stopping criterion is met, each remaining state represents a different speaker. A more detailed description of the system can be found in [Luque *et al.*, 2008a].

The performance of the speaker diarization was evaluated by means of the diarization error rate (DER) as defined by NIST [Fiscus and et al., 2007a]. The DER is a time-weighted metric composed of the sum of missed speaker time, false alarms and speaker error time as explained in chapter 2 in section 2.3.3.

Automatic Model Complexity Selection

At each iteration j, the number M_i^j of Gaussian mixtures to model the cluster i is updated by

$$M_i^j = \left\lfloor (\frac{N_i^j}{R_{\mathrm{CC}}}) + \frac{1}{2} \right\rfloor, \tag{5.3}$$

where N_i^j is the number of frames belonging to the cluster i. Whenever two segments are merged a new segment model is also trained pooling all the features from the merged segments and fixing the model complexity according to the R_{CC} value. Such automatic selection of the modeling complexity has demonstrated a successful performance while avoiding the use of the penalty term in the classical BIC metric [Anguera *et al.*, 2006c].

Iterative Viterbi Segmentation

The agglomerative clustering algorithm models the acoustic data using an ergodic hidden Markov model (HMM), where state corresponds to one of the initial clusters. Lets $X = \{x_1, x_2, ..., x_N\}$ the audio data to be segmented, we want to find the optimal number of clusters k^* and their acoustic models $\boldsymbol{\theta}_k^*$ that produce the "best" segmentation, in likelihood sense, of the data (X) according to:

$$\boldsymbol{\theta}_k^*, k^* = \underset{\boldsymbol{\theta}_k, k}{\mathrm{argmax}} \{\Pr(X, p_{best} | \boldsymbol{\theta}_k, k)\}, \tag{5.4}$$

where p_{best} is the Viterbi segmentation path with the highest likelihood, that is, a sequence of states/models which produce the maximum likelihood given the observations. Upon completion of the algorithms execution, each remaining state is considered to represent a different speaker. This step aims to refine the data partition obtained by the agglomerative clustering and improves the speaker segment boundaries [Tranter and Reynolds, 2006]. Thus, we want to find the set of clusters and their acoustic models that maximize the likelihood of the data and the associated segmentation based on this HMM topology. Since we do not want to consider all possible values for k, we start choosing a maximum value ($k = K$) by means an initial segmentation. Then, through the process of cluster merging, we reduce the value of k until we find an optimal number of clusters k^* and their acoustic models $\boldsymbol{\theta}_k^*$ according to equation 5.4.

In addition, a minimum duration (MD) constrain have been imposed in the HMM topology, see figure 5.7, which ensures a minimum length in the speaker turn duration. Each state in the HMM is composed by a set of sub-states, as seen in figure 5.7, imposing a minimum duration of each model. Each one of the sub-states has a probability density function modeled via a Gaussian mixture model (GMM). The same GMM model is tied to all sub-states in any given state. Once entering a state, at time n the model forces a hop to the following sub-state with probability 1.0 until the last sub-state is reached. In that sub-state, it can remain in the same sub-state with transition weight α, or jump to the first sub-state of another state with weight $\frac{\beta}{K}$, where K is the number of active states/clusters at that time. A justification for this values of the transition probabilities, β and α, in the chain of sub-states is given in [Anguera *et al.*, 2006d], where a disadvantage arises by using

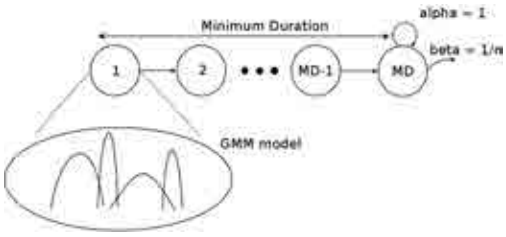


Figure 5.7: *Minimum duration constrain, which ensures a minimum length in the speaker turn duration, by stacking a set of substates on each state.*

$\beta + \alpha = 1$ as an artificial penalization of long speaker turns against turns with minimum MD frames.
In order to illustrate such a drawback, lets define the likelihood performed by an observation sequence $Y = \{y_1, y_2, \ldots, y_{2MD}\}$ of 2 times MD of duration in the case the speaker change does not occur, $lkld_{AA}$, and in the case there is a speaker change turn between A and B, $lkld_{AB}$:

$$lkld_{AA} = \Pr(y(0) \mid \Theta_A) \prod_{i=1}^{MD-1} (1 \cdot \Pr(y(i) \mid \Theta_A)) \cdot \prod_{i=MD}^{2MD-1} (\alpha \cdot \Pr(y(i) \mid \Theta_A)) \tag{5.5}$$

$$lkld_{AB} = \Pr(x(0) \mid \Theta_A) \prod_{i=1}^{MD-1} (1 \cdot \Pr(y(i) \mid \Theta_A)) \cdot \frac{\beta}{K} \Pr(y(MD) \mid \Theta_B) \prod_{i=MD+1}^{2MD-1} (1 \cdot \Pr(y(i) \mid \Theta_B)) \tag{5.6}$$

in which $y(i)$ stands for the acoustic observation i and Θ_j stands for the model of the j state. Equation 5.5 shows the computed likelihood given 2 times MD acoustic frames and remaining in cluster A during all of them. In contrast, equation 5.6, shows the total likelihood if we jump to a model B after the initial MD frames. The only way to ensure that equation $lkld_{AA} > lkld_{AB}$ given $A = B$ is setting $\beta = \alpha = 1$, [Anguera *et al.*, 2006d]. Thus, once a segment exceeds the minimum duration, the HMM state transitions no longer influences the turn length; turn length is solely governed by acoustics.
Whenever two clusters are merged, the total number of parameters in the HMM decreases. Modeling the same amount of data using fewer parameters yields a lower likelihood score. Given that the merging process can only result in monotonically decreasing likelihoods, we will not observe a maximum in the likelihood function 5.4 at any point rather than the starting point. Therefore we need to choose a likelihood threshold to tell us when to stop merging.

5.1.4 Merging Clusters and Stopping Criterion

Once the data has been segmented into many small pieces and each piece is assigned to a cluster, the system then iteratively merges clusters and stops when there are no clusters that can be merged. This procedure requires two different metrics: one to determine which pair of clusters to merge, and a second measure to

determine when to terminate the merging process. We used a modified BIC-based metric [Ajmera and Wooters, 2003] to decide in both cases: the most likely-pair of clusters to merge and the stopping point.
Ideally, we would like to find a method of selecting similar clusters which results in an increase of the objective function 5.4 when the pair of right clusters [1] are merged and an incorrect merge will result in a decrease. A common competing models selection method is to use the Bayesian Information Criterion (BIC).
Assuming that there are two clusters/segments to compare, the problem is to decide whether such two segments are uttered by the same speaker. Let $Z = X \cap Y$ and N_X, N_Y, N_Z be the numbers of samples in clusters X, Y and of their union cluster Z, respectively. Obviously, $N_Z = N_X + N_Y$. The modified BIC equation is defined as:

$$\Delta BIC(Z) = \quad BIC(X,Y) = \quad L_0 - L_1 \lesseqgtr 0, \tag{5.7}$$

which does not make use of the penalty term, which corresponds to the number of free parameters of a multivariate Gaussian process, see section 2.3.1. L_0 is defined as the log-likelihood performed by a model $\boldsymbol{\theta}_Z$ which takes into account the whole data Z and stands for the parameters of the Gaussian distribution, i.e., the mean vector μ_z and the full covariance matrix $\boldsymbol{\sigma}_Z$. Whereas L_1 is defined by the sum of log-likelihoods performed by two independent models $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Y$ on each data cluster X and Y, respectively:

$$L_0 = \sum_{i=1}^{N_x} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Z) + \sum_{i=N_x+1}^{N_z} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Z) \tag{5.8}$$

where $z_i \in \Re^d,\ i = 1, 2, \ldots, N_z$ which are assumed to be independent vector of acoustic features. The segments X and Y are modeled by distinct multivariate Gaussian densities, whose parameters are denoted by θ_X and θ_Y, respectively.

$$L_1 = \sum_{i=1}^{N_x} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_X) + \sum_{i=N_x+1}^{N_z} \log \Pr(\mathbf{z}_i|\boldsymbol{\theta}_Y) \tag{5.9}$$

In the case $BIC(X,Y) > 0$, it implies that the clusters X, Y are best modeled by one independent model $\boldsymbol{\theta}_Z$ rather than by two independent models $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Y$. That is, the clusters X and Y are candidates for merging. The equation 5.7 is similar to traditional BIC criterion, except by the lack of the penalty term and that the model $\boldsymbol{\theta}_Z$ is constructed in such a way that the number of parameters is equal to the sum of the number of parameters in $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Y$. By keeping constant the number of parameters on both sides in equation, we have eliminated the traditional BIC penalty term. Selecting candidates for merging using this criterion does indeed result in an increase in the objective function associated with 5.4. This increases the robustness of the system as there is no need to tune this parameter [Ajmera and Wooters, 2003].
The segmentation obtained at the output of the block B) (see figure 5.4) defines a new set of speaker

[1] That is, a merge involving clusters of data from the same speaker

clusters/states which will be retrained. After every new segmentation-training step, we look for the pair (X, Y) satisfying $BIC(X, Y) > 0$. In the case of many candidate pairs, we choose the pair that maximizes $BIC(X, Y)$. This method provides a fully automatic stopping criterion that does not require the use of any tunable parameters. However, there are a few hyper-parameters in this algorithm, namely the initial number of clusters (K), the initial number of Gaussian components in each cluster (M), the type of initialization used to create the clusters, and the set of acoustic features employed to parameterize the signal.
In the case of the stopping criterion, we decide to halt the iterative procedure based on the BIC values of the remaining cluster-pairs. Once we do not find a couple of cluster X, Y with $BIC(X, Y) > 0$ we stop the merging and the final clustering hypothesis is provided by the system . In the final clustering the remaining classes are considered as different speakers.

Multiple Cluster Merging Criterion

Most of the systems based on agglomerative clustering perform just one merge at each BIC iteration, in which they choose to merge those couple of clusters with higher BIC value. Since the computational cost of the agglomerative clustering increases roughly with the square of the number of clusters, we have explored strategies in order to increase the number of initial clusters without significantly impacting the run-time and performance. Instead of selecting the cluster-pair with higher BIC value as the merging candidate, a threshold is applied depending on the standard deviation of the set of BIC value obtained among the whole set of cluster-pairs. This strategy leads to a set of cluster-pairs candidates for merging instead of just one candidate. Therefore, the system might merge more than one pair of clusters per iteration yielding to a speed up in the agglomerative clustering. We expect that using this merging approach allows us to start with a much larger number of initial clusters without dramatically increasing the run-time or degrading the performance. In general, we decide to merge all cluster-pairs (X, Y) fulfilling:

$$BIC(X, Y) > BIC_{\mu} + \frac{3}{2} BIC_{\sigma} \tag{5.10}$$

where $BIC(X, Y)$ is the BIC value between the clusters X and Y, BIC_{μ} is the mean of $BIC(X, Y)$ for $X \neq Y$ and BIC_{σ} the standard deviation for the same set, that is, the mean and standard deviation of the whole set of BIC measures.

5.1.5 Inclusion of a Turn Taking modeling

The modeling of the turn taking or social interactions in a multi-party conversation has been addressed in several works in the past becoming an active field [Thomas P. Wilson and Zimmerman, 1984]. Recently, speaker roles and dynamics within the conversation and social interactions has attracted special attention from researchers [Valente *et al.*, 2011; Laskowski and Shriberg, 2012].
Turn-taking models use a truncated representation of past speech activity to specify how likely the speaker is to talk at the next instant. We study this question using the NIST RT database and the algorithms we developed

and presented in the Rich Transcription evaluation in 2009.

The turn taking modeling presented proposes to model the speaker sequence using n-grams of speakers occurrences which can be then combined with the acoustic information coming from MFCC features. The approach is largely inspired by the current Automatic Speech Recognition (ASR) framework where the acoustic information from the signal, i.e., the acoustic score, is combined with the prior knowledge from the language, i.e., the language model. The most common form of language model is represented by words n-gram.

Let us consider the meeting as a sequence of speaker turns, i.e. speech regions from same speaker, uninterrupted by pauses longer than 50 ms:

$$T = \{(t_1, \Delta t_1, s_1), \ldots, (t_N, \Delta t_N, s_N)\} \tag{5.11}$$

where t_n is the starting time of the n-th turn, Δt_n is its duration, s_n the speaker label and N the total number of turns in the recording. Once we obtain an initial estimation of the number of speakers in the meeting, that is, a partition of the data in clusters – see section 5.1.2 – we also get an estimation of the speaker sequence:

$$T^* = \{(t_1^*, \Delta t_1^*, s_1^*), \ldots, (t_N^*, \Delta t_N^*, s_N^*)\} \tag{5.12}$$

In our modeling we do not take into account the starting time and duration of the speaker turn, thus we are only interested on the sequence S of speakers occurrences. Such a sequence S can be modeled using n-grams of $\Pr(s_n|s_{n1}, \ldots, s_{np})$, i.e., the probability of the speaker n depends on the previous p speakers in the sequence, which is called the context. The duration of the mentioned context limits the computation of the probability of the sequence. Therefore the probability of a sequence S can be written as:

$$\Pr(S) = \Pr(s_1, \ldots, s_n) = \Pr(s_1, \ldots, s_p) \prod_{n=p}^{N} \Pr(s_n|s_{n-1}, \ldots, s_{n-p}) \tag{5.13}$$

After the initial clustering, the speaker sequence is re-estimated using an ergodic Hidden Markov Model/Gaussian

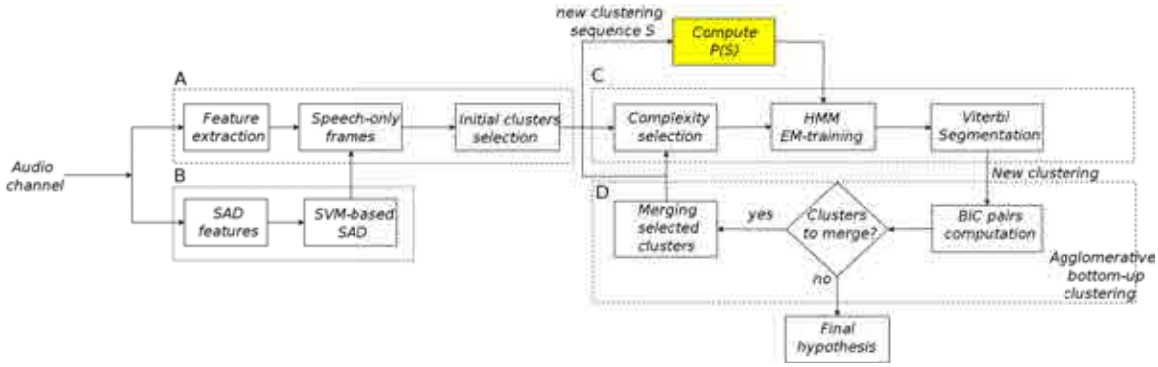


Figure 5.8: *Speaker diarization scheme, language modeling inclusion in the HMM/GMM decoding.*

Mixture Model where each state represents a speaker. The maximum likelihood sequence is decoded using a conventional Viterbi algorithm which implements a minimum duration constraint – see section 5.4. The optimal sequence is obtained maximizing the following likelihood, omitting the k number of clusters and the speaker model $\boldsymbol{\theta}$:

$$\boldsymbol{S}^* = \underset{\boldsymbol{S}}{\operatorname{argmax}}\{\Pr(X|\boldsymbol{S})\}, \tag{5.14}$$

and also neglecting the fact that not all speaker sequences S have the same probability. This new type of information can be included extending the maximization as :

$$\boldsymbol{S}^* = \underset{\boldsymbol{S}}{\operatorname{argmax}}\{\Pr(X|\boldsymbol{S})\Pr(\boldsymbol{S})\}, \tag{5.15}$$

in which the probability $\Pr(S)$ can be estimated from 5.13. This is somehow similar to what is done in Automatic Speech Recognition (ASR) where sentences (i.e. word sequences) are recognized combining acoustic information together with linguistic information captured in the language model (n-gram of words). Recalling previous equations 5.5 and 5.6 which stand for the probability of remain in state A (same speaker) or change from state A to state B (change the speaker turn) based uniquely on the acoustics, we can introduce the language model information as in equation 5.15 and solve it by means of the same Viterbi algorithm. In that sense, a speaker change occurs whenever the following inequality is fulfilled:

$$\begin{aligned}(1-LM)\cdot lkld_{AB} + LM\cdot\log(\Pr(s_p, s_{p-1}, \ldots, s_A, s_B)) > \\ > (1-LM)\cdot lkld_{AA} + LM\cdot\log(\Pr(s_p, s_{p-1}, \ldots, s_A, s_A))\end{aligned} \tag{5.16}$$

in which the maximization is expressed in the logarithm domain, the language/speaker transition probability $\Pr(S)$ depends on the p previous speaker turns and LM stands for the weight of the language model. As in ASR, the weight is introduced to scale $\Pr(S)$ values to comparable ranges with those obtained from acoustic observations. In addition, the LM weight is incorporated once the computation of the acoustics probabilities is performed ensuring the Minimum Duration (MD) constrain.
Several strategies were presented in the last RT evaluation in order to estimate the probability $\Pr(S)$. The most simple of them relies on the estimation of a transition matrix among clusters, that is, a 1-gram model equivalent to the number of occurrences of a cluster normalized by the total number of occurrences, i.e. the occurrence frequency of each cluster. Therefore we obtain a $k \times k$ matrix T, where k is the total number of clusters, which represents the probability of "jumping" from one cluster to other or just to remain in the same cluster. At each iteration of the agglomerative clustering, a new speaker sequence is obtained and the matrix T is updated with the information of the new clustering.

- **Unigram**: Count each transition among consecutive clusters and compute T.
- **Unigram updated with some trigrams**: Count the trigrams of the form ABA and increase the transition

probability from any speaker A to speaker B. That is inspired by the common behavior in a conversation with short speaker interruptions.

- **Trigrams weighted by interruption duration**: Depending on the duration of previous speaker interruptions, the transition probability of the speaker B is increased by a factor proportional to that duration. Speaker interruptions of 250ms and 150ms are taken into account in this approach.

5.1.6 Fast Logarithm and MPI Processing

In this subsection we give a brief idea of the implementation efforts carried out during the elaboration of this PhD thesis in order to speed up the agglomerative clustering approach. Two technical ideas were implemented: a fast logarithm based on a lookup table and the C++ code adaptation to MPI multi-core capabilities. First of them is inspired by the total amount of computational time spent by the AHC approach in computing logarithms, due to the fact that we are always working with log-likelihoods. The second is devoted to get benefit of the actual multi-core processors.
The proposed fast logarithm function is a fast single precision approximation of the natural logarithm with adjustable accuracy [Vinyals and Friedland, 2008]. Given an IEEE 754 floating point number, the main idea is to use a quantized version of the mantissa as a pointer into a lookup table. The amount of quantization of the mantissa determines the table size and therefore the accuracy. Current processors are able to store relatively large lookup tables in cache memory. Therefore an acceptable accuracy can be reached without too many main memory accesses.
Conceptually, a 32-bit IEEE 754 floating point number is stored as follows. A value V of a number is the product of a 23-bit mantissa m and an 8-bit exponent e. One bit is reserved for the sign, s. If $s = 0$ the sign is positive, otherwise it is negative. Since the real- valued logarithm is only defined for positive numbers, the sign bit can be ignored. We get:

$$V = 2^e \cdot m \tag{5.17}$$

We can use the multiplicative property of the logarithm function to decompose the logarithm computation as:

$$log_2(V) = log_2(2^e \cdot m) = e + log_2(m) \tag{5.18}$$

In order to calculate the natural logarithm, we can take advantage of the property that all logarithms are proportional to each other. This results in the following equation:

$$log_e(V) = (e + log_2(m))log_e(2) = e \cdot log_e(2) + log_2(m) \cdot log_e(2) \tag{5.19}$$

where $log_e(2) = 0.6931471805$ is a constant. Calculating the logarithm with respect to any other base only requires multiplying with a different constant. Extracting the exponent and the mantissa of a floating point number can be performed quickly using bit shift operations. Therefore, in order to calculate the left part of

the sum, only one multiplication is required. To calculate the right part of the sum, we store the results of the computation $log_2(m) \cdot log_e(2)$ in a lookup table. Unfortunately, this still requires a table with 2^23 entries with each entry needing 4 bytes, thus 32 MB. Nevertheless, using a table of this size increases the performance of the logarithm computation only very slightly since memory accesses take about the same time than the computation of the Taylor approximation [Vinyals and Friedland, 2008]. In order for the look up table to fit into cache, we quantize the mantissa, i.e. we ignore q least significant bits of the mantissa. The table is then indexed using the $23 - q$ most significant bits of the mantissa. The result is calculated by adding the value looked up in the table and the down scaled exponent. Accuracy is lost because of the quantization of the mantissa, however not a significant drop in DER performance is noticed while the computation time is reduced in a factor of 3 or more.

MPI is a library of message passing routines [Snir *et al.*, 1998]. The library allows a user to write a program in a familiar language, such as C, C++, FORTRAN77 or FORTRAN90, and carry out a computation in parallel on an arbitrary number of cooperating computers. Thus, this is the most remarkable feature: the user writes a single program which runs on all the computers. In addition to MPI, a High Throughput Computing (HTC) called Condor [Tannenbaum *et al.*, 2001] was also employed for computing distributively. Condor is a specialized workload management system for compute-intensive jobs. Like other full-featured batch systems, Condor provides a job queueing mechanism, scheduling policy, priority scheme, resource monitoring, and resource management. Users submit their serial or parallel jobs to Condor, Condor places them into a queue, chooses when and where to run the jobs based upon a policy, carefully monitors their progress, and ultimately informs the user upon completion.

5.2 AHC Multi Channel Diarization

The multi channel speaker diarization approach follows an strategy commonly applied in most of the state-of-the-art systems [Pardo *et al.*, 2012; Anguera *et al.*, 2011; Friedland *et al.*, 2011; Anguera *et al.*, 2007b]. In a multi-microphone environment, the use of redundant signals can improve the classical diarization systems. On the one hand, this information can be used for signal enhancement by applying a delay&sum algorithm as in [Anguera *et al.*, 2007a]. On the other hand, the speaker localization can directly perform the speaker segmentation and the clustering as in [Koh and et al., 2008] where a diarization based on Direction Of Arrival (DOA) information is proposed. In [Wooters and Huygbregts, 2008], the diarization is performed in an agglomerative way mixing the cepstral and the delay observations.

In this section, the previous described single-channel diarization is improved through the use of multiple microphone channels. It is achieved by means signal enhancement techniques like as Wiener filtering and signal beamforming, that mainly concentrate on obtaining a clean version of the speech wave or in focusing on the active speaker.

In addition, other information sources present in the speech, as the time delay of arrival (TDOA), are also employed. TDOA features are processed as a additional stream feature and combined jointly with MFCC

features leading to an strategy which has become one of the most popular system reported in the literature and in NIST Rich Transcription evaluations [Pardo *et al.*, 2007]. Moreover, the TDOA statistics and beahaviour are analyzed and their application in different parts of the diarization is demonstrated by **several original works** which apply them, e.g., to clustering initialization and speaker overlap detection.

5.2.1 Wiener Filtering

A Wiener filtering technique is applied on multiple distant microphone data (MDM). The combined use of Wiener filtering together signal beamforming has demonstrated to be a successful approach whereas in the previous case, in the single distant microphone (SDM) condition, no Wiener filtering is applied due no improvement was observed in experiments.

The noise reduction implementation from the QIO front-end [Adami *et al.*, 2002] is applied on each MDM microphone channel as can be seen in figure 5.9. In order to estimate the noise we also applied the same procedure than in the QIO front-end, in which the noise estimation is initialized from the beginning of each show and updated with those frames of the show selected as non-speech based on an energy threshold. Noise estimation is done in a causal fashion, so that each frame is noise-reduced using a noise estimate which does

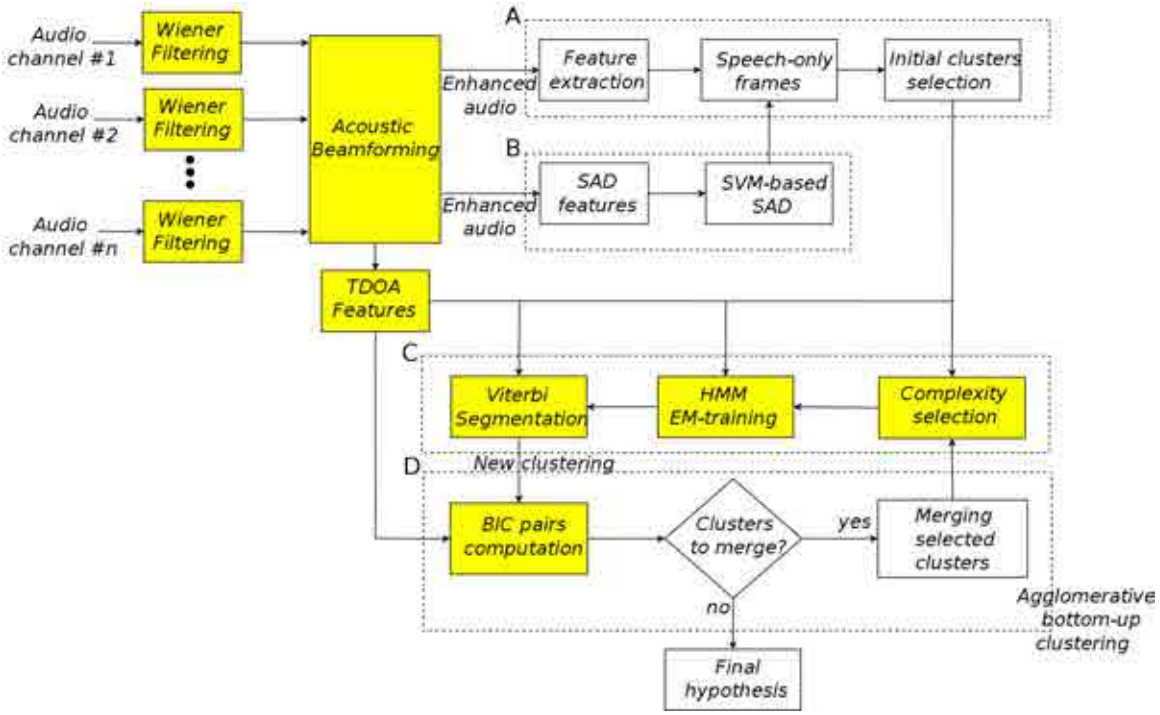


Figure 5.9: *Speaker diarization scheme, multi-microphone approach. In this approach TDOA features extracted from several channels are employed as an independent feature set. Statistical models of the TDOA distribution are employed in HMM training/segmentation as well as in BIC matrix computation.*

not depend on future frames.

5.2.2 Acoustic Beamforming and TDOA estimation

We applied the Weighted-Delay-and-Sum (W-D&S) technique [Flanagan *et al.*, 1985] to perform the signal enhancement as explained in section 2.4.1. It differs from the simpler D&S beamformer in that an independent weight is applied to each of the channels before summing them. We estimate the delay TDOA in order to synchronize two microphone signals for enhancing the signal to noise ratio and to obtain a second stream of information to combine with classical MFCC parameters in the diarization algorithm.
Given the signals captured by N microphones, $x_i[n]$ with $i = 0 \dots N-1$ (where n indicates time steps) if we know their individual relative delays $d(ref, i)$ (Time Delay of Arrival, TDOA) with respect to a common reference microphone x_{ref}, we can obtain the enhanced version of the signal by means equation:

$$y(n) = x_0[n] + \sum_{i=1}^{N-1} W_i x_i[n - d(ref, i)]. \tag{5.20}$$

In order to estimate the TDOA between two segments from two microphones we applied the generalized cross correlation with phase transform (GCC-PHAT) method as defined in section 2.4.1 by equation 2.73. The TDOA for two microphones is computed as in equation 2.74 using a window of 500 ms. at a rate of 250 ms. applied on the Wiener filtered channels. The weighting factor W_i applied to each microphone i and estimated depending on the cross correlation between each channel and the reference channel.
The TDOA information along with the MFCC stream are also combined throughout the diarization process, in the Viterbi decoding as well as in the computation of the BIC values among clusters. Due to the fact of different feature rates for each stream, recall that MFCC features are obtained at a rate of 10ms compared to 250ms in the TDOA case, and aiming to synchronize the TDOA and the MFCC streams, each TDOA value is repeated 25 times to match the rate of the MFCC stream.
In order to obtain a combined TDOA-MFCC score or likelihood we follow the same procedure than in [Pardo *et al.*, 2007], in where the TDOA stream is modeled as a GMM distribution and its log-likelihood is weighted with the MFCC log-likelihood considering the joint log-likelihood for any given set of frames X belonging to a cluster as:

$$\ell(X_{_{mfcc}}, X_{toda}|\Theta_{_{mfcc}}, \Theta_{tdoa}) = \quad \alpha\, \ell(X_{toda}|\Theta_{tdoa}) + (1-\alpha)\, \ell(X_{_{mfcc}}|\Theta_{_{mfcc}}) \tag{5.21}$$

where $\Theta_{_{mfcc}}$, $X_{_{mfcc}}$ are the acoustic model and the set of MFCC frames data respectively, Θ_{tdoa}, X_{tdoa} are the delay model and the TDOA data, and α the weight of each stream. ℓ stands for the log-likelihood, that is, $\log(\Pr(X|\Theta)$. It is worth to mention that, in this formulation, each stream is considered to be statistically independent from each other.

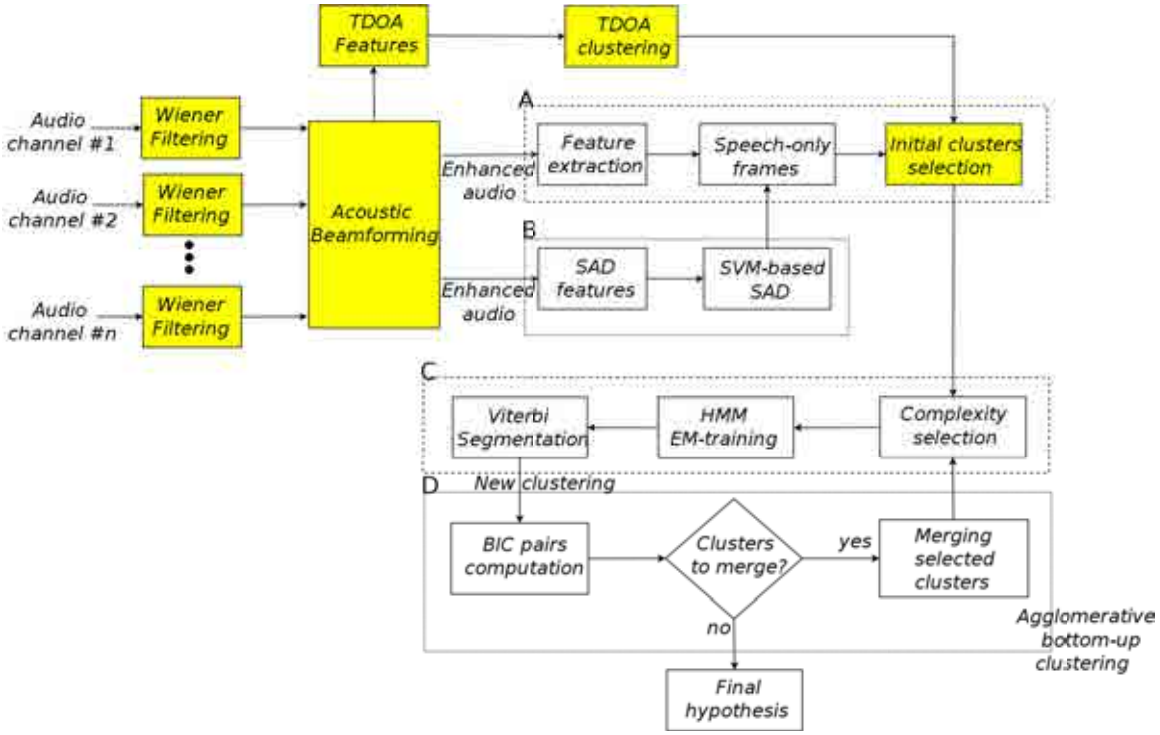


Figure 5.10: *Speaker diarization scheme, multi-microphone approach with cluster initialization based on TDOA features. In this approach TDOA features are employed to obtain a speaker clustering based upon speaker location. The resulting clustering initializes the MFCC-based diarization algorithm.*

5.2.3 TDOA Features for Clustering Initialization

In this section a TDOA-based initialization for the hierarchical diarization system is presented. The novel TDOA clustering proposed is based on the analysis of the temporal statistics of the distance among TDOA pairs. The aim is to obtain speakers clusters from a unique person. Such a cluster purity is based on the hypothesis that if a speech is produced from the same location during a certain period of time, it might come from the same speaker. The resulting TDOA clustering becomes the initial clustering condition for the integrated speaker diarization based on a HMM modeling of cepstral features as can be noticed in the scheme of the figure 5.10.

TDOA-based Clustering

In order to estimate the TDOA between segments corresponding to two microphones, we used a modified version of the Generalized Cross Correlation with phase transform GCCPHAT [Brandstein and Silverman, 1997]. GCCPHAT has been widely used in the acoustic localization task [Segura *et al.*, 2007] and in the blind signal separation field [Swartling and et al., 2006]. Moreover, it is known to perform robustly in reverberant

environments.

TDOA features by themselves permit short-term speaker segmentation by analyzing the delays steadiness. It can be employed for the tracking and the segmentation of individuals in a 2D or 3D metric space by making use of the geometrical information between microphones. Unfortunately, for the NIST database, where microphone locations are not available, a different strategy is needed.

Anyway, the delays between microphone pairs are related to a positions in the 3D space by a non-linear hyperbolic function. Without the knowledge of the microphone geometry, we can only conclude that a displacement of an active speaker in the room yields to different shifts in the estimated delays of each microphone pair. Such relationship with the real 3D geometry depends on the distance between microphones and the relative position of the speaker. Thus, the total dynamic range of the TDOAs of every microphone pair is associated to the maneuvering of the speakers in the room. With this assumption and analyzing the distribution of the TDOA values along time, a set of possible locations can give us the information of the speech events during a recording. The continuity of such values within a segment of time might be associated to the same speaker position. When a speaker change occurs, the TDOA changes and its value is associated to the new speaker location. Moreover, if a set of TDOA is available a smoothing of this strategy can be performed rejecting the wrong estimations of the time delays.

In terms of the number of microphones K in the recording, the total number of possible TDOAs at each time t is given by the combinatorial number $\binom{K}{2} = N$. We define the TDOA space as the set of points $\vec{\tau}_t = (\tau_{1t}, \tau_{2t}, ..., \tau_{Nt})$ where each τ_{it} is the estimated TDOA from a given pair of microphones i in time t. The range of values that can reach each component of $\vec{\tau}_t$ differs since the real distance between microphone pairs might be different. To avoid this problem each component of $\vec{\tau}_t$ is normalized, independently for each component, based on the dynamic range of the TDOA values during an interval of time T as follows:

$$\tau_{it\,norm} = \frac{\tau_{it}}{(\max_{t \in T}(\tau_{it}) - \min_{t \in T}(\tau_{it}))}, \quad i = 1, \ldots, N \tag{5.22}$$

After this normalization, the distance between two points $d(\vec{\tau}_t, \vec{\tau}_{t+1})$ is defined by means the Euclidean distance. Next the tracking and segmentation of moving speakers can be obtained by applying classical spatial data association and clustering techniques [Bar-Shalom and Fortman, 1988].

However, the Euclidean distance between $\vec{\tau}_t$ vectors is highly sensitive due the TDOA estimation errors. Thus, we assume that if a percentage of the TDOAs are not changing during a time segment, the same person is speaking.

With such assumption we propose to compute the distance in the normalized TDOA space as the Euclidean distance in a subset of N with dimension $n < N$. That subset represents the TDOA components of $\vec{\tau}$ which have the lowest variation in each time step. In order to select the n best TDOAs, the difference in two time steps, $\vec{\tau_t} - \vec{\tau}_{t+1}$, is computed. Next, those n closest TDOAs are selected based on a threshold and they are employed to compute the Euclidean distance as,

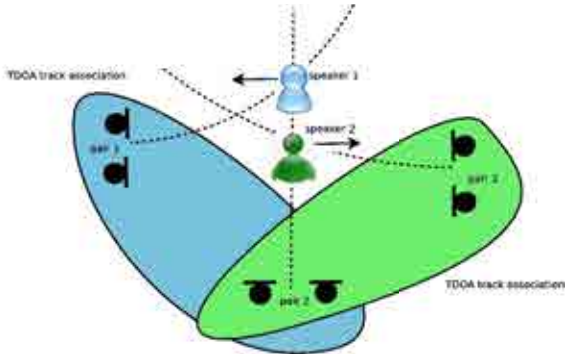


Figure 5.11: *Two speaker example where two tracks are associated to the same observed TDOAs, involving 3 microphone pairs. The arrow indicates the speech source orientation and, for this example, the distance is defined with dimension* $n = 2$

$$d(\vec{\tau}_t, \vec{\tau}_{t+1}) = \frac{1}{n} \sum_{i=1}^{n} \left(\tau_{it} - \tau_{i(t+1)}\right)^2 \tag{5.23}$$

Such TDOAs are more likely to focus on a particular source. Note that this criterion is evaluated in each t so the n selected components of $\vec{\tau}$ could be different at each iteration.
Finally, with the purpose of capturing the temporal evolution of distances $d(\vec{\tau}_t, \vec{\tau}_{t+1})$ and to associate several speakers detections along time, we track the temporal variations of $\vec{\tau}_t$ in the normalized TDOA subset in the following way:

- *Assignation* of new TDOAs to their corresponding active tracks, if they exist. The incoming TDOA estimates are associated to tracks based on the distance.
- *Check* the distance between tracks. Similar tracks are merged into one model. Tracks without an input association during a period of time are marked as non-active.
- *Search* of new tracks from TDOA history based on a distance threshold. Coincidence with non-active tracks are taken into account. The potential tracks are tested with deactivated tracks based on both TDOA distance and an exponential function of the elapsed time since the deactivation.

With this strategy, the non-active tracks are more likely to be merged with the new potential track. That assumption is based on the hypothesis that the tracking of a person that is not speaking may lead to errors since the person might have moved to other position losing the association between the speaker and his observable TDOAs. A time constrain is also imposed on the minimum duration of the tracking, fixing it to 3 seconds. Tracks with less duration are not considered.
As a example, the figures 5.11 and 5.12 show an hypothetical case involving two persons that are speaking simultaneously. We assume that due to the orientation of the speakers, the estimated TDOA at the microphone pair 1 is steered to person 1 and the TDOA at the pair 3 focus to speaker 2, while the pair of microphones 3

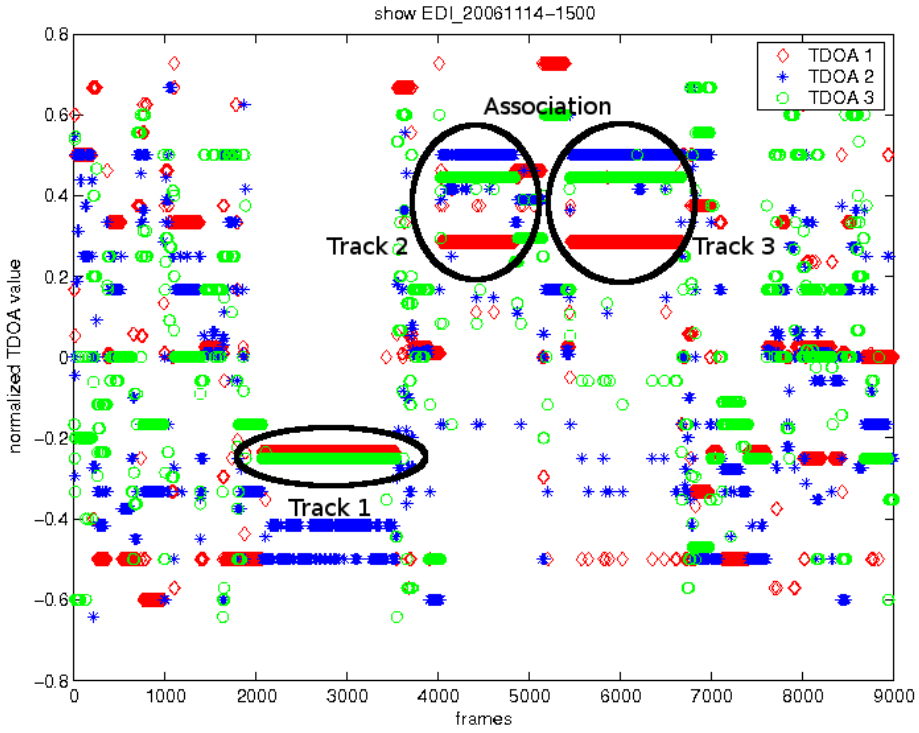


Figure 5.12: *Two speaker example where two tracks are associated to the same observed TDOAs, involving 3 microphone pairs. In the example, the TDOA subsets Track 1 and Track 2 are associated since they present lower variation in such a 3-dimensional TDOA space, compared to Track 3.*

has a TDOA that points to both speakers. In this hypothetical case and using $n = 2$, two overlapped tracks are associated to the speakers and the set of TDOA pairs. So the algorithm also detects overlapped speakers in the recording.

Finally and based on the previous processing of the TDOA values, a speaker clustering is obtained which will be employed as the initial segmentation of the data for the classical agglomerative clustering, see figure 5.10.

5.2.4 TDOA Features for Detection and Handling Speaker Overlap

In this subsection we propose the use of more spatial information in the overlap detection stage previous to the speaker diarization system itself. This strategy is concerned with applying a set of cross-correlation-based spatial features for simultaneous speech detection on distant channel data. Spatial features are extracted from the Generalized Cross-Correlation with Phase Transform weighting (GCC-PHAT) from all channel pairs, and consist of the main peak magnitude of the cross-correlation, the rate of change of the TDOA and a dispersion ratio that measures the energy dispersed in the neighborhood of the main peak in the GCC-PHAT [Zelenák *et al.*, 2010]. These spatial features pose the problem that the dimensionality of spatial feature space is dependent on the number of microphone channels available and can be eventually very high or can vary across different

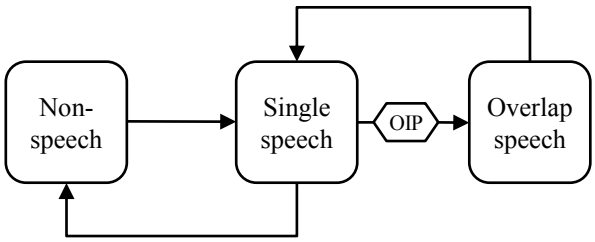


Figure 5.13: *Word network topology in decoding process.*

rooms. In order to deal with these issues, we suggest a reduction and normalization of the size of spatial feature vectors by using principal component analysis (PCA). Afterwards, spatial information is integrated into a spectral-based overlap detection system. The PCA-based microphone-pair fusion is additionally compared to an alternative strategy involving a multilayer perceptron (MLP) neural network, whose output classification score is used as an extra spatial feature.

Our aim is to improve the baseline diarization system by handling detected simultaneous-speech segments, as in [Boakye *et al.*, 2008d],[Boakye *et al.*, 2008b]. Two techniques are considered to accomplish this purpose. In the former approach, also referred to as overlap exclusion, overlaps are discarded from training, hoping to achieve purer cluster models and thus a more precise segmentation. The latter technique allows to assign two speaker labels in segments with simultaneous speech. In the latter case, the overlap hypothesis needs to be sufficiently precise, since all of the falsely detected overlaps will contribute to diarization error and only a perfect selection of speaker labels would recover the missed overlapping speaker time.

Baseline Overlapped Speaker Detection

The baseline overlap detection utilizes a number of spectral-based features. Cepstrum is successfully applied in various speech-related tasks and forms a good basis for a feature set. For that reason, 12 mel frequency cepstral coefficients (MFCCs) were extracted every 10 ms over a window of 30 ms.

Another spectral-based feature is the spectral flatness (SF), which was extracted over 30 ms windows. Spectral flatness was applied for discrimination between speech and non-speech [R. Yantorno, 2001], but can eventually convey information about the number of speakers speaking [Boakye *et al.*, 2008d]. It is defined as the ratio between geometric and the arithmetic mean of N spectral magnitudes ($N = 100$ in our case)

$$M_{\mathrm{SF}} = 10 \log_{10} \frac{\sqrt[N]{\prod_{i=0}^{N-1} \mathrm{mag}(i)}}{\sum_{i=0}^{N-1} \mathrm{mag}(i)}. \tag{5.24}$$

Linear predictive coding (LPC) analyzes the speech signal by estimating the formants of a speaker. It is assumed that LPC of a reasonably chosen order can model the spectrum of a single speaker quite well, but will fail for a region with multiple speakers [Sundaram *et al.*, 2003; Boakye *et al.*, 2008b]. Consequently more energy is left in the residual signal (prediction error) in the later case. In our system, residual energy of

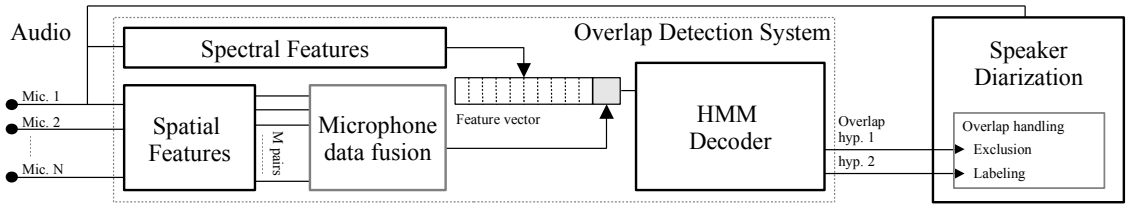


Figure 5.14: *Overlap detection system diagram.*

a 12th-order LPC (LPCRE) was computed over a 25 ms window. All features vectors were mean-variance normalized according to statistics of the training data and first order differences were added totalling 28 features with 10 ms frame rate.

The system considers three acoustic classes representing non-speech, single-speaker speech and overlapping speech. For each class an HMM is defined. For a more accurate modeling of transitions between classes the HMM has three states, which also works as a minimum duration constraint. Every state is modeled with a GMM using diagonal covariance. Since the amount of training data is not balanced among classes, we use 256 Gaussian components for single-speaker speech and 64 components for overlapping speech and non-speech. GMMs are created by iterative Gaussian-splitting technique and subsequent re-estimation.

Detection hypothesis is obtained by Viterbi (maximum-likelihood) decoding and applying a word network whose topology is depicted in the figure 5.13. It is worth to mention that the transition probabilities between different HMMs states are not trained. They are set manually. In order to increase the precision, the transition from single-speaker speech to overlapping speech might be penalized with an overlap insertion penalty (OIP), e.g. imposing that certain transitions are completely forbidden. The detection hypothesis are then fed to the speaker diarization system as shown in the figure 5.14.

Spatial Features for Overlap Detection

The cross-correlation function is well-known as a measure of the similarity between signals for any given time displacement and ideally its maximum lies in correspondence to the delay between the pair of signals [Svaizer and others, 1997]. A commonly used technique to estimate the time delay between two acoustic signals that performs robustly in reverberant environments is the Generalized Cross-Correlation with Phase Transform weighting (GCC-PHAT) [Brandstein and Silverman, 1997; T. Gustafsson and B. Rao and M. Trivedi, 2003]. Although it is a general purpose technique and not fully adapted to speech, it has turned out to be the most successful state-of-the-art approach to speaker localization and it has been employed by some researchers in the field of speaker diarization, including [Luque *et al.*, 2008b; Araki *et al.*, 2008]. See chapter 2, section 2.4.1. The GCC-PHAT function exhibits a prominent peak at the elapsed time corresponding to the dominant sound source in the room, minimizing the peaks of the non-dominant sources and reverberation at the same time, see figure 5.15. The value of the GCC-PHAT peak provides a measure of the coherence between signals independently of the microphone gains or the signal power, and varies with the distance between microphones,

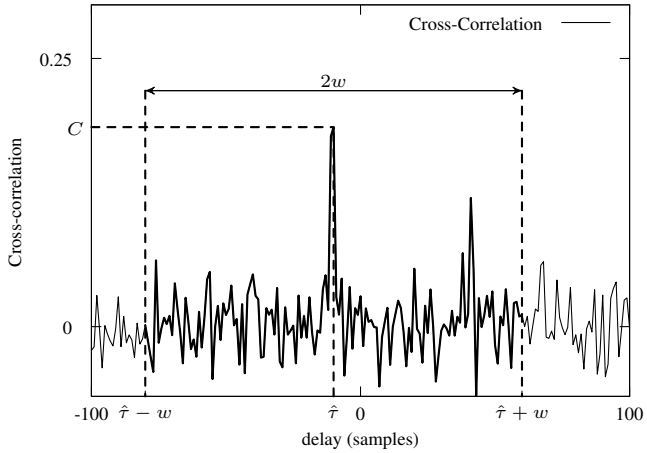


Figure 5.15: *Example of the cross-correlation between a pair of microphones involving two concurrent speakers. The value of the main peak is the confidence feature C, and its time displacement $\hat{\tau}$ corresponds to the TDOA. The ratio between C^2 and the quadratic sum of the values in bold under the window is used as the dispersion feature D.*

the distance between the acoustic source and the microphone pair, and with the environmental noise and reverberation conditions.

In situations dealing with multiple, possibly moving, concurrent speakers, we have observed that the time delay estimates produced by the GCC-PHAT jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. The maximum value of the cross-correlation sequence is also lower than in the single speaker situation, since multiple speakers introduce random peaks, which attenuate the main peak. Based on these observations we are proposing several cross-correlation-based spatial features for every microphone pair that provide some degree of information on speaker overlaps.

An easily observable feature is the *coherence value*, defined in equation 5.25. This is the value of principal peak of the GCC, and in ideal conditions should be high for single-source situations, while the presence of noise, reverberation and concurrent acoustic sources attenuate this value.

$$C_{mn} = \max(R_{mn}(\tau)) \tag{5.25}$$

Derived from the coherence value, we are also proposing to extract the coherence *dispersion ratio*, as follows,

$$D_{mn} = \frac{C_{mn}^2}{\sum_{t=-w_{mn}}^{w_{mn}} R_{mn}^2(t+\hat{\tau}_{mn})}. \tag{5.26}$$

This value is computed as the ratio between the square of the main peak value and the square quadratic sum of the cross-correlation values under a time delay window w_{mn}. The size of the window w_{mn} varies for different microphone pairs and it is set to the TDOA standard deviation of each pair. In this way, the dispersion ratio measures the relation between the energy of the main peak and the energy that is scattered in its neighborhood. Similar to the coherence feature C_{mn}, the dispersion ratio is close to 1 in the case of a single speaker and ideal

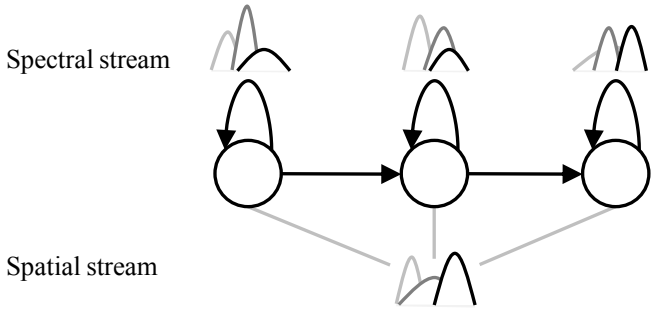


Figure 5.16: *HMM diagram with two feature streams.*

conditions, while it has a lower value in reverberant conditions or concurrent acoustic sources situations. Finally, the *delta of TDOA* obtained by equation 2.74 for every microphone pair also carries information on overlaps. The derivative of the TDOA is high in situations where the speaker is moving, multiple non-concurrent speakers change turns at talk or multiple speakers talk simultaneously. An illustration of the cross-correlation between a pair of microphones and the proposed spatial features can be seen in the figure 5.15.

One of the main problems that arise is the high dimensionality of the spatial feature vectors. A recording involving 12 microphones yields to 66 pairs and 198 features. Also the number of microphones differs from site to site, making it difficult to train a general model. Other issue is that the proposed spatial features are, in general, not commensurable across different microphone pairs, since they are intrinsically tied to physical characteristics of the pair like the inter-microphone distance.

Our first strategy for dimensionality reduction and normalization is the application of a sequential PCA [Ross *et al.*, 2008], originally introduced in [Levy and Lindenbaum, 2000], which transforms the original feature space into a new coordinate system with the greatest variance lying on the first component. PCA was used for similar issues in diarization in [Otterson, 2007]. We estimated a separate transformation matrix for every discussed spatial feature kind per each site and then we use just the first principal component. Hence, in the given example with 12 microphones, we would end up with one transformed coherence, one dispersion and one TDOA delta, which are finally added to spectral vector.

We also considered an alternative approach to reduce the spatial vector dimensionality based on a neural network with a four-layer perceptron. The input of the MLP is composed by 6 input neurons, 3 for spatial features and 3 for normalization values (*mean of coherence*, *variance of coherence*, *variance of TDOA*) for every pair. The output is a binary score classifying between overlap and non-overlap, which is commensurable across microphone pairs. For a given frame the average score was taken and merged with corresponding spectral feature vectors. Spatial information is modeled in the overlap detection HMM-based system with a separate Gaussian mixture as the spectral features (see figure 5.16). Furthermore, the spatial GMM shares means and variances across the three states.

5.3 Other Diarization Approaches: Spectral Clustering

Agglomerative hierarchical clustering (AHC) has become one of the most widely applied approach to speaker diarization task. Clusters are represented by parametric probability densities like Gaussian mixture models (GMMs). Hidden Markov Models (HMM) together with Viterbi perform segmentation and clustering of audio data in an iterative bottom-up fashion [Ajmera and Wooters, 2003]. In such a framework, Bayesian information criterion (BIC) is one of the most popular metrics to estimate which couple of clusters merge at each agglomerative iteration. BIC is usually also employed as a stopping criterion for the agglomerative process [Anguera *et al.*, 2011]. Metrics like as Generalized likelihood ratio (GLR), Kullback-Leibler (KL) divergence, information change rate (ICR), amongst others, has been also proposed, but all of them with same Achilles' heel, that is, a high computational cost and a performance heavily depending on the choice of the metric [Han *et al.*, 2008].

To overcome this drawback, we propose a speaker diarization approach method based on spectral clustering (SC) avoiding the use of computationally demanding statistical metrics like BIC. Spectral clustering refers to a class of techniques which rely on the eigenstructure of a similarity matrix to partition points into disjoint clusters. Points having high similarity are pooled together in the same cluster whereas they evidence a low similarity among other points grouped in different clusters. SC has been successfully applied in blind source separation, separating speech mixtures from a single microphone [Keshet and Bengio, 2008] with no requirement of explicit models for speakers. However, there are a few recent works which use SC to infer speaker clusters specifically in speaker diarization task [Ellis and Liu, 2004; Ning *et al.*, 2006; Ning *et al.*, 2010; Iso, 2010].

Instead of making assumptions on data distribution, SC relies on analyzing the eigenstructure of an affinity matrix [Keshet and Bengio, 2008; Luxburg *et al.*, 2007] which models the similarity among the clusters. Nevertheless, in contrast to classical AHC clustering approaches, such affinity matrix is treated as part of the learning problem. Our proposal is based on a parametric segment representation through a Gaussian super vector (GSV). The GSV vector is composed by stacking just the means of the Gaussians [Campbell *et al.*, 2006b]. The classical BIC metric in AHC is replaced by Ng-Jordan-Weiss (NJW) spectral clustering algorithm [Keshet and Bengio, 2008]. In our work, the affinity matrix is built by defining the similarity between segments through the Euclidean distance in the GSV space of segments representation. We employ spectral clustering algorithm with cluster number estimation based on eigenstructure analysis, searching the drop in the magnitude of the eigenvalues as in [Ning *et al.*, 2006; Iso, 2010].

Our clustering algorithm still depends on HMM/GMM modeling and Viterbi segmentation as pre and post - processing for spectral clustering. For instance, they are used for obtaining the GSV vector representation per each segment which feed the SC algorithm. In that case, the initial segmentation is computed through a initial partition in homogeneous segments. Such segments are realigned by an HMM/GMM model together with Viterbi decoding up to no variation in segmentation structure is noticed. Finally, it is also applied as a post-processing of spectral clustering results. This approach generates results comparable to AHC+BIC ones but achieves much higher speed than the latter.

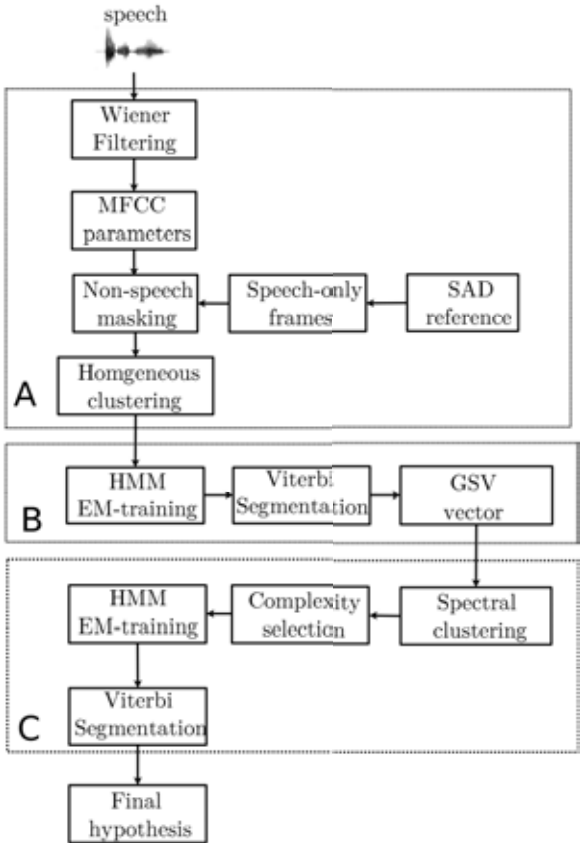


Figure 5.17: *Speaker diarization scheme based on spectral clustering with Viterbi HMM/GMM initialization and clustering refinement.*

Despite of the good results achieved by popular AHC systems, an important drawback arises in the case of long duration audio documents. AHC approach is a highly time consuming approach. The processing time for audio recordings depends directly on the number of initial segments taken into account. For instance, augmenting the initial number of segments in long audio documents considerably increases the size of the BIC comparison matrix and, therefore, the total time processing of the iterative approach. Reducing the number of initial segments drastically makes smaller such time but at the expense of the speaker detection accuracy due to the initial cluster impurity. So there exists a trade-off between computational cost and detection performance in AHC based systems. To overcome such drawback we propose a clustering approach based on spectral clustering that, despite of its computing time is still dependent on the number of initial segments, it avoids statistical metrics to build the similarity matrix yielding to a faster algorithm than AHC+BIC one.

In the figure 5.17 we draw the scheme of the proposed system based on spectral clustering. As in the AHC approach, we keep as prior modules, the oracle Speech/Non-Speech detection module and a Wiener filtering implementation from the QIO front-end. Cluster initialization is still based on an homogeneous splitting of

data but, in contrast to AHC approach, no automatic selection of number of clusters is performed. Number of initial cluster is tuned with development data.

5.3.1 Segments Representation

The core of the proposed system is shown in blocks B and C in the figure 5.17. Before spectral clustering was carried out, initial segments are modeled by a mixture of Gaussians with fixed complexity, that is, number of Gaussians is independent of the duration of the segment. Following, a Viterbi decoding is performed by means an ergodic HMM. Once initial segmentation is stabilized, the segments presents a great variety of durations. To overcome this drawback, a Gaussian super vector (GSV) modeling is proposed [Campbell *et al.*, 2006b]. Furthermore, segments lesser than 3 seconds are discarded in order to ensure statistical significance in Gaussian parameter estimation. Such segment discarding is motivated by characteristics of our data. The estimated probability density for a speech segment is assumed to represent speaker characteristics. However for conversational speech recordings, plenty of short utterances and changes in speaker turns, the density estimation by means GMM will be strongly biased by their phonemic variations. In any case, initial discarded segments will be assigned to discovered clusters by the SC through Viterbi alignment in last step of the approach, see block C in the figure 5.17.
Only the means of the Gaussians μ_{ik} are stacked in a vector to build the GSV. The μ_{ik} means are normalized through the corresponding variance σ_{ik} and weight of the Gaussian as follows:

$$GSV_{ik} = \sqrt{w_{ik}}\ \Sigma_{ik}^{(-1/2)}\mu_{ik}, \qquad (5.27)$$
$$k = 1, \ldots, D\ ,\ i = 1, \ldots, M$$

where w stands for the weight of the Gaussian, Σ is the corresponding variance and μ represents the mean of the Gaussian. Indexes i and k stand for the number of Gaussian in the mixture model and the Gaussian dimension respectively. Therefore, stacking normalized Gaussians' means in a vector leads to a length of the GSV vector equals to the number of Gaussian M employed to model i-th segment (which is always the same for all segments) times the number of dimensions D.
Other segment representation has been proposed for spectral clustering in diarization task. In [Ning *et al.*, 2006] GMM parameters adapted from a UBM, trained on the whole audio data, are employed as representation for speech segments whereas KL distance is used for building the affinity matrix. In [Iso, 2010], author employed a non-parametric representation of speech segments based on Vector Quantization (VQ) in which the VQ codebook is created from the audio recording and utterances are represented as a vector of frequencies in VQ space. The affinity matrix is constructed by means cosine similarity distance.
In our approach we have decided to apply Gaussian super vector model due his excellent results in speaker verification tasks and its robustness against trials involving segments of different duration [Campbell *et al.*, 2006b]. In addition, no statistical measure as KL is proposed to construct the affinity matrix but Euclidean

distance is computed in GSV space, consequently saving in computational time.

5.3.2 Spectral Clustering

Once a initial segmentation and a segment representation is computed, a speaker clustering is performed to join those segments which belong to same speaker. We use a modification of the Ng-Jordan-Weiss (NJW) algorithm [Ng *et al.*, 2001] and a modified implementation in C++ programming language taken from [Chen *et al.*, 2011], which we first briefly review. Given a set of speech segments $S = \{s_1, \ldots, s_n\}$ represented by n points $X = \{x_1, \ldots, x_D\}$, in this work the GSV vector, that we want to cluster into k subsets:

- Form a similarity graph defined by the affinity matrix $A \in \mathbb{R}^{n \times n}$ where $A_{ij} = \exp(\frac{-d^2(s_i, s_j)}{\sigma^2}) \quad \text{if } i = j, \text{ and } A_{ii} = 0$, where $d(s_i, s_j)$ is distance function and σ^2 is a scaling parameter.

- Define D to be the diagonal matrix whose (i, i)-element is the sum of A's i-th row, and construct the normalized symmetric graph Laplacian matrix $L = D^{-1/2} A D^{-1/2}$.

- Select the number of clusters k.

- Find $\{u_1, u_2, \ldots, u_k\}$, the k largest eigenvectors of L, and form the matrix $U = \{u_1, u_2, \ldots, u_k\} \in \mathbb{R}^{n \times k}$.

- Re-normalize the rows of U to have unit length yielding $Y \in \mathbb{R}^{n \times k}$, such that $Y_{ij} = U_{ij} / (\sum_j U_{ij}^2)^{1/2}$.

- Cluster the points Y_{ij} with k-means algorithm into clusters $C_1, \ldots, C_k$.

The main idea behind spectral clustering algorithm relies on changing the representation of data points x_i in $y_i \in \mathbb{R}^k$, that is, mapping x_i into a space where the simple k-means clustering algorithm has no difficulty to detect clusters. Nevertheless, such a situation only occurs in an ideal case whether data is enough clean and consequently no overlap among different classes takes place.
In order to form the affinity matrix, it is required to define a similarity function d on the data and a scaling parameter σ. In this work, the Euclidean distance among GSV vectors has been employed, fulfilling distance requirements such as: be non-negative, be low for similar segments and high otherwise. Euclidean distance has clearly an intuitive sense in GSV space, giving an idea of how far are Gaussian mixtures among different segments. In addition, all distances amongst segment-pairs has been considered leading to a fully connected graph. The scaling parameter σ is some kind of measure of when two points should be considered similar and controls how rapidly the affinity matrix A_{ij} falls off with the distance between s_i and s_j segments. As the work presented in [Ning *et al.*, 2006], we calculate a scaling parameter depending on the pair of segments (s_i, s_j) involved in distance computation, by considering the second order statistics of distances to all other data segments as follows,

$$\sigma_{ij} = \sqrt{Var(d(s_i, s_n))Var(d(s_j, s_m))}, \quad \text{with } n \neq i,\ m \neq j \tag{5.28}$$

where $Var(\cdot)$ computes the variance and $d(s_i, s_n)$ are distances from segment s_i to all other segments. In contrast to [Ning *et al.*, 2006] we do not include the scalar parameter β in computation of σ_{ij}.
As part of the diarization task, the number of clusters has to be estimated automatically. In model-based clustering approaches, such decision is usually based on the likelihood performed from data as in the previous AHC system. In this work, number of clusters is estimated by analyzing the magnitude of the eigenvalues of the normalized Laplacian matrix L as in [Ning *et al.*, 2006; Iso, 2010]. It is known as eigen gap heuristic, where the objective is to select k clusters as the number of k maximum eigenvalues of the Laplacian L matrix,

$$\gamma_k = |\lambda_k - \lambda_{k+1}| > \Theta, \tag{5.29}$$

where γ_k is the eigen gap between two consecutive eigenvalues $\{\lambda_k, \lambda_{k+1}\}$ and Θ is a threshold we tune with development data. There exists different explanations to the use of such criterion, as those from perturbation theory or geometric graph invariants, due to the fact that similarity information can be compacted with just first eigenvalues/eigenvectors of the Laplacian matrix L [Keshet and Bengio, 2008; Luxburg *et al.*, 2007].
Finally, in the last step of the SC approach and once we have selected the number of k clusters, a k-means algorithm is employed to link up segments in clusters into the new space representation, $y_i \in \mathbb{R}^k$

Clustering Refinement

As we can see in block C in the figure 5.17, the resulting clustering obtained by SC feeds a last HMM alignment step. In contrast to the initialization step, a complexity selection as in AHC system is employed, and the newly clusters are modeled by an HMM/GMM. Several Viterbi alignments are performed until no variation in the segmentation is perceived and a final clustering hypothesis is obtained.

5.4 Experiments

This section verses about the experimentation of the different proposed techniques in order to evaluate its suitability in the task of speaker diarization for meetings. It is done by first defining a baseline system to compare all algorithms to. Such baseline system is derived from the broadcast news mono-channel system with several improvements that were considered standard and necessary to the system as adapted to meetings. Then a set of metrics used in the evaluation of the different techniques are described in detail. Next, the databases that are used to compare the algorithms performance with that of the baseline and the reference segmentations which are used in the experiments are explained and reasoned. Finally, the different experiments with the proposed algorithms are performed and results are explained.

5.4.1 Meetings Domain Experiments Setup and Corpora

As has been reported in NIST Rich Transcription evaluations and for researches in speaker diarization, there are two phenomena that were common to all diarization systems. These are the big variance of the scores among all evaluated shows and the extreme susceptibility of the scores to experience big changes upon small modifications of their tuning parameters. Such a evidence forged a common term inside the speaker diarization community: the ”flakiness“, a term started being used for speaker diarization during the RT04f workshop. The DER results depend on many factors and, as it is reported in [Mirghafori and Wooters, 2006] where some of these factors are studied, they refer to the high variability of the DER values as show flakiness.
The ”flakiness” corrupts the conclusions obtained when comparing the performances of several algorithms to a baseline and it mights yield to an incorrect selection of the optimum baseline parameters and test conditions. In many cases, due to flakiness, testing the same algorithms with two different databases or baseline systems derives into two very different results, one proving the validity of the proposed algorithm and one otherwise. In order to run meaningful and fair experiments using the algorithms proposed in this thesis one needs to carefully define:

- A baseline system, which acts as the comparison ground to all systems proposed and tested.
- A common development and test datasets, based on the NIST RT evaluations datasets, in order for results to be comparable between experiments and to systems outside of the thesis.
- A set of metrics in order to evaluate such systems with commonly used and available techniques.

In the following subsections each of these items is described as it has been used in this thesis for most of the experiments with the main blocks of the system. Taking as a reference the block diagram in the figure 5.1, experiments were conducted on three of the main blocks, namely the single channel diarization, the speech/non-speech module and the multiple channel speaker diarization module. For each block a baseline was defined to suit its characteristics and to allow for the development of its optimum parameters selection. The initial Wiener filtering of the signal was not analyzed as it was used without modification from its original implementation outside of the scope of this thesis.
The single channel diarization depicted in figure 5.1 is considered as the baseline system. It is employed for comparison in SDM experiments. It is worth to mention that the complex selection algorithm is not used as baseline. The speech parametrization is conformed of 16 static MFCC parameters without energy and without deltas. The number initial of cluster is set to 65 at each show and the number of Gaussians per cluster is fixed to 5.
During the agglomerative clustering processing the same speaker turn minimum duration is applied, that is 2.5 seconds. Before the output of the resulting segmentation, a final segmentation step is performed using the same speaker models but reducing the minimum duration to 1.5 seconds to allow for smaller speaker turns to be properly detected. The merging criterion is driven by the BIC value among cluster and just the highest BIC pair-value selects the couple of clusters to merge. In the same way, the stopping criterion is reached based on the BIC value as explained in previous sections, see section 5.1.4.

Dataset	# excerpts	Ave. duration in seconds	Meeting Sources						
			CMU	ICSI	TNO	NIST	AMI/EDI	VT	IDI
RT05s	10	722.21	2	2		2	2	2	
RT06s	9	1080.43	2		1	2	2	2	
RT07s	8	1352.19	2			2	2	2	
RT09s	7	1551.21				3	2		2

Table 5.1: *Summary of datasets used in the experiments.*

The fast logarithm implementation, see section 5.1.6, was employed in the experiments in order to speed-up them. The Q parameter was set to 12.
The baseline system used for experiments in the multichannel-approach is mainly the system submitted to RT09s NIST evaluation campaign. This contains all the modules explained for the baseline SDM system, those explained in the MDM section 5.2 and their parameters were optimized using a subset of 17 meetings from the development data available for RT06s and RT07s. Differences between official results [Fiscus and et al., 2009b] and those reported in this PhD thesis proposal are due to a "bug" in the processing of the audio data during RT09s evaluation participation.

Corpora

In the experiments in this thesis the datasets used were obtained from the data available for the Rich Transcription (RT) evaluations for meeting domain. This databases were delivered by the National Institute of Standards and Technology to participants in RT evaluations. During the development of this thesis the author participated in the later two evaluations, RT07s and RT09s whereas datasets from previous evaluations was
So far the evaluations on meetings have been RT02, RT04s, RT05s, RT06s, RT07s and RT09s. On the later four editions only the conference room type data has been used as it contains a richer variety of speakers and with characteristics matching more closely the aim of the algorithms presented in the thesis. From all available datasets, two groups have been defined as development and test. The RT05s, RT06s and RT07s sets form the development set, with a total of 26 meeting excerpts, ranging from 10 to 12 minutes in duration each. The RT05s data was only used as training material for the SVM-based speech detector. In addition, the meeting evaluation data from CHIL 2006 database and the Spanish Speecon database [2] was also employed for speeh/non-speech model estimation. The datasets RT06s and RT07s (with 17 meetings) was used for algorithm development and tuning of diarization parameters. Finally, the RT09s dataset has been used as a test set (with 7 meetings), to compare the system improvements on data not used to tune its parameters, "unseen" data. Table 5.1 summarizes the data available in each one of the RT sets used. For a complete list of the individual files refer to Appendix C.

[2] http://www.elda.org/catalogue/en/speech/S0160.html

SAD SVM-based						
RT'05 sdm	RT'06 sdm	RT'07 sdm	RT'09 sdm	RT'07 mdm-softsad	RT'07 mdm-hardsad	RT'09 mdm
8.03 %	4.88 %	7.03 %	6.02 %	5.39 %	4.72 %	5.9 %

Table 5.2: SAD error official results in previous RT Evaluation, conference data condition.

5.4.2 Speech/Non-Speech Detection based on SVM Classifier

Experiments for the speech/non-speech module were obtained for the SDM case to make it directly comparable with the baseline system results reported in the previous section. The training set for model estimation was the RT05s dataset + CHIL06 development meeting set [3] + Spanish Speecon database. The development set consisted on the RT06s + RT07s datasets (17 meeting excerpts) and the test is the RT09s set. In the development of the SVM-based speech/non-speech detector there is a main parameters that need to be set, that is the bias b of the separating hyperplane, see section 2.2.2, which allows to adjust the speech detection in the testing stage without modification of the trained SVM models.

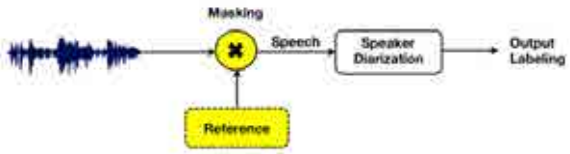
This section summarizes the **official** results in RT evaluations of the SVM-based speech detector and some post experiments aiming to study the SAD performance impact on the global diarization error of the agglomerative clustering. Mainly, we focus on two open issues: Examining the trade-off between misses and insertions introduced by the SAD system and the difference between a SAD module as a pre-processing step to the agglomerative against a SAD as a post-processing of the agglomerative output. Furthermore, we examine the results not only in the sdm condition but also in the multi channel approach.

The table 5.2 reports the official performance results of the SAD module in the different Rich Transcription evaluations conducted in the conference room environment. The Rich Transcription evaluation was composed by an specific SAD evaluation until the RT06s evaluation. Therefore, the results reported in the table 5.2 refers to those results performed in RT05s and RT06s specific SAD evaluations, respectively, and corresponds to the participation of a UPC team previous to the work presented in this PhD thesis. For the rest of the evaluations, RT07s and RT09s in where no specific SAD evaluation was conducted during Rich Transcription, the results reported correspond to the diarization system output scored considering any speaker labeling in the NIST official references as speech.

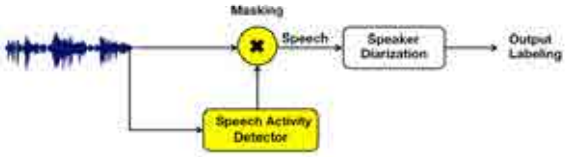
Pre-processing SAD Versus Post-processing SAD

Speech/non-speech activity detection is a key component in any diarization system which handles data in a real multi-party environment. The SAD module gives to the diarization system those frames which are considered speech and therefore the data to be clustered. This is the case of our baseline approach, in where the SAD is considered as a pre-processing step to the agglomerative clustering. In such a situation, the errors introduced by the SAD module can not be recovered in following processing since no strategy has been implemented

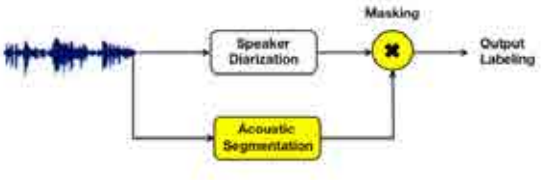
[3] It corresponds to RT06s lecture data set



(a) Oracle SAD with NIST references



(b) Classical pre-processing SAD



(c) Post-processing SAD

Figure 5.18: *Three main schemes about SAD are studied. (a) An oracle SAD, provided by the NIST official references, serves us to study the effect of non-perfect SAD references and gives us an idea of the room for improvement. (b) The classical pre-processing SAD which discards non-speech frames before feed with them the diarization system. (c) A post-processing approach which masks the output of the diarization engine.*

System	MISS (%)	FA (%)	SER (%)	DER (%)
Oracle SAD (NIST reference)	3.7	0	6.5	10.27
SAD as pre-processing	5.7	2.3	4.6	12.61
SAD as post-processing	5.7	2.3	4.5	12.41

Table 5.3: *DER error rates obtained in RT07s dataset. Results reported at the RT09s workshop [Fiscus and et al., 2009b].*

to deal with this issue. Therefore the miss speech, that is, the speech not detected by the SAD module is no further processed by the AHC system and, as a consequence, it sums directly to the final DER error, see section 2.3.3. For the false alarms errors, those non-speech regions detected as speech by the SAD, the situation is identical as for misses but, in addition, the fact that the AHC has to deal with false speech degrades the purity of the speakers clusters and it likely leads to increasing diarization errors.

Three main schemes around the SAD scheme are studied in this section, focused on assess the impact of a pre-processing SAD compared to a post-processing SAD. The figures 5.18 stand for these two strategies. First of all, an oracle SAD is taken into account by using the official references provided by NIST. It counts for the effect of non-perfect SAD references, giving us an idea of the possible room for improvement. The second scheme, depicted in the figure 5.18 (b) shows the classical pre-processing SAD which discards non-speech frames before they are further processed by the diarization system. This is the most employed strategy by

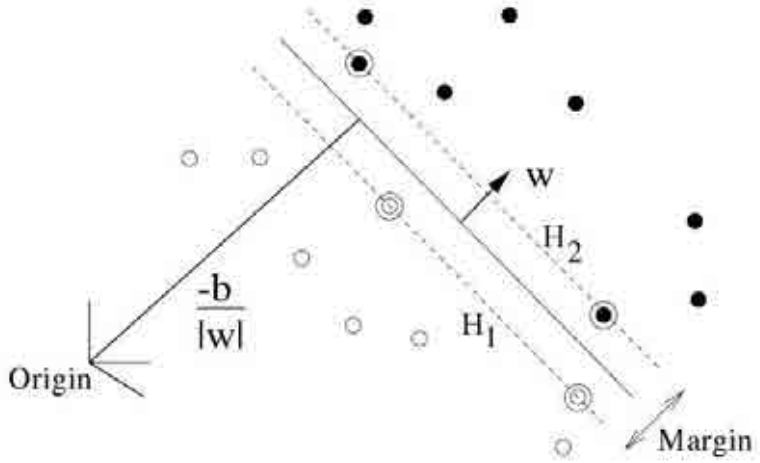


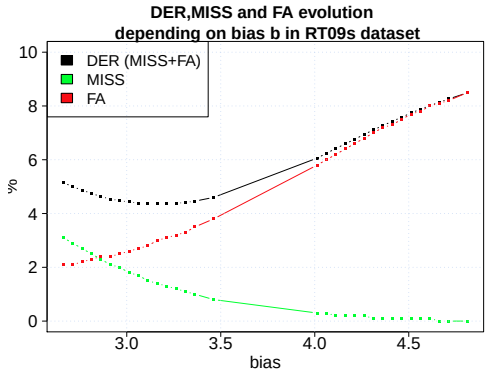
Figure 5.19: *One a SVM model for speech is trained, in this case using proximal SVM technique see section 2.2.2, the corresponding separation hyperplane between classes can be modified by introducing a bias* b*. Geometrically, it can be interpreted as a displacement of the hyperplane* H_0 *towards one of the classes, that is, favoring the detection of that class.*

most of the diarization systems reported in the literature. Finally, in figure 5.18 (c) a post-processing approach is proposed in where the diarization output obtained by clustering the whole data (speech and non-speech frames) is masked by the SAD output labeling.

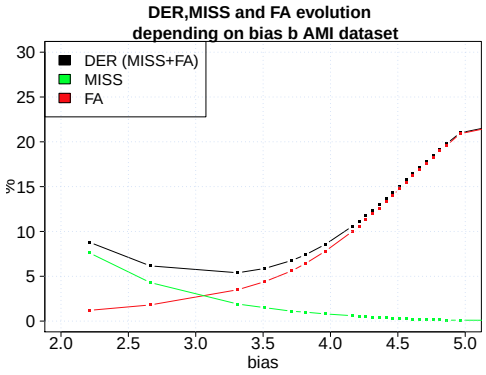
The table 5.3 reports the diarization errors rates obtained in RT07s dataset. These are official results reported at the RT09s workshop [Fiscus and et al., 2009b]. The diarization scheme is comparable to the mdm version depicted in figure 5.9, for further details see the workshop presentation in [Fiscus and et al., 2009b]. Both pre and post processing SADs just differ in the way how they combine with the agglomerative clustering. The SAD was trained with the corpus data explained in the introduction of this section and developed using RT06s and RT07s datasets in order to choose the best bias parameter b, see following section, for SVM modeling w.r.t DER obtained in RT07s dataset.

Two main conclusions can be extracted from the previous table 5.3. First, the DER obtained by the system with oracle SAD references give us an inferior threshold for the improvement of the global DER of the system. It can be noticed that up to a 2.5% absolute improvement can be reached with a SAD perfectly adapted to the NIST references. It worth to mention that the miss speech in this case is far from be 0.0%. It is due to the fact that overlap speech is taken into account in the computation of the DER and therefore all overlap speech is considered as miss speech since the SAD references does not provide information about more than one speaker at the same time.

A second impression reflects that there is no a statistical difference between discard non-speech frames before the clustering or just masking the clustering output with the non-speech frames detected by the SAD. Only a 0.20 absolute improvement is reported by the post-processing strategy that seems insufficient to decide which



(a) RT09 dataset



(c) AMI dataset

Figure 5.20: *Effect on misses and false alarms by shifting the corresponding separation hyperplane between classes through the bias* b. ***(a)*** *Results in RT09s dataset and in* ***(b)*** *AMI dataset by using the same SAD trained with Speecon, CHIL06 evaluation dataset, RT06s and RT07s datasets.*

of the two approaches is better. The RT07s dataset is only composed of 8 excerpts and a great number of shows should be necessary to elucidate what is the better strategy.

Speech Misses versus Speech Insertions

The bias b of the separating hyperplane, see section 2.2.2, is chosen according to the DER performance obtained in the RT06s and RT07s datasets. The trade-off between miss speech and false alarms produced in the results of the development data can be observed in the figure 5.20 and the table 5.4. The bias b acts as a weighting parameter which controls the decision boundary between the speech and non-speech classes. Thus, by adjusting the bias of the hyperplane benefits the detection of one class with respect the other one and, as can be seen in the table 5.4, it has a noticeable impact on the total DER error of the diarization process itself. It is worth to mention that such a parameter tuning can be done after model estimation which means that no new model computation is required, see figure 5.19.

The effects of varying the bias b is visible in the trade-off between misses and false alarms. The figure 5.20 clearly illustrates this impact which also can be noticed in the two first columns in the table 5.4. The sum of misses and false alarms corresponds to the third column. The impact on the speaker error is also reported in the fourth column and the total sum of errors (DER) in the last of them. The speaker error seems to follow a trend in which false alarms errors has a higher impact on the DER to those errors caused by misses. If we focus on the results with a sum of errors belonging to the range $[8.0, 8.2]$ it can be observed that the DER oscillates inside the interval $[6.7, 4.6]$, that is, the standard deviation for the DER is higher compared to that obtained from the sum of misses and false alarms. The minimum speaker error 4.6 is reached at the operation point $(MISS, FA) = (5.7, 2.3)$ which correspond also to the minimum DER error, 12.61. Summarizing, such a results suggest a higher sensibility of the DER error against the false alarms, yielding to the conclusion

MISS (%)	FA (%)	MISS+FA(%)	SPK ERR (%)	DER (%)
7.0	1.6	8.6	7.2	15.79
6.6	1.7	8.3	9.5	17.76
6.1	2.0	8.1	6.6	14.72
5.7	2.3	8.0	4.6	12.61
5.3	2.7	8.0	6.7	14.72
5.2	2.8	8.0	4.7	12.76
5.0	3.2	8.2	5.0	13.1
4.7	3.8	8.5	6.3	14.80
4.4	4.7	9.1	5.6	14.77
4.2	6.4	10.6	5.3	15.94

Table 5.4: *Development results on the RT07s conference dataset depending on the SVM model's bias* b*. The trade-off between miss speech (MISS) and false alarms (FA) and the impact on the total DER error can be felt by examining the corresponding columns.*

that the diarization are more robust to misses errors than to false alarms. In that sense, the tuning of speech activity detection module during the elaboration of this thesis follows the conclusions presented in this section.

5.4.3 Baseline AHC Single Channel Diarization

The diarization system described in the section 5.1 and depicted in the figure 5.1 is taken as reference system for comparing the improvements by applying different techniques. In the figure 5.21 the diarization error rate per excerpt and its standard deviation is reported. Such a result is summarized in the table 5.5, reporting DER per both development and evaluation datasets. The diarization systems corresponds to a SDM system without any of the improvements proposed in this PhD thesis and without the use of any speech activity detection module. Aiming to study the influence in the system performance due the false alarms and misses caused by a non perfect SAD, an oracle SAD is employed by taking as SAD references the official NIST reference results. It is worth to recall that, despite of there is none dependence of the system on training data for development, some parameters of the diarization system has been tuned using the RT06s and RT07s data, such as initial number of clusters, the cluster complexity or the minimum duration, see previous sections, which still depend on the nature of the meetings. Finally, a "blind" test is conducted on RT09s database to assess the generalization of the algorithm to unseen data.

Looking at the figure 5.21, it can been observed the DER degradation by taking into account the speech overlap, that is, segments of time in where two or more speakers are talking at the same time. This increment can be noticed in both the total DER, the red color bar at the right side of the picture, and the standard deviation per excerpt, σ. The total DER raises 8.75% absolute, from 35.65% to 44.4%, just counting as miss errors the overlap speech appearing in the recording. As expected, the standard deviation per excerpt also increases,

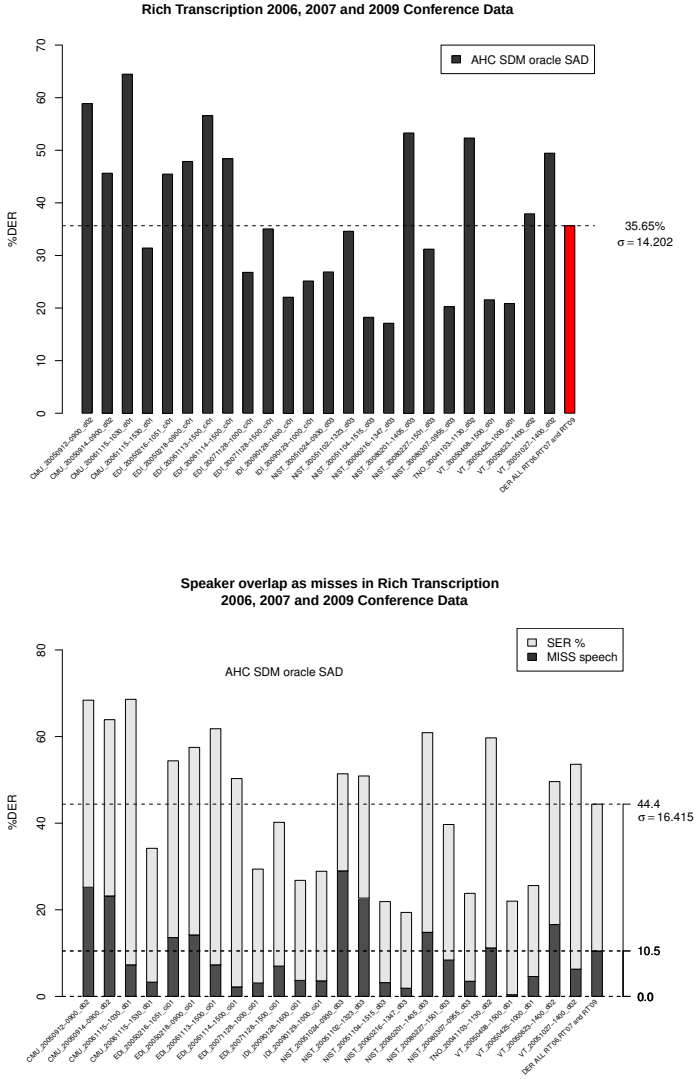


Figure 5.21: *Diarization error rate (DER) results on NIST Transcription evaluation conference data. At the top, baseline system without any improvement and with an oracle sad. At the bottom, DER degradation taking into account the speaker overlap regions as misses due to the fact that just one label is provided by the diarization engine.*

from 14.20 to 16.41. These results evidence the great impact on the diarization performance due the speaker overlap issue and suggests mechanisms to handle such situations in order to reduce the error caused by the speaker overlap regions. This issue will be analyzed in the next sections.

It is worth to mention the high percentage of error obtained by the baseline system in both development and

	AHC SDM Baseline %DER / σ	AHC SDM Baseline overlap %DER / σ	Total Time seconds (seconds)	Time μ/ σ
RT06	44.33% / 9.73	56.62% / 6.39	84,119	9,346 / 5,355
RT07	34.83% / 18.92	37.95% / 19.56	94,550	11,818 / 6,002
RT06+RT07	39.86% / 15.10	47.83% / 16.74	178,669	10,509 / 5,632
RT'09	30.54% / 11.24	35.66% / 12.78	105,351	15,050 / 7,645

Table 5.5: *DER results and standard deviation (σ) per set in Rich Transcription 2006, 2007 and 2009 conference data by the baseline system.*

evaluation sets. Focusing on the results with overlap, that is the NIST RT official metric see section 2.3.3, we can see that the worst results are obtained in development datasets whilst better performance is observed in the evaluation RT09s set. Similar results has been reported in the literature [Anguera *et al.*, 2011] and in the RT workshops [Fiscus and et al., 2009b]. This DER variation among datasets may be due to the data characteristics itself such as the excerpt duration or the difference in speakers' interactivity. Nonetheless, the impact of overlap is noticed in any of the RT datasets.

In addition, the effect of "flakiness" is also also visible in the same table by looking at the σ value. The DER variance obtained per dataset counts for more than 30% of the total DER. It translates into a big difference in DER among recordings, even from the same site. For instance, in non-overlap metric, the DER obtained after processing the show NIST_20051104-1515 is around 2.21% while that obtained by processing the NIST_20080201-1405 recording is over 40.76%. Same situation is also noticed in the DER variation by the parameters tuning of the diarization approach.

Table 5.5 gives a summary of the results independently per both development and evaluation sets. The higher error due the speaker overlap is perceived in the RT06s, not a surprising result since this dataset presents the higher number of overlap regions compared to others RT datasets, see the Appendix C. Overall, the overlaps degrades the total DER by two main factors: Directly as *miss speech* since the baseline diarization approach just provides one speaker label at the same time and by *cluster purity* due to the fact that clusters might be contaminated by speech coming from several speakers corresponding to those overlap regions.

The configuration used to obtain the results for the diarization baseline includes fixing a variety of parameters. Among them, the initial number of clusters was fixed to 65 and initialized following a homogeneous splitting of the recording data. The number of Gaussians to model each cluster was set to 5 and the minimum duration to 250 milliseconds. The feature set is composed of 16 MFCC static parameters and an oracle speech activity detection module based on the NIST references was employed. The parameter Q which defines the precision (number of bits) to compute the logarithm function was set to 12.

Following, a brief summary of experiments related to techniques discussed in previous sections is given. The behavior of the agglomerative hierarchical clustering system is studied to determine which parameters and modules impact accuracy most. Therefore, it aims to quantify the impact on the DER by applying the different strategies and algorithms with respect to the results of the baseline single channel diarization reported above.

Cluster Initialization

The cluster initialization is a key feature in the AHC diarization approach. Most of the system makes uses of an uniform splitting of the data to obtain a fixed number of initials clusters for further processing. Nonetheless, such approach combined with the agglomerative hierarchical clustering is highly dependent on the number of clusters and the quantity of data belonging to each of them, that is, to the total length of the recording to process.

In addition, Gaussian mixtures models, as generative models, specially suffer from the cluster impurity. In the case a cluster composed with speech from two different speakers, there exists a trade-off between over fitting and under fitting the cluster data. In the overfitting case, the overlapped data will lead, at the end of the agglomerative clustering, to a erroneous speaker cluster that likely will join two different speakers. In the

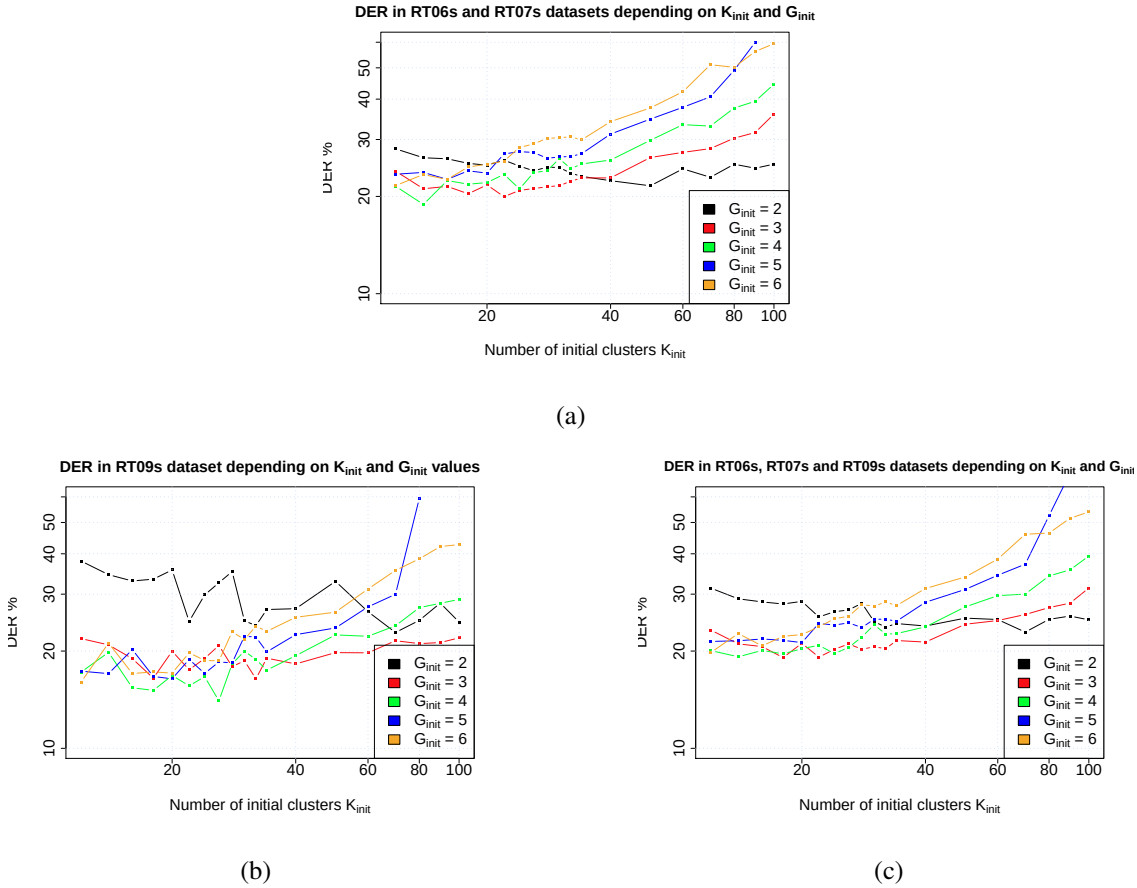


Figure 5.22: *Diarization error rates (DER) for the baseline diarization system varying the number of initials clusters, K_{init}, and the number of Gaussians mixtures used to model each of them, G_{init}. (a) In the development RT06s and RT07s datasets, (b) In the evaluation RT09s dataset and (c) In all the three RT datasets. The oracle SAD has been employed and no overlap is taking into account.*

underfitting case, a not good estimated model mights contaminate other clusters models leading to erroneous cluster merges during the agglomerative procedure. In any case, it yields to raise of the speaker diarization error of the system.

The figure 5.22 draws the DER curves in log-log scales for the baseline diarization system varying the number of initials clusters, K_{init}, and the number of Gaussians mixtures G_{init} used to model each of them, see section 5.1.2. The figure (a) shows results on the development RT06s and RT07s datasets, while the the figure (b) does on the evaluation RT09s dataset. In the figure (c) the DER error curves are depicted for all RT datasets. The oracle SAD has been employed and no overlap is taking into account to compute the DER metric. As can be noticed looking at the results on the development set, the figure (a), lowest values of DER are obtained using 3

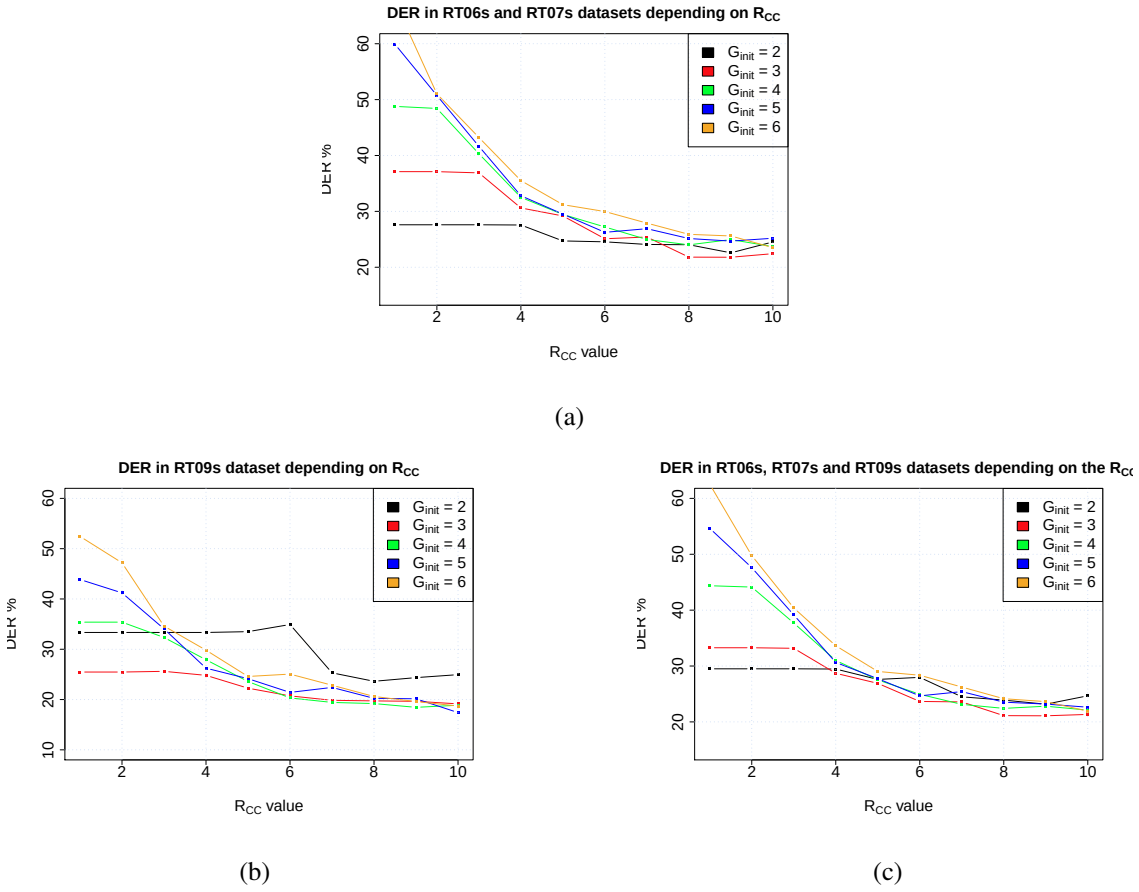


Figure 5.23: *Diarization error rates (DER) for the baseline diarization system with automatic estimation of the number of initial clusters. The figures depict the DER curves varying the number of Gaussians mixtures used to model each cluster, G_{init} and the R_{CC} which defines the number of frames assigned to each Gaussian. (a) In the development RT06s and RT07s datasets, (b) In the evaluation RT09s dataset and (c) In all the three RT datasets. The oracle SAD has been employed and no overlap is taking into account.*

or 4 Gaussians and a number of initial clusters ranging between 15 and 30, which correspond to the red and green curves. A trend is also observed by raising the number of Gaussians which model the clusters. From $M = 3$ forward, the values of DER increase as the number of cluster does. Same conclusions can be extracted from the figures 5.22 (b) and (c).

As the accuracy of the diarization system is indeed very sensitive to the values chosen for the initialization parameters and factors such as the duration of speech in the recording, an automatic method is also implemented. The automatic selection of initial clusters is based in equation 5.1 and computes the number of clusters based on the number of Gaussians mixtures G_{init} initially assigned per cluster and upon R_{CC} which defines the number of frames per each Gaussian. Such a parameters have still to be fixed in this estimation of initial clusters. The figure 5.23 draws the DER curves depending on the number of Gaussians mixtures used to model each cluster, G_{init} and the R_{CC} value. (a) In the development RT06s and RT07s datasets, (b) In the evaluation RT09s dataset and (c) In all the three RT datasets.

As in the previous curves, the oracle SAD was employed and no overlap was taking into account to compute the DER. The results on the development set, see the figure 5.23 (a), show that lowest values of DER are obtained using 3 or 4 Gaussians and among $6 - 8$ seconds of training speech per each of them. It corresponds to the red and green curves. Anyway, both 5.22 and 5.23 figures show a similar behavior, evidencing a trade-off between the initial number of segments and the available data to compute Gaussian parameter statistics. In conclusion, $20 - 30$ initial segments seems to be a good starting point for the agglomerative algorithm in the RT datasets. The latter along with models build of 4 Gaussians appear to obtain lowest DER results based on our experiments.

Finally and looking for a method to reduce manual tuning of previous presented parameters, a new initialization algorithm [Imseng and Friedland, 2010] is compared, see section 5.1.2. In order to estimate R_{CC}, we employ the linear regression:

$$R_{\text{CC}} = 0.01 \cdot \text{speech in seconds} + 2.6 \tag{5.30}$$

The figure 5.24 depicts the DER per recording obtained using the estimated R_{CC} based upon the length of the show. The improvement in DER is over 3% absolute, from 24.4 to 21.11. The baseline SDM system used for comparison just differs in the R_{CC} and G_{init} parameters which were fixed to 7 seconds and 5 Gaussians respectively per each show, it corresponds to a point in the blue curve in figure 5.23 (c). It is worth to recall that in the case of the automatic approach such a value is estimated independently per each recording.

In overall and comparing to the manual tuning results per different couples of values $(R_{\text{CC}}, G_{\text{init}})$, represented in previous figures (a), (b) and (c) in 5.23, the automatic estimation of R_{CC} proposed reaches comparable results than those obtained by the best values of $(R_{\text{CC}}, G_{\text{init}})$ couples. As can be seen in the figure 5.23 (c), the couple $(R_{\text{CC}}, G_{\text{init}}) = (8, 3)$ reaches around 21% of DER whereas even lower DER values are obtained in figure 5.22. DER around 20% are observed by fixing the couple $(K_{\text{init}}, G_{\text{init}})$ in the range $(15 - 30, 3 - 4)$ which corresponds to the green and the red curves.

As conclusion, both the automatic estimation of K_{init} in the first approach and the estimation of the couple

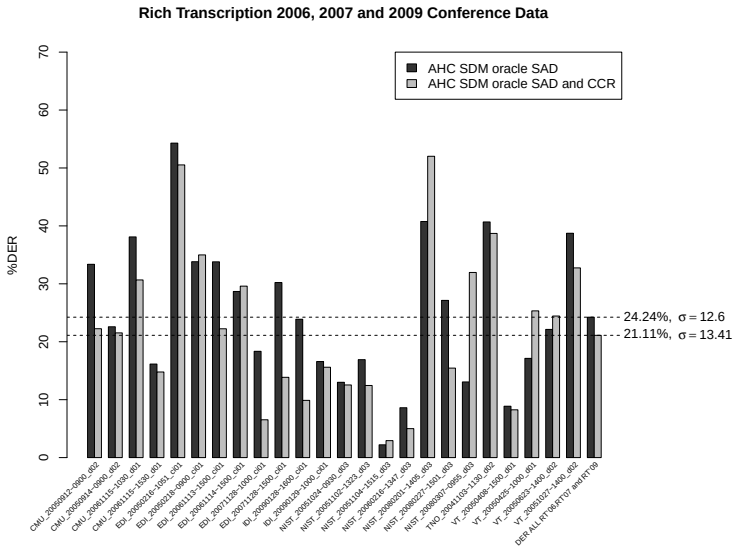


Figure 5.24: *Diarization error rate (DER) results on NIST Transcription evaluation conference data. DER improvement by using CCR.*

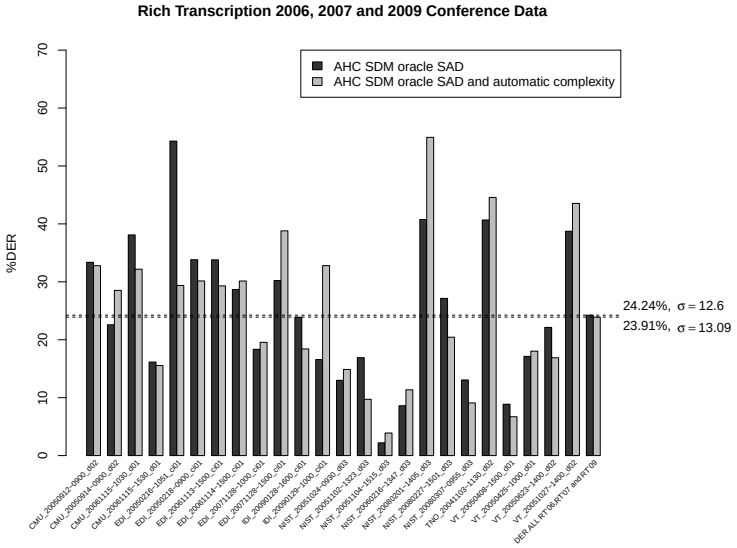
$(R_{\mathrm{CC}}, K_{\mathrm{init}})$ in the second seems to be well suited to the NIST data while avoiding the manual tuning of such a parameters. In next subsections will be discussed an original approach to initialize the cepstral diarization algorithm based on the speaker position estimation inside the room.

Model Complexity Selection and Multiple Merging Criterion

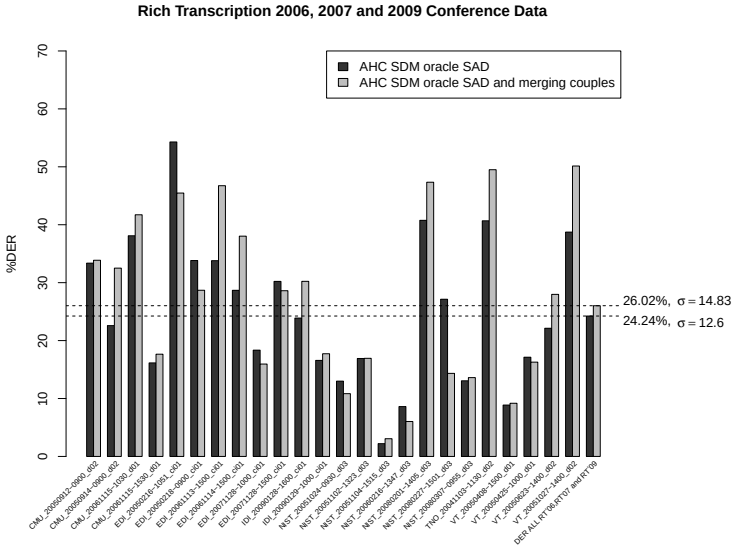
The model complexity selection is not a new algorithm [Anguera *et al.*, 2006d] but its importance has been addressed in several works and diarization approaches based on AHC clustering. It is also used as a core module in the diarization system employed in this PhD thesis. This subsection is devoted to report results about complexity selection algorithm, see section 5.1.3, aiming to assess the impact of the algorithm in our diarization approach. In addition, experiments about the new merging criterion proposed in section 5.1.4 which merges more than one cluster in each AHC iteration are also reported. The figures in 5.25 show the DER per each technique.

For the model complexity a factor of $R_{\mathrm{CC}} = 7$ was fixed to compute at each agglomerative iteration the number of Gaussians which compose each cluster. In the case of the merging criterion, equation 5.10 sets the threshold on the BIC for selecting those couples to merge.

The DER results for the complexity selection algorithm does not show a significant improvement over the SDM baseline. For the merging criterion the situation is even worst by degrading the baseline DER over 1.7 absolute. Nonetheless, we will see in next subsections that these techniques along all other improve the



(a)



(b)

Figure 5.25: *Diarization error rate (DER) results on NIST Transcription evaluation conference data. (a) DER improvement by using complexity selection and (b) DER decreasing by using merging more than one couple at each iteration depending upon BIC values.*

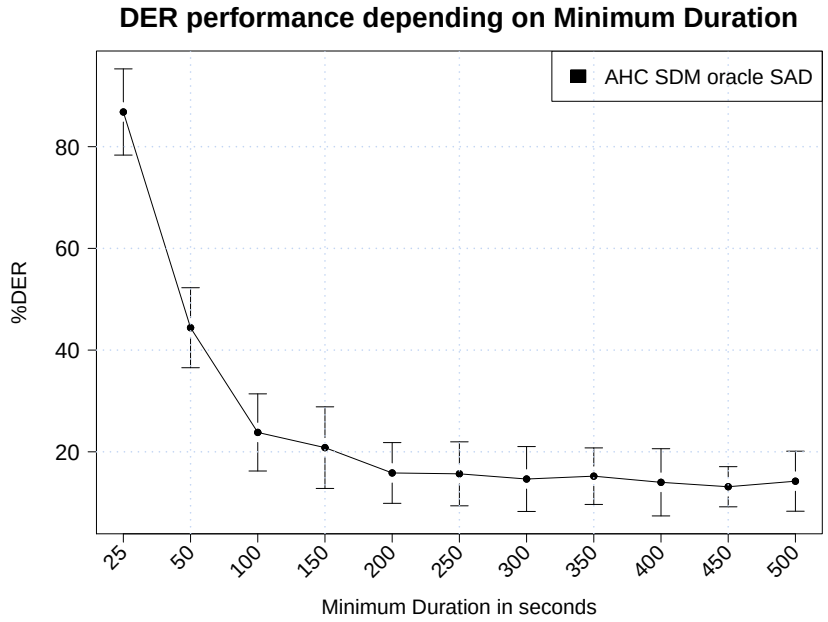


Figure 5.26: *Diarization error rate (DER) results on NIST Transcription 2006 and 2007 evaluation conference data depending on the minimum duration taken into account in the HMM decoding.*

baseline in spite of their individual performances.

Minimum Duration in the Speaker Turn and Turn Taking Modeling

The minimum duration constrain imposed upon the HMM topology, see figure 5.7 in section 5.1.3, guarantees a minimum length in the speaker turn duration. It avoids so short speaker turns and ensures enough speech data for estimation of speakers models. The figure 5.26 draws the diarization performance depending on this parameter. The baseline SDM system with oracle SAD was used for drawing the picture. However, the baseline was augmented with previous techniques: the automatic initialization based on the regression 5.5 and with the two techniques presented in the previous section.

It is worth to mention that once the agglomerative clustering has finished, a last Viterbi alignment using a minimum duration value of 1 second is conducted until no variation in the clustering is noticed.

Such a last Viterbi step is in charge of detecting shorter speaker interventions once a reliable estimation of speakers cluster has been done. The curve clearly shows a drastic drop in the diarization error from 25 to 150 milliseconds, then it stands around the minimum DER value reached at 450 milliseconds. Despite of this, most of the experiments reported in this PhD thesis proposal were carried out fixing the minimum duration value to 250 milliseconds.

In the modeling of the speaker turn-taking, the n-gram sequences of speakers occurrences were combined

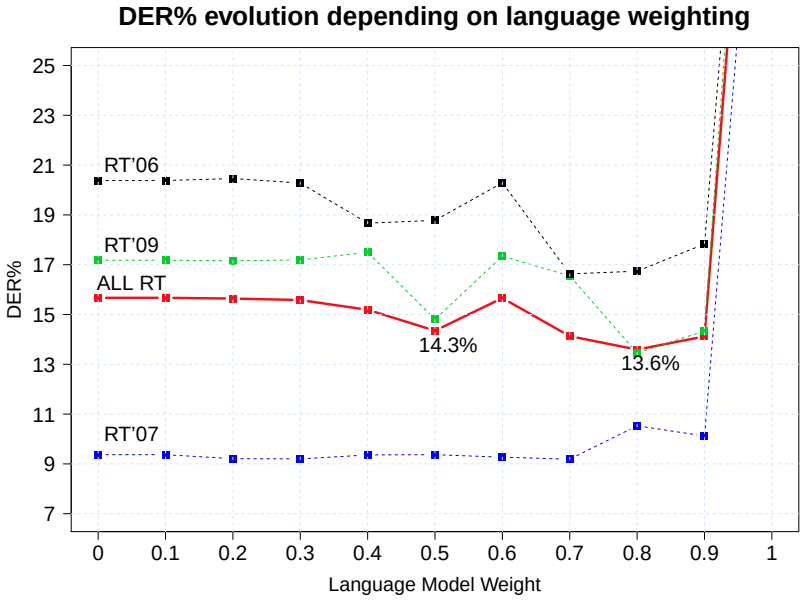


Figure 5.27: *Diarization error rate (DER) results on NIST Transcription evaluation conference data depending on the language model threshold. All three language model strategies are implemented.*

LM technique	RT'06 & RT'07	RT'09	RT'06	RT'07	RT ALL
Unigram ($w = 0.7$)	12.98 %	17.57 %	16.85 %	**9.04 %**	14.49 %
Unigram + Trigrams ($w = 0.9$)	12.91 %	14.06 %	15.84 %	9.92 %	13.29 %
Unigram + Weighted Trigrams (350ms)	12.54 %	**14.03 %**	14.98 %	10.04 %	**13.03 %**
Unigram + Weighted Trigrams (250ms)	**12.33 %**	14.73 %	14.99 %	9.62 %	13.12 %

Table 5.6: *Development and evaluation best results of the language model based on weighting the speaker transition probability, $\Pr(S)$. In brackets the weight for which the DER value is reached. The* LM technique *column stands for the way to compute such a transition matrix:* ***Unigram****, counts the frequency of occurrence of that speaker,* ***Unigram + Trigrams****, also takes into account trigrams of the form A?A oriented to anchor speakers who tend to speak before and after an intervention. The* ***Unigram + Weighted Trigrams*** *cases of 350 ms. and 250 ms. give even more weight to that speaker interruptions with shorter interventions, see section 5.1.5 for further details.*

with the acoustic information from MFCC features. As commented in the section 5.1.5, several strategies are proposed based on weighting the speaker transition probability of changing speaker turn, $\Pr(S)$. The figure 5.27 depicts the DER curves per each RT depending upon the weight assigned to the language model scores. In this case, the speaker-turn system makes use of unigrams along with weighted trigrams as explained in section 5.1.5. Looking at the picture, the simple speaker-turn modeling seems to work since it obtains significant DER reduction in RT'06 and RT'09 datasets. In such a situations, the DER is improved by 4 absolute points. Nonetheless, in RT'07 dataset the DER value degrades over 2 absolute points. Anyway and in overall, the red curve, that shows the DER per all RT data, shows a significant drop for weight values within the range $(0.5, 0.8)$. It is worth to mention that previous weight values not directly represent the portion of the final score since scores from acoustics and speaker-turn were not normalized.

Table 5.6 compares the different results by using several techniques in order to model the speaker-turn transitions. As can be observed, the combination of acoustic probabilities along with the probability of speaker turn improve the DER for all the techniques and in most of the datasets, except for the RT'07 data. Anyway, such a degradation does not impact significantly the global DER for the development dataset, RT'06+RT'07, leading to an overall improvement in DER which is also noticed in the evaluation set RT'09. The *LM technique* column stands for the way to compute the speaker transition matrix. **Unigram**, counts the frequency of occurrence of that speaker, **Unigram + Trigrams**, also takes into account trigrams of the form A?A oriented to anchor speakers who tend to speak before and after an intervention of the same speaker. The **Unigram + Weighted Trigrams** cases of 350 ms. and 250 ms. give even more weight to that speaker interruptions with shorter interventions, see section 5.1.5 for further details. The results give evidence of the importance of modeling speaker transitions in speaker diarization and that there is still a room for improvement in the combination with acoustic probabilities.

Speed-up depending on Logarithm Precision

In this subsection experiments adjusting the number of Q most significant bits employed to quantize the mantissa in the logarithm approximation, see section 5.1.6, are reported. The figure 5.28 shows the DER behavior aiming to assess the influence of the logarithm precision on the diarization error and the trade-off with the time consumption.

The logarithm lookup-table is indexed using the Q most significant bits of the mantissa. Accuracy is lost because of such a quantization of the mantissa, however not a significant drop in DER performance is noticed while the computation time is reduced in a factor of 3 or more for some recordings. Furthermore, the DER results in the figure 5.28 show that the best DER is reached using the $Q = 15$ most significant bits while worst error even is obtained by means the full mantissa, that is, $Q = 23$ bits.

The figure 5.29 depicts the mean time spent by the diarization algorithm to process a recording. The curve was obtained by computing the mean time consumption per show using 24 recordings belonging to NIST RT 2006, 2007 and 2009 datasets. In addition, the standard deviation is also reported. As can be noticed, there is a clear trade-off between the logarithm precision employed, number of Q bits, and the time consumption of the

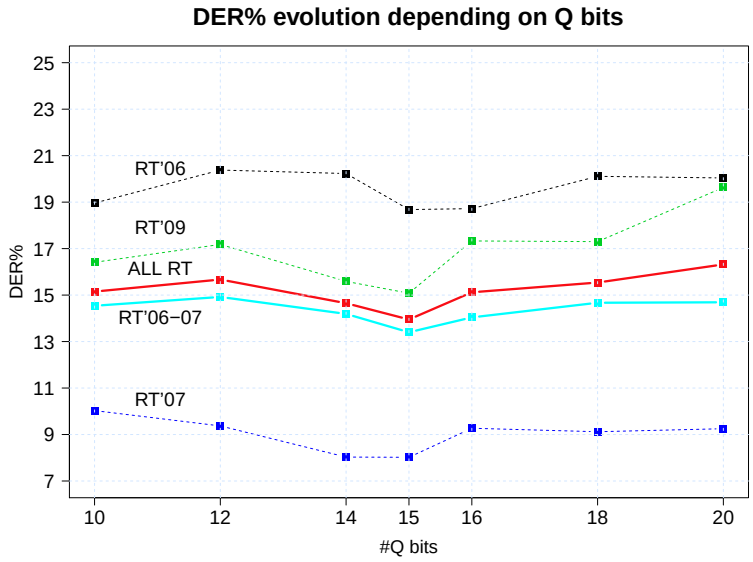


Figure 5.28: *Diarization error rate (DER) results on NIST Transcription evaluation conference data depending on the number of Q most significant bits employed to quantize the mantissa, see section 5.1.6.*

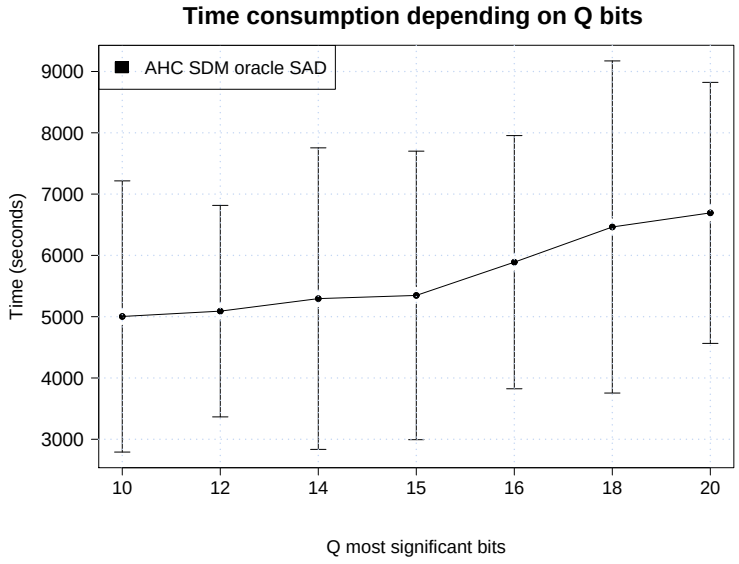


Figure 5.29: *Mean time consumption per show in seconds on NIST 2006,2007,2009 conference data depending on the number of Q most significant bits employed to quantize the mantissa, see section 5.1.6. It is depicted the mean duration in processing a recording and the standard deviation. Curves were obtained with the AHC system augmented with all techniques and working with* 22 *MFCC parameters.*

algorithm. Using $Q = 10$ bits save in mean more than 25 minutes in processing a recording w.r.t. the same processing but using $Q = 20$ ans without a drastic drop in DER performance, see the figure 5.28 above.
Aiming to put emphasis in the time consumption reduction rather than improved DER results, in the experiments performed during this PhD thesis a value of $Q = 12$ was selected in the quantization of the mantissa in order to build up the lookup-table.

Single Channel Algorithms Agglomeration Performance

This subsection is devoted to report the different improvements reached by the agglomeration of the previous algorithms in the single diarization channel condition.
One of the most common parameter to tune in speech applications is the kind of parametrization to use and the number of speech parameters or feature vector dimension. In the experiments presented in this section, we have decided to employ classical MFCC parameters, despite of the work submitted to RT'07 [Fiscus and et al., 2007a] where similar results were obtained by means FF parametrization. The use of delta and delta-delta coefficients is also not considered since worse results were obtained by augmenting the feature vector with these parameters. In such a situation, the results presented can be directly compared with most of the state of the art systems which employ similar speech parameters.
The figure 5.30 depicts the DER curve obtained by shifting the number of static MFCC parameter from 8 to 30. The diarization algorithm employed joins together all the techniques previously commented. As can be seen in the picture, the lowest DER is reached at 22 MFCC features. A wide variance is also noticed among the different recordings, translated to a standard deviation per feature set ranging between 10.89 and 16.17 in DER. Due this fact, the MFCC set of size 22 was selected to perform the baseline experiment in this subsection and the successive agglomeration steps aiming to offer an easy comparison. The baseline experiments showed in the following tables 5.7 and 5.8 contrasts to results presented in the figure 5.21 which corresponds to a 16 MFCC parameter set and a different group of cluster initialization parameters.
Table 5.7 shows the DER improvement among the development and evaluation subsets by adding the previous

Techniques	Development		Evaluation	
	RT'06 & RT'07	r.i. %	RT'09	r.i. %
Baseline	25.24% (13.95)	–	24.29% (9.45)	–
+ Estimation $(R_{\text{CC}}, K_{\text{init}})$	22.88% (12.72)	9.35%	20.76% (15.95)	14.53%
+ Automatic complexity	22.62% (14.97)	1.13%	21.82% (16.92)	−5.10%
+ Merging couples	20.94% (13.92)	7.42%	20.61% (15.41)	5.54%
+ Language modeling $w = 0.8$	18.61% (11.13)	6.35%	18.45% (20.91)	10.48%

Table 5.7: *Development and evaluation results, in terms of DER and standard deviation, of the successive agglomeration of the previous techniques and relative improvement (r.i) per cent w.r.t. the previous system.* 16 *MFCC parameters were employed in all systems.*

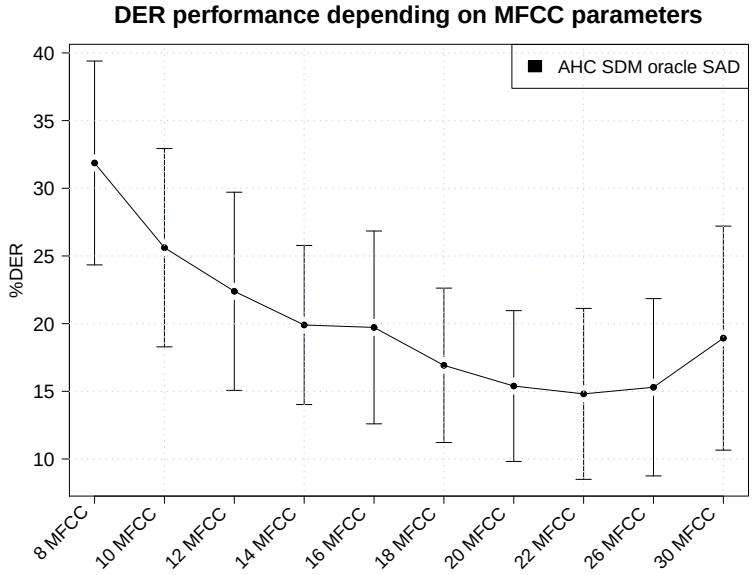


Figure 5.30: *Diarization error rate (DER) results on NIST Transcription evaluation conference data depending on the number of MFCC parameters extracted from the speech waveform. Baseline system without any improvement and with an oracle sad.*

Techniques	development		all data
	RT'06	RT'07	RT ALL
SDM Baseline	26.17% (14.63)	24.17% (14.06)	24.40% (12.60)
+ Estimation $(R_{\mathrm{CC}}, K_{\mathrm{init}})$	24.58% (13.56)	20.98% (12.32)	21.11% (13.41)
+ Automatic complexity	22.57% (15.36)	22.66% (15.58)	21.08% (15.19)
+ Merging couples	22.06% (15.09)	19.68% (13.38)	19.72% (14.02)
+ Language modeling $w = 0.8$	21.74% (9.63)	15.53% (10.05)	18.59% (10.15)

Table 5.8: *The two first columns show the individual development subset results. The global results per all datasets are also showed in the last column. Results are in terms of DER and standard deviation – value between brackets –, of the successive agglomeration of the previous techniques.*

techniques to a common baseline in order to assess the joint reached improvement. Looking at the results, both the linear regression estimation of values for clustering initialization and the language modeling techniques show the biggest DER improvement w.r.t. the baseline system. Such a situation is comparable for both datasets reaching similar DER improvements around 4% absolute. Nonetheless, by augmenting the system baseline with the different techniques, the standard deviation in the RT'09 data considerably arises. It suggests a greater variance in the individual DER results compared to the baseline which along with the lower DER, likely means that some particular shows in the RT'09 data reduce considerably its DER compared to the rest.
In contrast, the complexity and merging couples algorithms does not report a drastic DER reduction. In the first case, we shall see it is an important algorithm for the combination of acoustics and TDOA features while in the "merging" algorithm its use is supported by the reduction in terms of time computation whereas it keeps DER error constant.
Table 5.7 evidences same effects we have already commented in previous paragraph. It can be noticed in the last column which shows the DER for the whole RT datasets. Analyzing the development data individually, first and second column, it can be noticed a drastic DER drop, around 6% absolute, by applying the language modeling in RT'06 dataset whereas DER in RT'07 raises over 3%. It suggests a really different structure among data shows, with highly interactive meetings and discussions in the RT'06 set compared to the next one. Despite of evaluation and global results show the usefulness of the agglomeration of the different algorithms, the individual results still suggest they could be improved and applied in a data dependent way somehow as is the case of the language modeling.

5.4.4 AHC Multichannel Diarization

This subsection is devoted to provide results on the Rich Transcription datasets by improving previous single diarization system with the information provided by the TDOA features as explained in subsection 5.4.4. Main differences w.r.t. SDM baseline are the Wiener filtering of multiple channels, the diarization on a beamformed channel and inclusion of second feature stream composed of TDOA values among pair of microphones. The TDOA features were obtained as in equation 2.74 and they were combined jointly with acoustic observation at the score level as stated in equation 5.21. The MDM baseline system employed in the following experiments is mainly depicted in figure 5.9.

Weighting TDOA Features

The figures 5.31 and 5.32 report the diarization error rate results reached for the multichannel diarization system. The TDOA features computed for each show were combined with the spectral features in order to compute the output HMM probabilities. They are weighted in the range: $w = 0$ that uses just spectral features, to $w = 1$ in where diarization is performed by means only the TDOA stream. The DER obtained by varying the TDOA stream weight reach lower values at low values of the weight. Worst diarization results were obtained as the system more relies on TDOA values instead of spectral information. The MDM with weight value $w = 0.1$ reached the best results on development data and the same improvement can be noticed

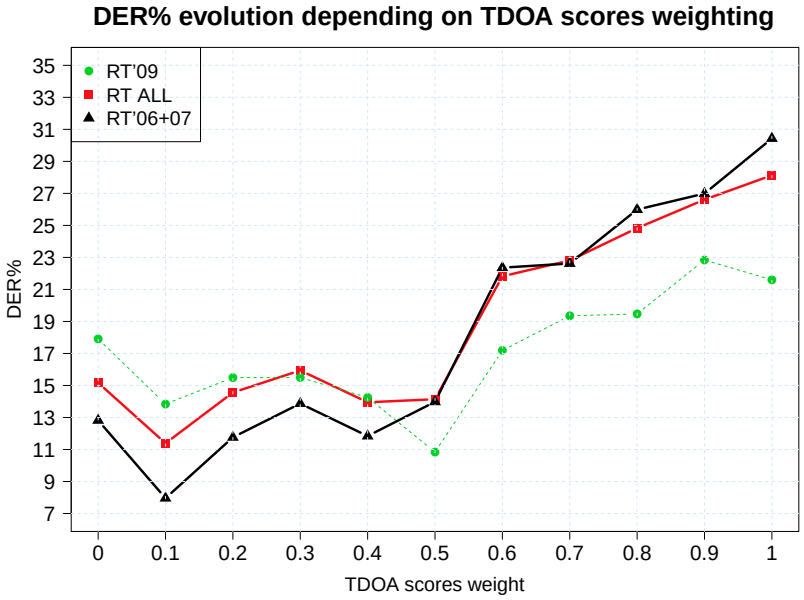


Figure 5.31: *Diarization error rate (DER) results on NIST Transcription conference data depending on the TDOA score weighting. Results are reported with an oracle speech detection based upon NIST references. The experiments reported make use of a MFCC stream of dimension 22 and a TDOA stream of variable dimension per show (number of couples of mics).*

in RT'09 evaluation data. Despite of the good results obtained on the evaluation data, TDOA weight seems highly dependent on the show recording and a specific weight per show should likely improve the global DER, getting close the development and evaluation curves in the figure 5.31 for all weight values.
The figure 5.32 reports the DER results provided by the same system and using a real speech activity detector. The DER curve evolution follows same trend as the system employing reference speech/non-speech labels. The inclusion of a real SAD arises in a DER degradation around 10% and 4% per cent on development and evaluation respectively. Such a result evidence a clear difference from recordings among the two subsets possibly due to differences in room setup and speaker interactivity. In overall, the figure 5.33 reports the decomposition of DER in terms of misses, false alarm and speaker error for all NIST RT data. The error due to SAD inclusion is around 8% which is comparable to the SER error around 11% which enforces the idea of the importance of a reliable speech activity detection algorithm to improve diarization results. It is also worth to mention the high value of DER variance per show. The recordings NIST_20080201-1405, TNO_20041103-1130 and VT_20051027-1400 present the higher errors contributing to the high variance value. The individual analysis of previous recordings gives some clues about such particular bad performance. The NIST_20080201-1405 show is composed uniquely of six female speakers with high interactivity and speech overlap. For the TNO_20041103-1130 and VT_20051027-1400 recordings possibly low level signal from far-field mics along with highly interactivity are the main reasons for the low DER obtained.

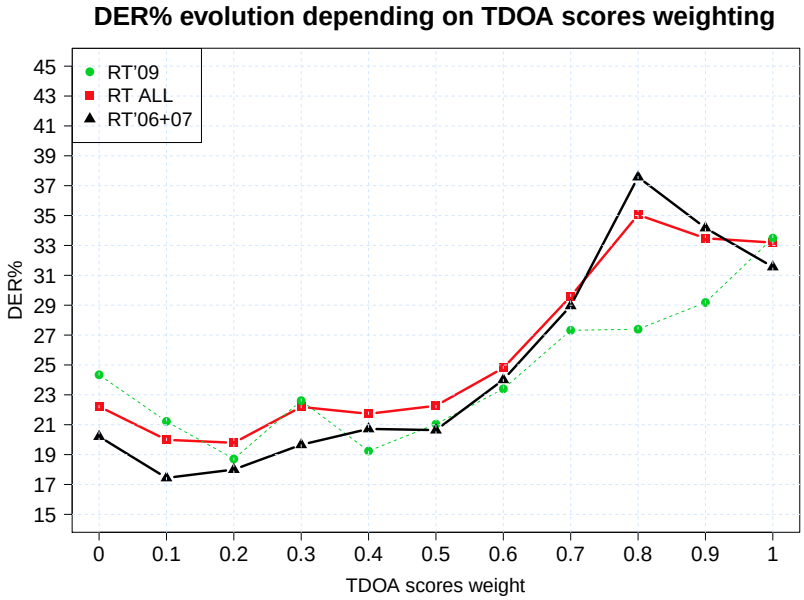


Figure 5.32: *Diarization error rate (DER) results on NIST Transcription conference data depending on the TDOA score weighting. A speech/non-speech detector is employed.*

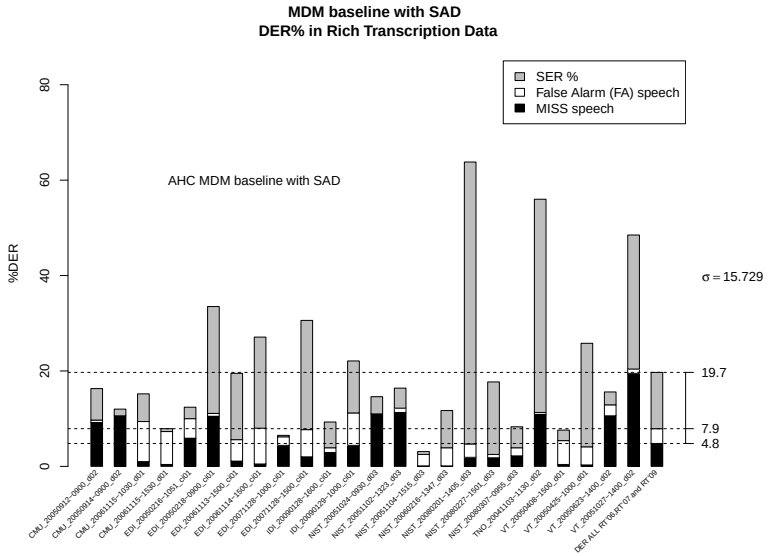


Figure 5.33: *Results in terms of speaker error (SER), miss speech (MISS) and false alarm (FA) speech results on NIST Transcription evaluation. Multiple Distant Microphone (MDM) baseline with TDOA-weight* 0.1 *and automatic speech/non-speech detection.*

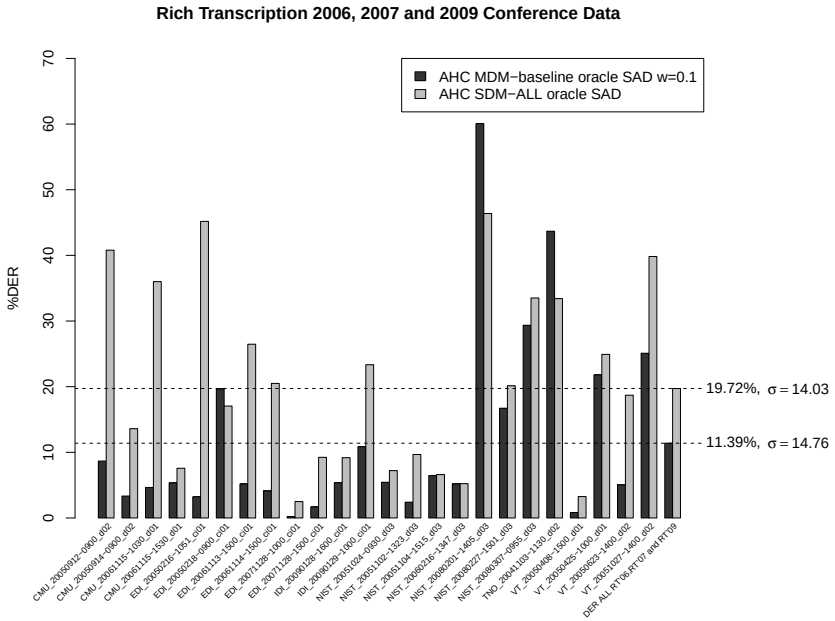


Figure 5.34: *Comparison in terms of Diarization Error Rate (DER) results on NIST Transcription evaluation conference between the Single Distant Microphone (SDM) system augmented with all techniques and the Multiple Distant Microphone (MDM) baseline with TDOA-weight* 0.1.

Comparison with SDM Baseline

The figures 5.34 and 5.35 report a comparison between the SDM system including all techniques, except language modeling, and the MDM system with TDOA weight fixed to 0.1. It can be observed a significant improvement w.r.t. the single channel system. Diarization error rate decreases from 19.72% in the SDM system up to 11.39% in the MDM case, around a 42.3% relative improvement. Such a improvement confirms results reported in diarization literature and for submitted RT systems [Pardo *et al.*, 2007; Fiscus and et al., 2009a; Fiscus and et al., 2007a].

The figure 5.35 reports the comparison between SDM and MDM systems in terms of speaker error, misses and false alarm time obtained by applying a real speech activity detector. It can be noted that in MDM case the speech detector is best adapted than in SDM case, obtaining low miss and false alarms rates, 2% absolute improvement. Anyway, it just account for a few part of the global DER improvement reached by the MDM with respect to SDM system, 19.7% and 25% respectively.

Finally, in order to assess the improvement reached by the agglomeration of techniques and by incorporating multi-microphone information, the table 5.9 provides a comparison between previous reported approaches. The table shows DER per cent results SDM baseline without any improvement, for SDM with most of them

Techniques	Development RT'06 & RT'07	Evaluation RT'09
SDM Baseline	25.24% (13.95)	24.29% (9.45)
SDM ALL	18.61% (11.13)	18.45% (20.91)
MDM Baseline ($w = 0.1$)	7.96% (7.36)	13.84% (21.19)
MDM Baseline ($w = 0.2$)	12.51% (14.18)	14.47% (17.04)

Table 5.9: *Development and evaluation results, in terms of DER and standard deviation, of the diarization system applied on the Single Distant Microphone (SDM baseline) condition and on the Multiple Distant Microphone (MDM) condition. The SDM ALL system is the same than SDM baseline system but augmented with the agglomeration of all previous techniques. The MDM baseline system is the same than SDM ALL system but incorporating the TDOA weighted scores to compute output HMM probabilities scores. SAD references were used.*

and for two MDM systems with best weight threshold tuned during development experiments. The best TDOA system which makes use of a weighting $w = 0.1$ outperforms SDM baseline in both development and

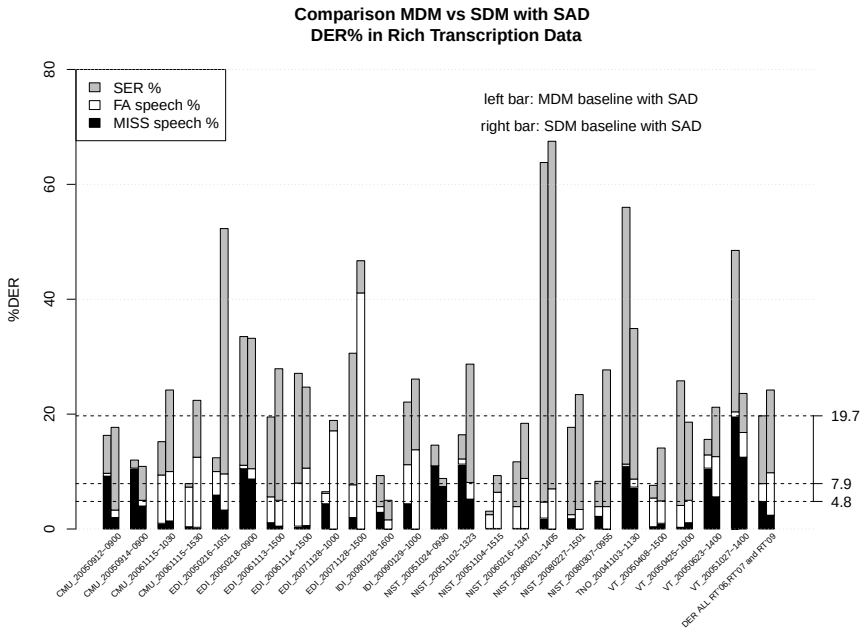


Figure 5.35: *Comparison in terms of speaker error (SER), miss speech (MISS) and false alarm (FA) speech results on NIST Transcription evaluation conference between the Single Distant Microphone (SDM) – right bar of each show corresponds – system augmented with all techniques and the Multiple Distant Microphone (MDM) – left bar of each show corresponds – baseline with TDOA-weight* 0.1*. The automatic speech/non-speech detection has been employed.*

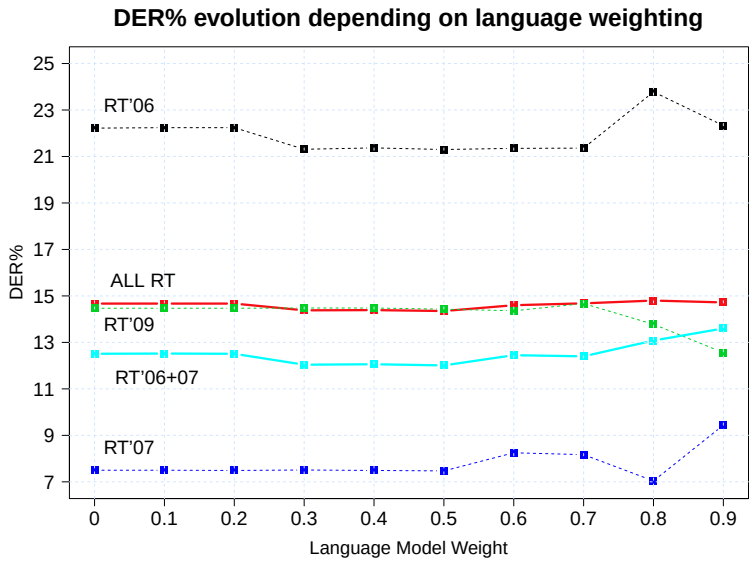


Figure 5.36: *Diarization error rate (DER) results on NIST Transcription evaluation conference data depending on the language model threshold. All three language model strategies are implemented*

evaluation sets. In development data, MDM improves DER over 31% relative w.r.t. SDM baseline and over 57% relative w.r.t. SDM ALL system (which uses all techniques, except language modeling). In the evaluation RT'09 data, the relative improvement is reduced compared to development results but still significant: 43% and 25% respectively.

Language Modeling

Results obtained in MDM condition by incorporating language modeling (LM) techniques are reported in figure 5.36. The inclusion of language modeling we did not report significant improvements in our experiments. Despite the improved results in the SDM case, same situation does not occur in MDM case. One possible reason for this low effect of LM in MDM condition could be the lack of normalization among acoustic and language probabilities which is performed in the HMM. Further research should be conducted in this topic in order to discern such a hypothesis. The figure 5.36 depicts the different DER curves for several datasets and depending upon the weight of the language modeling.

Agglomerating Techniques

The table 5.10 reports the results obtained by upgrading the MDM baseline system with initialization methods and algorithms as complexity selection or multiple merging. In addition an automatic weighting strategy [Anguera, 2006] was also used employed and compared with best results obtained by manually tuning. When setting the values by hand they are normally defined for all meetings equally and therefore they do not account

for peculiarities due to the meeting room or to the nature of the meetings. The automatic weight setting algorithm is able to compute the optimum values for each meeting independently. The weight is iteratively computed at the each segmentation iteration setting it proportional to the inverse of standard deviation of the BIC values among TDOA clusters. Its value remains without modification along the clustering process:

$$W_{\text{TDOA}} = \frac{\sigma^{-1}(BIC_{\text{TDOA}})}{\sigma^{-1}(BIC_{\text{TDOA}}) + \sigma^{-1}(BIC_{\text{MFCC}})} \; ; \; W_{\text{MFCC}} = 1 - W_{\text{TODA}} \tag{5.31}$$

where σ stands for standard deviation. The initial TDOA weight was set to same value, $w = 0.2$ as baseline system. Table 5.10 reports development and evaluation results, in terms of DER and standard deviation, of the agglomeration of several techniques to the MDM baseline system. The relative improvement (*r.i. %* column) per cent is also reported w.r.t. the baseline system.

It can be observed a different trend compared to the single channel approach results. In the case of cluster initialization methods, two methods were evaluated. The former based on linear regression – *MDM/*I_1 – and the second one based on the automatic computation of clusters setting $(R_{\text{CC}} = 7, G_{\text{init}} = 5)$ values – *MDM/*I_2. The former outperforms the second method in development data experiments. Nevertheless, results on evaluation dataset show a lack of generalization of the linear regression method despite of the improvement reached w.r.t. MDM baseline system.

Automatic complexity algorithm and multiple merging of clusters per iteration did not reach better results in development data w.r.t. not apply them. The DER reached corresponds to the lowest relative improvement with respect to MDM baseline, 3.52% and 9.91% respectively. However, *MDM/*A_1 and *MDM/*A_2 systems outperforms both the baseline and previous system reaching, in the case of *MDM/*A_2, the best relative improvement w.r.t the baseline system: 66.99%. The automatic TDOA weighting algorithm is shown as the most successful approach in terms of lowest DER in both sets of data. Last two rows in the table 5.10 report its DER per cent. It obtains the lower DER results independently the initialization method employed. In

Name (ID)	Techniques	Development		Evaluation	
		RT'06 & RT'07	r.i. %	RT'09	r.i. %
(**MDM/B**)	MDM Baseline $(w = 0.2)$	12.51% (14.17)	–	14.47% (17.04)	–
(**MDM/I$_1$**)	MDM/B + LR Estimation $(R_{\text{CC}}, K_{\text{init}})$	8.63% (7.43)	31.01%	13.54% (17.57)	6.42%
(**MDM/I$_2$**)	MDM/B + $(R_{\text{CC}} = 7, G_{\text{init}} = 5)$	9.96% (11.01)	20.38%	8.06% (9.28)	35.57%
(**MDM/A$_1$**)	(MDM/**I$_1$**) + Automatic complexity + Merging couples	12.07% (11.34)	3.52%	11.73% (23.63)	6.23%
(**MDM/A$_2$**)	(MDM/**I$_2$**) + Automatic complexity + Merging couples	11.27% (11.41)	9.91%	$\mathbf{4.13}\%$ $(\mathbf{5.57})$	66.99%
MDM-ALL$_1$	(MDM/**A$_1$**) + Automatic TDOA weighting	9.88% (9.54)	21.02%	6.76% (22.06)	45.96%
MDM-ALL$_2$	(MDM/**A$_2$**) + Automatic TDOA weighting	$\mathbf{7.95}\%$ $(\mathbf{9.50})$	36.45%	$\mathbf{5.60}\%$ $(\mathbf{12.98})$	55.24%

Table 5.10: *Development and evaluation results, in terms of DER and standard deviation, of the agglomeration of several techniques to the MDM baseline system. The relative improvement (r.i. % column) per cent is also reported w.r.t. the baseline system.*

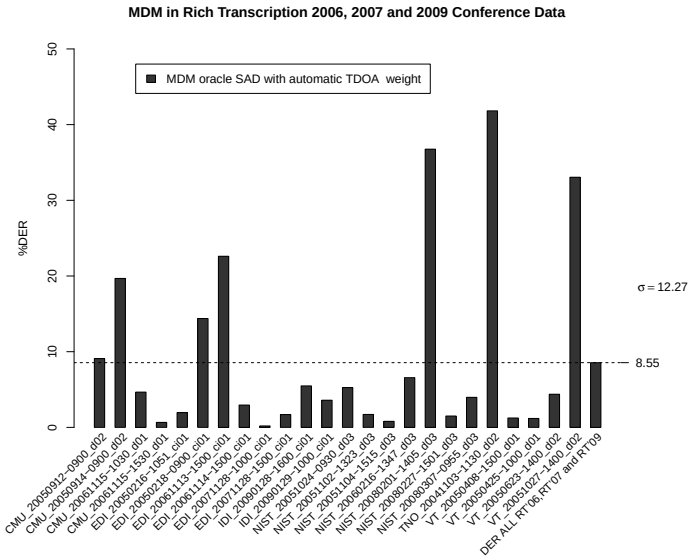


Figure 5.37: *Diarization error rate (DER) results on NIST Transcription evaluation conference data by the MDM system augmented with automatic computation of initial clusters, automatic complexity, multiple merging and automatic spatial feature weighting based on standard deviation of TDOAs, i.e.,* ***MDM-ALL***$_2$ *system in the table 5.10.*

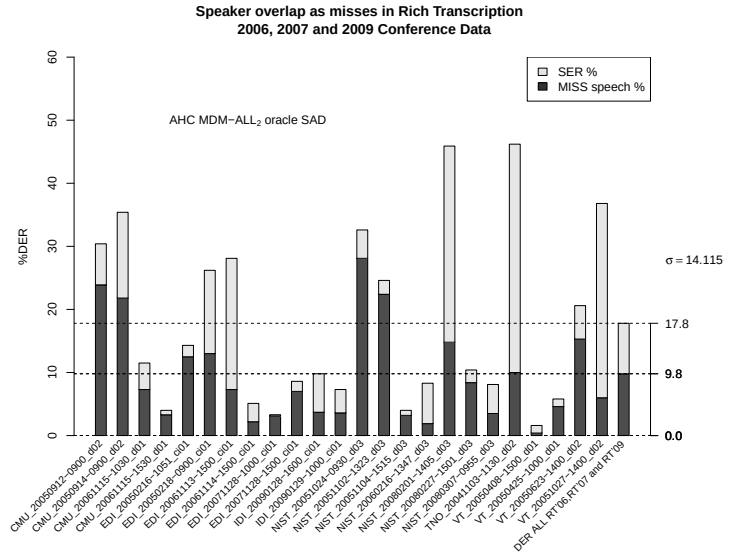


Figure 5.38: *Diarization error rate (DER) results on NIST Transcription evaluation conference data taking into account speaker overlap. The* ***MDM-ALL***$_2$*, see table 5.10, system results in terms of DER are reported.*

Show Name	ref. #SPK	system #SPK	diff	DER %
CMU_20050912-0900	5	4	1	9.09 %
CMU_20050914-0900	9	4	5	19.68 %
CMU_20061115-1030	4	4	0	4.66 %
CMU_20061115-1530	4	4	0	0.66 %
EDI_20050216-1051	5	4	1	1.95 %
EDI_20050218-0900	5	4	1	14.38 %
EDI_20061113-1500	5	4	1	22.61 %
EDI_20061114-1500	4	4	0	2.95 %
EDI_20071128-1000	4	4	0	0.19 %
EDI_20071128-1500	4	4	0	1.69 %
IDI_20090128-1600	5	4	1	5.48 %
IDI_20090129-1000	5	4	1	3.59 %
NIST_20051024-0930	5	9	-4	5.26 %
NIST_20051102-1323	8	8	0	1.71 %
NIST_20051104-1515	4	4	0	0.8 %
NIST_20060216-1347	7	6	1	6.57 %
NIST_20080201-1405	6	5	1	36.76 %
NIST_20080227-1501	6	6	0	1.5 %
NIST_20080307-0955	8	11	-3	3.97 %
TNO_20041103-1130	3	4	-1	41.82 %
VT_20050408-1500	5	5	0	1.24 %
VT_20050425-1000	4	4	0	1.18 %
VT_20050623-1400	7	5	2	4.38 %
VT_20051027-1400	6	4	2	33.05 %
DER ALL RT'06,RT'07 and RT'09	128	119	25	8.55 %

Table 5.11: *Results summary for the MDM-ALL$_2$ system. The second column ref #SPK gives the number of speaker present in the reference. The third column system #SPK the number of detected speakers by the system. The labeled column diff reports the difference between detected speaker and the reference. Last column DER% reports individual DER per cent values per recording.*

the case of manually adjusted $(R_{\mathrm{CC}}, G_{\mathrm{init}})$ values, the full enhanced MDM baseline system – *MDM-ALL*$_2$ – reaches 7.95% and 5.60% in development and evaluation data respectively. The figures 5.37 and 5.38 report the DER% per show in the whole RT data set, i.e., NIST RT'06-09. The former reports the speaker error (SER) per cent by the *MDM-ALL*$_2$ system which reaches 8.55% SER%[4] with a standard deviation

[4]It is worth to note that this result is not the arithmetic mean of the two SER% values reported in last column of the Table 5.10. DER and SER are time metrics which value is normalized by the total time in the data. That is the reason cause SER% computed taking into

$\sigma = 12.27$. This is the lowest SER% value reached by any of the systems during experiments. It can be observed the speaker overlap contribution to the diarization error (DER) in the figure 5.37. Last bar in the graph, dark gray part, corresponds to the speaker overlap error, 9.8% whereas the speaker error is around 8%. The flakiness effect is still perceived and reported by $\sigma = 14.11$ value, 6 out 24 recordings have DER% values higher than 30%. Most of the higher DER% values are due a high percentage of misses produced by speaker overlap. It can be note in CMU recordings, first two bars, and in first two NIST recordings – NIST_20051024-0930 and NIST_20051102-1323. Once more time, as in the MDM baseline see figure 5.33, the recordings NIST_20080201-1405, TNO_20041103-1130 and VT_20051027-1400 report the higher errors. As explained above, possibly reasons are high interactivity and and low level signal from far-field mics.
Finally, the Table 5.11 reports individual SER% values per each recording and the number of detected speakers by the *MDM-ALL*$_2$ system. In overall, the number of speaker is underestimated and there is not a clear relationship between number of speaker detected errors – *diff* column – and higher DER% values.

TDOAs for Clustering Initialization

In this subsection experiments of the proposed TDOA initialization, see section 5.2.3, are reported. In order to assess the TDOA initialization algorithm, it is compared to the uniform initialization based on R_{CC} value. The uniform initialization defines K_{init} initial clusters by splitting the input signal into equal parts and then iterates over model training and segmentation on the data in order to obtain acoustically homogeneous initial clusters. Both initialization techniques are compared using the data distributed for the NIST Rich Transcription 2007 Spring meeting Recognition Evaluation, RT07s as evaluation set whereas RT06 was employed for development and tuning of systems' parameters. The diarization approach is applied to the single reference channel given by NIST, SDM condition, or to the enhanced signal from all available microphones, that is the MDM condition without TDOA feature stream. Complexity selection algorithm with R_{CC} value fixed at 7 was also employed in both of the systems. The speech activity detection was the same as the presented in the RT07s evaluation analyzed in section 5.4.2.
For the case of the initialization based on TDOA analysis, a set of parameters was to be tuned in order to compute "source of speech" segmentation. Following, a brief setup of the parameters employed is given as well as the values and percentages they were fixed for the experiments reported:

- The size of the TDOA analysis window is set to 50 frames (at a rate of 250 ms. per frame)
- The percentage of TDOA considered to compute the geometric distance is set 60% of the TDOA components of that track.
- In order to create a new track and during non-speech, the minimum percentage of TDOAs which have to remain close together into the analysis window is fixed to 50%.

account speaker overlaps in figure 5.38 does not match SER% without overlap. Overlap segments contribute to the total evaluation time. Nonetheless, all σ values computed in this PhD thesis proposal equally weight each recording.

	SDM				MDM			
NAME SHOW	% SER		% DER		% SER		% DER	
	TDOA	UNI	TDOA	UNI	TDOA	UNI	TDOA	UNI
CMU_20050912-0900	15.9	30.2	19.4	36.2	11.6	11.4	22.6	25.7
CMU_20050914-0900	14.5	28.2	18.4	30.3	10.3	12.5	21.3	24.5
EDI_20050216-1051	15.7	32.4	20.1	35.3	11.5	13.2	23.6	25.1
EDI_20050218-0900	14.6	27.9	18.8	30.1	10.6	12.1	20.3	24.1
NIST_20051024-0930	4.1	6.9	9.2	13.2	3.7	6.5	7.9	10.6
NIST_20051102-1323	3.4	5.2	8.7	12.1	2.8	6.1	6.5	8.8
TNO_20041103-1130	15.8	33.8	21.7	36.1	11.7	13.6	24.3	25.1
VT_20050623-1400	4.5	6.8	9.4	12.7	4.1	5.9	8.3	10.2
VT_20051027-1400	16.5	34.1	22.4	36.7	12.2	14.1	24.7	26.2
ALL	16.7	19.8	**22.75**	26.55	9.5	12.3	**19.22**	23.50

Table 5.12: *Diarization Error Rate (DER) per cent and Speaker Error (SER) per cent in the development RT06s dataset for the initialization based on clustering of TDOAs features – TDOA column – and for the classical uniform initialization – UNI column.*

- Once a track is created, the percentage of TDOAs which have to remain near is also fixed to 50% of the analysis window, in this case and during such a condition is fulfilled, the associated track is considered as "alive".
- The maximum time in silence or without TDOA correspondence to previously created tracks is 10% of the analysis window. In this case, the track is discarded for further processing and the temporal segment associated is considered as over.
- The minimum duration output of a detected speech segment is 3 seconds.
- The threshold for the SRP-PHAT value is set to 0.9.

The figure 5.39 depicts the DER% evolution of a recording for both initialization methods and channel conditions. It is worth to note that initial DER% observed at the first iteration is directly related to the employed initialization method. Furthermore, as can be seen in the figure, the lowest DER is obtained in the TDOA initialization case and such a improvement remains constant during next iterations. That situation clearly illustrates the importance of the selected initialization method prior AHC. To take an obvious example, it inspires research in clustering algorithms which aims to improve cluster "purity" in any stage of the AHC clustering [Anguera *et al.*, 2006b; Bozonnet *et al.*, 2010a; Nwe *et al.*, 2012].
The Tables 5.12 and 5.13 report the SER and DER per cent results obtained by the TDOA initialization in both conditions in the RT06s and RT07s datasets. Overall, the use of the TDOA initialization reduces the global

	SDM				MDM			
NAME SHOW	% SER		% DER		% SER		% DER	
	TDOA	UNI	TDOA	UNI	TDOA	UNI	TDOA	UNI
CMU_20061115-1030	5.5	24.9	17.20	38.26	7.3	9.4	18.92	21.04
CMU_20061115-1530	1.3	13.7	8.37	21.27	6.3	9.5	13.42	16.61
EDI_20061113-1500	11.9	24.6	23.29	46.63	19.5	29.8	30.85	41.15
EDI_20061114-1500	13.6	12.2	19.83	29.02	9.9	11.0	16.10	17.26
NIST_20051104-1515	2.0	3.9	7.45	8.31	5.7	1.2	11.12	6.63
NIST_20060216-1347	29.8	2.5	34.28	6.90	7.1	1.8	11.56	6.33
VT_20050408-1500	12.4	6.6	18.21	12.36	1.1	7.1	6.92	12.92
VT_20050425-1000	21.1	7.1	28.97	15.44	10.5	3.7	18.39	11.64
ALL	12.3	11.8	**19.73**	21.99	8.2	9.0	**15.70**	16.50

Table 5.13: *Diarization Error Rate (DER) per cent and Speaker Error (SER) per cent in the evaluation RT07s dataset for the initialization based on clustering of TDOAs features – TDOA column – and for the classical uniform initialization – UNI column.*

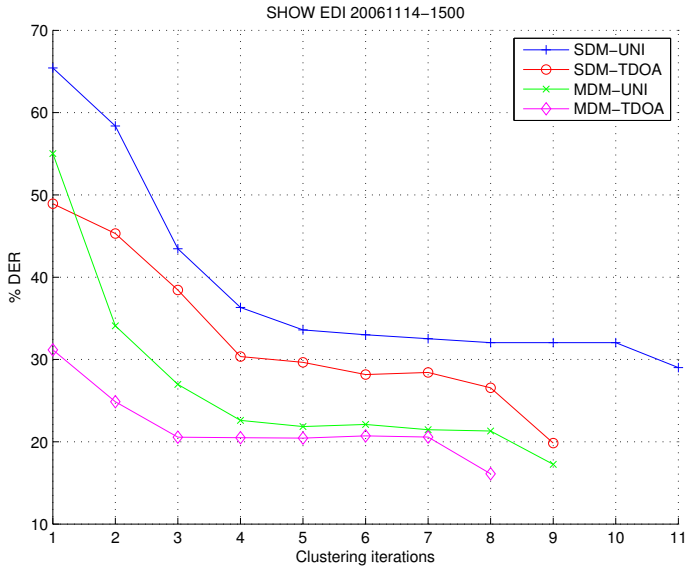


Figure 5.39: *DER % improvement per iteration for the recording EDI_20061114-1500. The different curves reports the improvement by applying the TDOA initialization on both Single Distant (SDM) and Multiple Distant Microphone (MDM) conditions.*

DER as well as the SER error [5].

The lowest DER 15.70% is obtained by the MDM system with the TDOA initialization (MDM-TDOA) improving a 4% relative with respect to the uniform initialization. In the SDM system, the TDOA initialization also performs better with a 10% of relative improvement. However, the partial results per recording show a high variability of the DER and some results show an improvement in the SDM system but not in the MDM, and viceversa. The reduction of the SER also shows the benefit of the initial TDOA clustering. The initial estimations of the speaker models are more accurate which results in a SER reduction in the most of the recordings.

In conclusion, the initial clustering provided by means the analysis of TDOA dynamics supplies "pure" segments in which speech belongs to a unique speaker avoiding, therefore, overlapped speech.

Overlapped Speech Detection

In this section we present overlap detection experiments for the multi-site scenario conducted on the NIST RT evaluation data and compare different overlap detection setups in terms of their effect on diarization improvement. More a detailed discussion is given in [Zelenák *et al.*, 2011] in where we are also discussing the behavior of overlap labeling and exclusion in relation to changing overlap detection properties in a wide dataset, the AMI meeting corpus [Mccowan *et al.*, 2005]. For the experiments reported, the RT '05–'07 data were employed for training of the overlap detection system and the RT '09 corpus for testing.

The Multiple Distant Microphone system was employed in overlap detection experiments. The output HMM probabilities were weighted in the ratio 0.80 and 0.20, for spectral and spatial feature stream, respectively. Performance is measured with Recall—the ratio between true detected and reference overlap time, Precision—the ratio between true and all detected overlap time, and with Error—the sum of missed and false overlap time divided by the reference overlap time. Note that these metrics are very strongly influenced by the overlap insertion penalty, since this penalizing parameter controls the number of overlap segments the system will hypothesize. .

Overlap detection experiments were performed for different feature setups including spectral-only system (*Scpt*) and some combinations of spectral and the spatial features, i.e., coherence (*Spat C*), dispersion (*Spat D*) and delta TDOA (*Spat dT*) and combinations of them. The detection performance on NIST RT 2009 site recordings for several of these feature setups is given in the figure 5.40. We can see that combined setup (*Spct+Spat CdT*), spectral with coherence and delta TDOA, outperform the spectral-only *Spct* in error and precision in all penalty regions. Setup *Spct+Spat CDdT* performs better than *Spct* in the lower penalty regions. In general, the best system is (*Spct+Spat CdT*) achieving lowest error and corresponding precision and recall of 95%, 58%, 34%, respectively, at OIP of -50. The possible reason for the worse performance of feature setups involving the cross-correlation dispersion ratio is the fact that this parameter may be closely related to the spatial distribution of microphones in a room. And the generally worse performance on multi-site data indicates, that the PCA is probably a too simple technique to compensate for the variability of this scenario.

[5] The speaker error time (SER) is responsible for the speaker identification error time. DER not only includes the SER error but it also takes into account the missed speaker time error and the false alarm speaker time error, see section 2.3.3.

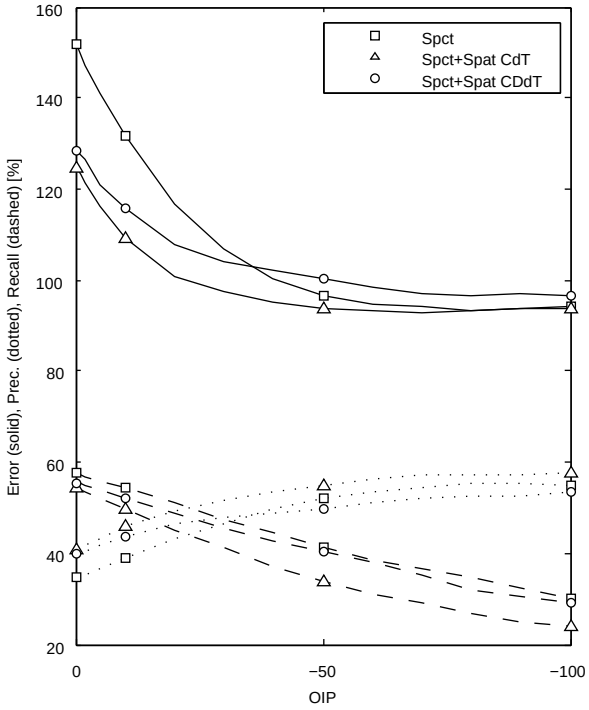


Figure 5.40: *Overlap detection performance for NIST RT '09 data. Feature setups are as follows, spectral features only (*Spct*), combination of spectral features with spatial coherence and delta TDOA (*Spct+Spat CdT*), and finally spectral features with all three spatial features (*Spct+Spat CDdT*). Error is delineated with solid line, precision with dotted line and recall with dashed line.*

The results reported in [Zelenák *et al.*, 2011] show a great dependence of the overlap detection algorithm w.r.t. the room setup and the differences among sites which degrade the detection performance compared to the single site adapted system. DER error around 95% in NIST RT database, which is a multi-site database, is coherent with those obtained in the AMI meeting corpus. Despite the high percentage of misses overlap segments and false alarms introduced by the system, the output overlap regions detected could still be used to apply exclusion and labeling techniques in order to improve diarization error rate.

The complement of the overlap detection error tells us how much the diarization can possibly gain with labeling using a particular overlap hypothesis, since all of the overlap false positives will be propagated to the DER, but only a perfect labeling would transform all true positives into a reduction of missed speaker time. Sufficiently high precision is also important for obtaining good results. Overlap hypotheses, which were produced for development recordings for several OIP values with the baseline overlap detection system, were subsequently applied in diarization system for assigning second labels.

In addition to labeling technique, segment exclusion was also considered aiming to reduce DER results. In the overlap exclusion approach, overlaps are discarded from training, hoping to achieve purer cluster models and thus a more precise segmentation. The labeling technique allows to assign two speaker labels in segments with simultaneous speech. In the latter case, the overlap hypothesis needs to be sufficiently precise, since all of

System	MS	FA	SPKE	#Spks. (Det/Miss/False)	DER (with collar)	DER
BASELINE, TDOA ($w = 0.20$)	15.1	0.1	17.4	29-9-3	19.6	**32.5**
Ovlp Excl + Labl. (Spct system)	11.1	3.3	16.1	30-8-0	17.5	**30.6**
Ovlp Excl + Labl. (Spct+Spat CDdT)	11.3	3.4	15.5	31-7-0	17.5	**30.2**
Ovlp Excl + Labl. (Spct+Spat CdT)	11.9	2.4	19	27-11-0	21.4	**33.4**

Table 5.14: *Improved speaker diarization with labeling of simultaneous speech segments on multi-site data, missed speaker-time error (MS), false alarm error (FA), speaker error (SPKE), DER and relative improvements over the new baseline (in %). In addition the number of correctly detected speaker, speaker misses and false speakers are also reported in column (#Spks.)*

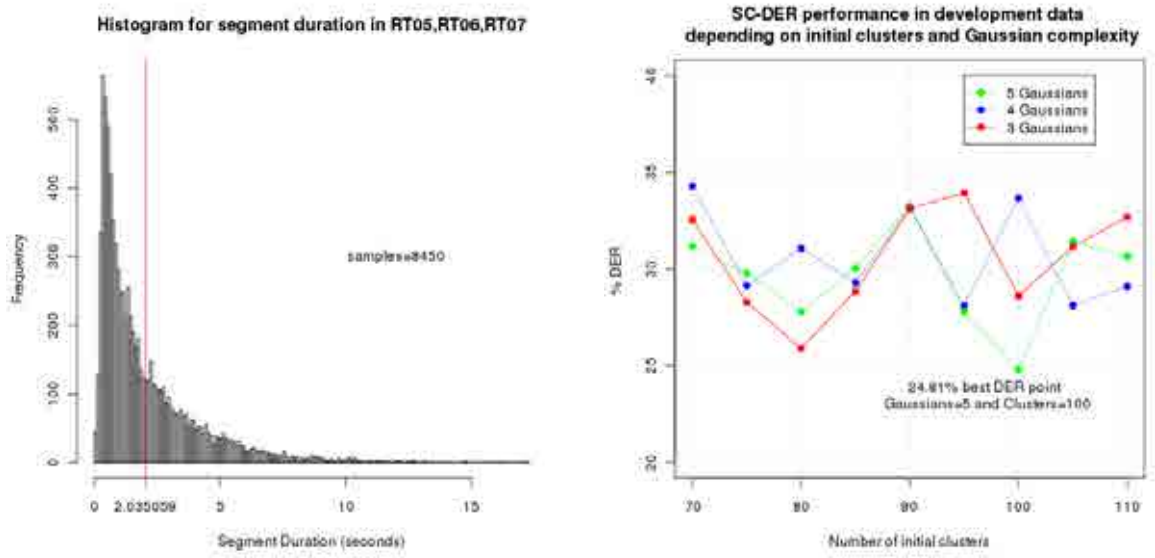
the falsely detected overlaps will contribute to diarization error and only a perfect selection of speaker labels would recover the missed overlapping speaker time. In practice, it is useful to have one overlap hypothesis for overlap exclusion and another for overlap labeling.
In a series of preliminary experiments, we spent some effort to obtain results on the NIST RT '09 conference meetings. We selected the overlap hypotheses presented in the figure 5.40 for the OIPs values: no penalization for overlap exclusion and -100 for labeling, respectively. The baseline diarization performance with the improved system utilizing beamforming and TDOAs is 32.5%. The application of the overlap handling techniques reduced the error to 30.6% for the *Spct* overlap detection setup, and to 30.2% for the *Spct+Spat CdT* setup. Again, these DER results were computed with and without any scoring collar. When the standard NIST RT evaluation collar of 0.25 s is used, the corresponding error reduction is from 19.6% to 17.5% DER for the latter of the two overlap setups.

5.4.5 Diarization based on Spectral Clustering

This subsection is devoted to provide results on RT datasets about the speaker diarization approach based upon spectral clustering technique applied on Gaussian superverctor space, see section 5.3.

Tuning System Parameters

Some parameters were tuned through a set of experiments on the development data, such as: The minimum duration turn per speaker, the initial number of segments, its GMM model complexity and finally the threshold Θ as maximum eigengap.
In the figure 5.41 (a) we depict the histogram for the segment duration in NIST RT data for the evaluations in 2005, 2006 and 2007. It takes into account any speaker segment in the evaluation time, that is, all consecutive speech from the same speaker without silences greater than 0.5 seconds. Speaker overlapped segments are also considered to draw the picture yielding to a total of 8450 samples. As we can see at the red line in the histogram, the mean duration of the segments is around 2 seconds. The minimum duration constrain for HMM/Viterbi alignment is set to such value in both SC and HMM+BIC implementations.



(a) (b)

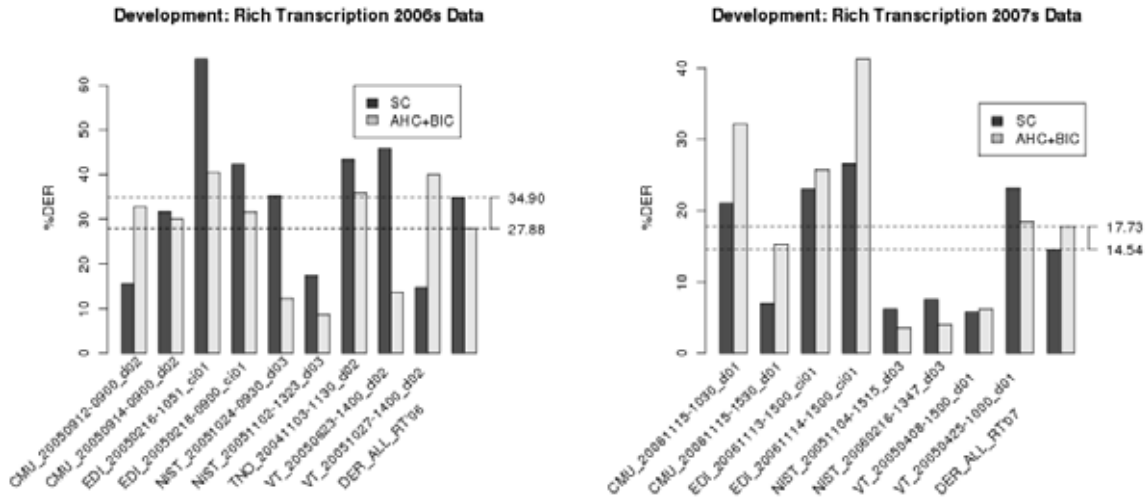
Figure 5.41: *(a) Histogram for segment durations in RT'05, RT'06 and RT'07 data. The tick marked as the red vertical line stands for the mean duration of a speaker segment. (b) Spectral clustering performance in terms of % DER depending on the initial number of clusters and the Gaussian complexity for building the GSV vectors.*

The initial number of segments and the number of initial Gaussians per segment has been also tuned using the development data sets. The figure 5.41 (b) depicts the impact on diarization error rate for the spectral clustering algorithm for different GMM model complexities: 3,4 and 5 Gaussians respectively; and for a number of initial segments ranging from 70 to 110 segments. The DER curves are obtained on the development data RT'06 and RT'07. The lowest DER is reached by using 100 initial clusters and employing GMM models composed by 5 Gaussians. These values are selected in the SC approach applied to the RT'09 evaluation data. Finally the threshold Θ, which is used to select the number of clusters, is also tuned based on %DER performance in development data. Thus the first maximum γ_k eigengap is fixed to 0.001.

Comparison with Agglomerative Clustering

The figures 5.42 (a) and (b) display the results per each show obtained on RT'06 and RT'07 conference data respectively. In both data sets, the DER errors produced by the SC-based implementation are only slightly worse than those obtained by the AHC+BIC approach. In general, AHC system obtains a better performance for both development and evaluation data sets. Nevertheless and depending on the development subset, SC outperforms the results obtained by classical AHC+BIC.

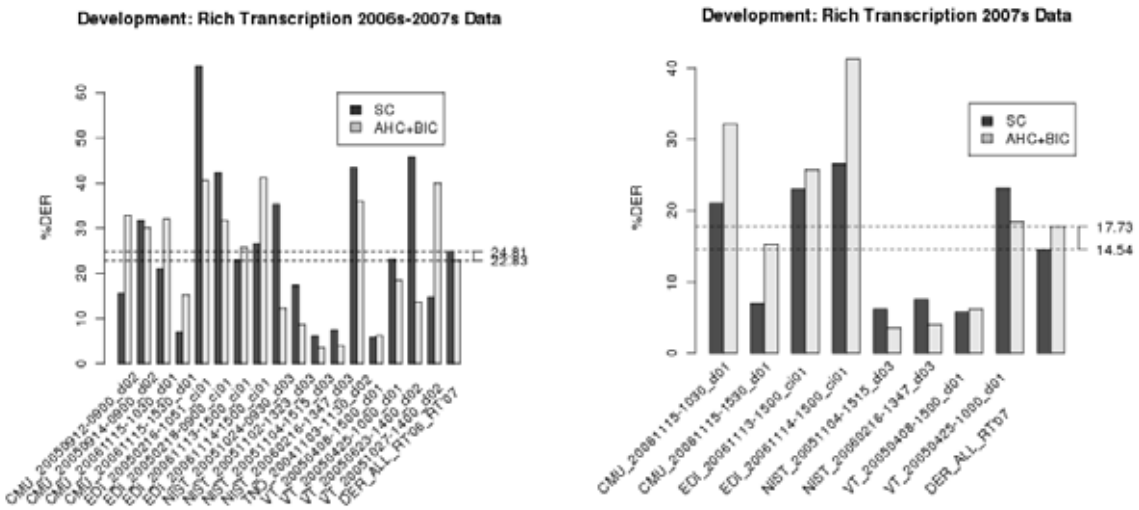
As we can see in the RT'07 data, SC obtains a 14.54% DER outperforming the AHC+BIC system with a 17.73% DER. Nonetheless, the same does not happen in RT'06 data in which SC performance fall off



(a) (b)

Figure 5.42: *Development results on (a) Rich Transcription 2006s and (b) 2007s evaluation data.*

compared to the AHC approach. In overall, in the figure 5.43 (a) we report the DER per show and the total error computed on the development data in which AHC+BIC obtains slightly lower DER results.
Finally, the figure 5.43 (b) shows the generalization of the results to unseen data in the evaluation data set. As in the case of development experiments, the AHC+BIC approach outperforms slightly the SC-based one, 25.19% compared to 27.52% DER.
Table 5.15 summarizes the results performed by both approaches on the different data sets. DER error rates and the associated standard deviation (σ) per set are also reported. It is worth to mention the lowest deviation (σ) observed in the SC results compared to the AHC approach. The SC implementation seems to perform more robustly across different shows than AHC does, specifically in RT'09 and RT'07 data sets. In addition and aiming to verify that the SC clustering provides a significant reduction in terms of complexity, we report in Table 5.15 computational relative time on the different RT evaluation data sets for AHC+BIC and SC approaches. Feature extraction processing is common for both methods and it was not taken into account for measuring time consumption. Processes were run on a Intel(R) Xeon(R) CPU E5540 2.53GHz machine. Experiments conclude that SC based clustering runs around 3 times faster than the AHC+BIC system.
The main advantage of spectral clustering is that it does not build any statistical metric for deciding if two clusters should be merged. This avoids explicit BIC or KL computation at each merging step, by employing a Euclidean distance among super vector representation of clusters, thus significantly reduces the complexity of the clustering algorithm. Experiments are performed on RT'06, RT'07 and RT'09 conference evaluation data and results are provided in terms of diarization error rate and using an oracle speech detector.



(a) (b)

Figure 5.43: *Development results on (a) Rich Transcription 2006s and 2007s evaluation data and (b) evaluation results on RT09s data.*

To summarize, the spectral clustering based algorithm along with Viterbi realignment is found to achieve DER results slightly worse than the conventional AHC+BIC system but with reduced computation. Results presented also display a great variance among different shows as well as between evaluation data sets. It may be due to data characteristics, e.g., number of speakers involved, room setups, SNR levels and speech overlap segments. For instance, the worst performance is reported in RT'06 data which exhibits both lower SNR and higher overlap time than the other databases used, see table C.1.

Finally, this work on spectral clustering theory is based on a series of assumptions that will be further investigated in future works. For instance, the similarity matrix is built based on Euclidean distances among Gaussian super vector representation of clusters obtained without MAP adaptation of an universal background model. In addition, we employ a full connected graph weighted by a scalar parameter for building the affinity matrix. These steps can be improved by means a more robust cluster representation and a best adapted metric distance among them, or by a most suitable scalar parameter to improve the merging step. Furthermore, we used an initial uniform segmentation in blocks of fixed duration. In this case, the system may be improved through the use of a speaker change detection algorithms to obtain "pure" initial segments containing a single speaker. Further research comparing implementation with multiple channels and different sets of features will also be addressed in future works.

Dataset	AHC+BIC %DER / σ	SC %DER / σ	xfaster
RT06	27.88% / 12.38	34.90% / 16.98	3.02x
RT07	17.73% / 13.90	14.54% / 9.13	2.43x
RT06+RT07	22.83% / 13.51	24.81% / 16.82	2.67x
RT'09	25.19% / 15.33	27.52% / 13.19	3.24x

Table 5.15: *DER results and standard deviation (σ) per set on Rich Transcription 2006, 2007 and 2009 conference data and number of times that SC implementation is faster than classical AHC+BIC.*

5.5 Conclusions

In this chapter we have described a speaker diarization system based on agglomerative clustering performed by a HMM/GMM model applied to meeting domain data. We have described several algorithms and tuning of different parameters and we assessed its performance in the NIST RT'06-09 conference databases. Speech activity detection along with cluster initialization and speaker overlap detection has been shown as key points in the improvement of the global diarization. Due this fact, experiments related to these questions have been carefully reported.

Some novelties and original works has been also highlighted. The use of complementary cues of information to classical spectral MFCC features has been shown as an efficient strategy in order to increase the quality of the diarization. TDOA features has been successfully applied to signal enhancement, for computing output HMM probabilities or for creating the initial clustering. Furthermore, we found that spatial information can be used to perform speaker overlap detection and we proposed three new cross-correlation-based features.

Furthermore, a newly method based on spectral clustering was proposed for speaker diarization in meeting domain. The system switches classical BIC pair computation by euclidean distances in a transformed space spawned by the Gaussians means of the clusters. The experiments reported show a comparable performance with respect to the AHC-HMM diarization based on the BIC metric.

Most of the work reported in this chapter was performed within the framework of NIST Rich Transcription Evaluations 2007 and 2009. It was published in several book chapters, international conference proceedings and a journal paper.

Chapter 6

Speaker Tracking and Diarization in Broadcast News

Since some years we can observe a much easier access to digitalized spoken documents e.g. broadcast shows or meeting recordings. This situation created a growing demand for applications of human language technologies. The common goal of these technologies is to provide a complete transcription, also known as rich transcription. Transcription in this sense means not only the words uttered by the involved speakers, but also information about the audio conditions, channel conditions, acoustic events, sex of the speakers or their identities. This additional information allows us to perform effective accessing, searching and indexing of such spoken documents. Development of techniques for its extraction is a challenging task and has also woken the interest of the research community up.

Specifically, *speaker tracking* consists of segmenting and identifying target speakers in a continuous audio stream. In contrast to diarization task in which there is no prior knowledge about the speakers, it involves audio segmentation and speaker identification/verification, this latter performed in a supervised way. Usually, acoustic data from a set of target speakers are used to estimate the corresponding acoustic models. Then, input streams are segmented, and each segment is classified as belonging to one of the target speakers or as *Unknown* (unknown speaker or non-speech). This latter category (acoustically complementary to the set of target speakers) can be easily integrated in the formalism by estimating its own acoustic model.

This chapter is devoted to describe and to assess the performance of a speaker tracking and a speaker diarization algorithms in the broadcast news domain. In the section 6.1 it is described a couple of speaker tracking systems submitted to the Albayzin Evaluations, organized by the Red Temática de Tecnologías del Habla (RTTH) [Segundo, 2006]. A speaker tracking approach based on HMM/GMM-MAP adapted distributions is compared with a step to step approach, based on BIC segmentation and GMM speaker identification techniques. In the section 6.2, a speaker diarization algorithm jointly with an acoustic event detection system, are described and their joint integration results on TV broadcast news data are presented. The detection and handling

of background conditions as complementary information for speaker diarization was assessed within the framework of the Catalan government founded project Tecnoparla.

6.1 Speaker Tracking on Spanish Radio Broadcast News: RTTH Albayzin Evaluations

This section describes two speaker tracking approaches presented to the challenge organized in November 2006 by the Spanish Network on Speech Technology, Red Temática de Tecnologías del Habla (RTTH) [Segundo, 2006]. The task consisted on a speaker tracking with speaker open set, i.e., segmentation and identification in a continuous radio stream of a previously known set of target speakers. In addition, other speakers rather than the known speakers could be present as well as commercials, noises, different channel background condition and so forth. All these conditions made the evaluation a really hard challenge.

6.1.1 Albayzin Speaker Tracking Evaluation

The Albayzin Evaluations, among others tasks, aims to promote research work on speaker segmentation of multiple speakers together with identification of some of them. The evaluation was focused on tracking, segmentation and identification, of 5 different speakers which could be surrounded by other non-target speakers, music or commercials. Spanish broadcast news data from a radio show was collected to assess the performance of the different approaches submitted to the evaluation. The goal of the speaker tracking system is to find answer to the question: *who speaks when?*, discriminating among several speakers, some of them known and other being totally unknown. In conclusion, it might be seen as a speaker indexing audio task.
The evaluation database consisted of audio tracks taken from radio broadcasts in Spanish. It includes several speakers, around 400 turn changes, music, movie excerpts, commercials, overlaps, etc. Training data were available only for 5 target speakers. It was composed of 5 short utterances per speaker, 4 of them artificially distorted with noise and reverberation. The training material for each speaker had an average length of 12.8 seconds (64 seconds joining all utterances). The test corpus was composed of 20 tracks, lasting 4 minutes in average, yielding to a total of around 77 minutes of speech. Speaker and silence transcriptions were obtained manually. The silences between interventions of the same speaker and smaller than 500 ms. were not labelled. Finally, one of the testing tracks was delivered to participants prior the test evaluation and was used for developing purposes. Target speakers, unknown speakers and other classes of acoustic events were present in the recordings. Some examples of these acoustics events are: Music, background music, overlapping of speech, channel and environment changing, studio and telephonic conditions, ...
It was the first edition of the speaker tracking evaluation and just a couple of participants submitted contributions. The first submitted system was developed by the Software Technology Group of the University of the Basque Country (EHU). It performed audio segmentation and speaker identification in a completely decoupled fashion. Segmentation was done in a fully unsupervised way, by locating the most likely change points in

the acoustic signal. Then the available speaker data was used to estimate single-Gaussian acoustic models which were finally employed to classify the audio segments. Speaker identification was done by computing the score of each segment with regard to the speaker models, which were trained beforehand starting from labelled speaker data. Each segment was assigned the label of the most likely speaker or, alternatively, the label *Unknown*, if none of the speakers was likely enough.

The second submitted system, developed at the TALP Research Center of the Technical University of Catalonia (UPC), was based on the Agglomerative Hierarchical Clustering (AHC) system described in the previous chapter, i.e., on an iterative segmentation through a HMM/GMM. In contrast to the EHU's approach, this strategy allowed for a global clustering of the audio sequence, coupling with the segmentation and identification tasks at the same time. In order to adapt it to the speaker tracking task, the speaker diarization algorithm was adapted by including modeling techniques such as Maximum a Posteriori Adaptation (MAP) and some scores strategies coming from the speaker verification task.

6.1.2 Sequential Approach: XBIC-based Tracking

In the following we briefly describe the speaker tracking approach developed at EHU. Let's consider two segments of speech, X and Y, of the same length, and the corresponding sequences of spectral feature vectors, $x = x_1, \ldots, x_N$ and $y = y_1, \ldots, y_N$. Assuming that the acoustic vectors are statistically independent and that can be modeled by a multivariate Gaussian distribution, the acoustic models are defined as $\lambda_x = N(O; \mu_x, \Sigma_x)$ and $\lambda_y = N(O; \mu_y, \Sigma_y)$ and the *dissimilarity measure* between X and Y is defined as follows:

$$d(X,Y) = -\log\left(\frac{P(x|\lambda_y)P(y|\lambda_x)}{P(x|\lambda_x)P(y|\lambda_y)}\right) \tag{6.1}$$

where $P(z|\lambda) = \prod_{i=1}^{N} N(z_i; \mu, \Sigma)$ is the likelihood of the acoustic sequence z given the model λ. In other words, if X and Y are acoustically close, their respective models will be quite close too, which means that $d(X,Y) \approx 0$. On the other hand, the more X and Y differ, the greater $d(X,Y)$ will become.

The audio segmentation algorithm considers a sliding window W of N acoustic vectors and computes the likelihood of a change at the center of that window, then moves the window n vectors ahead and repeats the process until the end of the vector sequence. To compute the likelihood of change, each window is divided in two halfs, W_l and W_r, then a Gaussian distribution (with diagonal covariance matrix) is estimated for each half and finally the dissimilarity measure between W_l and W_r is computed and stored as likelihood of change. This yields a sequence of likelihoods which must be post-processed to get the hypothesized segment boundaries. This involves applying a threshold τ and forcing a minimum segment size δ. In practice, a boundary t is validated when its likelihood exceeds τ and there is no candidate boundary with greater likelihood in the interval $[t-\delta, t+\delta]$. An example of audio segmentation is depicted in the figure 6.1.

Once the segmentation is done, each segment must be given a speaker label or, alternatively, the special label *Unknown* when no speaker is likely enough. Assuming that a certain amount of training data is available for L target speakers, speaker models can be estimated beforehand. In the approach developed at EHU, speaker

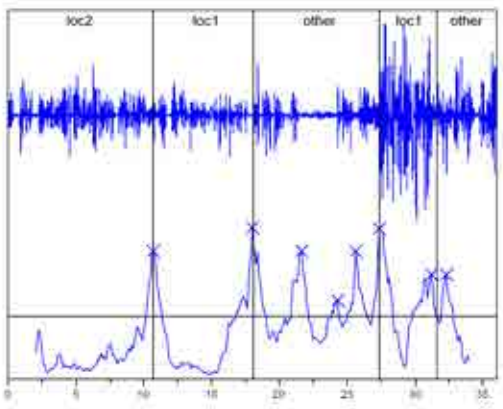


Figure 6.1: *An example of audio segmentation at EHU. Vertical lines represent actual boundaries. The local maxima marked with 'X' represent the boundaries hypothesized by the system. Figure courtesy of Professor Luis Javier Rodriguez Fuentes.*

models are multivariate single-Gaussian distributions: $\lambda_i = N(O; \mu_i, \Sigma_i)$, for $i = 1, \ldots, L$. To classify any given segment X, firstly the *segment model* is estimated (again as a single-Gaussian distribution with diagonal covariance matrix) $\lambda_X = N(O; \mu_X, \sigma_X^2)$, starting from the sequence of acoustic vectors $x = x_1, \ldots, x_N$. Note that $P(x|\lambda_X) \geq P(x|\lambda_i) \ \forall i$. The label $l(X)$ is given the value:

$$k = \arg \max_{i=1,\ldots,L} P(X|\lambda_i) \tag{6.2}$$

if $\frac{1}{N} \log \left(\frac{P(x|\lambda_k)}{P(x|\lambda_X)} \right) > \epsilon$, where ϵ is a heuristically fixed margin. Alternatively, if the likelihood ratio of the most likely speaker does not exceed ϵ, the label *Unknown* is assigned to X.

6.1.3 Integrated Approach: HMM/GMM-MAP based Tracking

The system developed at UPC to perform speaker tracking was based on the Agglomerative Hierarchical Clustering (AHC) algorithm introduced in previous chapter. It was specifically adapted to the speaker tracking task by the inclusion of speaker identification and verification techniques. Nevertheless, it still relied on the modeling and segmentation of the input speech by means of a Hidden Markov Models (HMM), in an iterative strategy of training and segmentation cycles. The stopping criterion in this case is met whenever the clustering output remains constant between two consecutive iterations.

We should consider the differences of this evaluation with respect to the RT diarization evaluation from NIST in order to understand differences among the diarization and tracking approaches. In NIST RT, there is no prior knowledge about the speakers involved in a conversation or how many of them are in the audio stream. Therefore systems developed to perform diarization in RT do not use any previous speaker data but they create speaker models from the scratch. In contrast, in the speaker tracking task we seek for specific target speakers into the audio stream. The speaker indexing task is carried out by comparing input speech to target speaker

data in a repository. Hence the HMM topology was adapted to the speaker tracking task.

The topology of the HMM was composed of 6 states, corresponding to the 5 target speakers and a General Model (GM) representing jointly unknown speakers and other acoustic events, see figure 6.3 (a). The enrollment data available from target speakers was used to train initial target speaker models, by applying the iterative Expectation-Maximization (EM) algorithm. The GM_1 model, composed of 32 mixtures, was estimated using the whole test data, once again by means Maximum Likelihood algorithm. In order to deal with mismatch conditions, another General Model (GM_0) was also EM-trained through testing data and using a *split-vanish* initialization. Afterwards, Maximum A Posteriori (MAP) mean adaptation [Bimbot *et al.*, 2004] to GM_0 was performed for each target speaker and also for the GM_1 model, see figure 6.2.

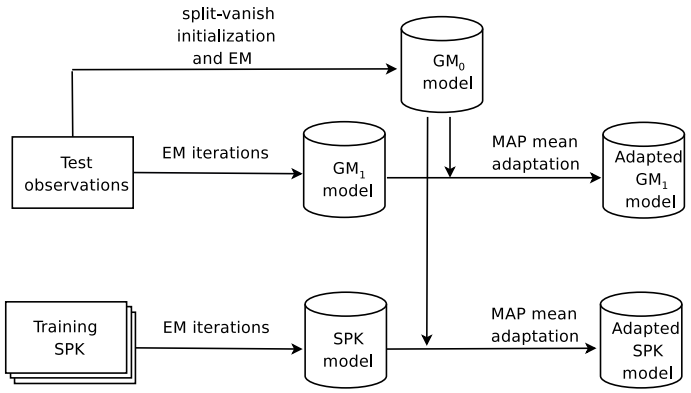
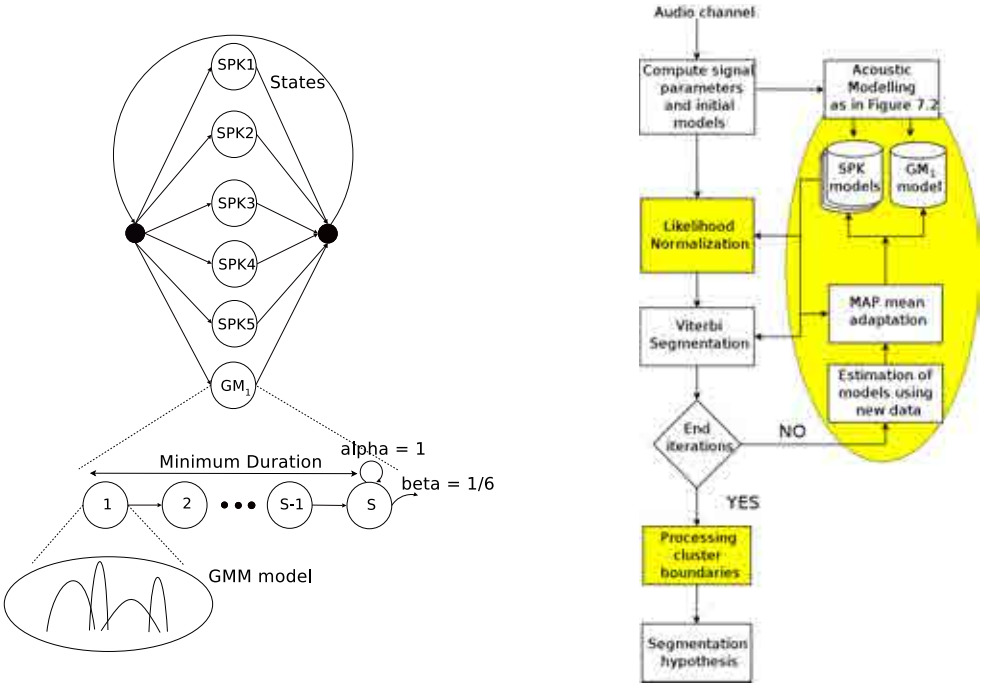


Figure 6.2: *Acoustic modeling in the HMM/GMM tracking system. The target models and the GM_1 model are MAP-mean adapted to the GM_0 model, moving them closer to the acoustics of the test data.*

It is also worth to mention that a frame pruning strategy was also developed. It aimed to avoid frames of silence or non-speech data which were not masked by the SVM-based speech detector prior to the enrollment stage. Around 5% of the frames corresponding to those frames with lowest energy were pruned, that is, discarded to enroll speaker models.

As it is represented in the figure 6.3 (a), each HMM state is composed of a sequence of S sub-states which imposes a minimum duration of the speaker turn. The initial probabilities of the states are set manually to make the classes equally likely at the beginning of the stream. The sub-states of any given state share the same probability density function, that is, the target speaker GMM model. Upon entering a state at time n, the model forces a switch to the following sub-state with probability 1.0 until the last sub-state is reached. Then, the model can cycle on that sub-state with transition probability α, or switch to the first sub-state of a different state with transition probability $\beta = 1/M$, where M stands for the number of target speakers. In this case, α and β were set to 1.0. Thus, once a segment exceeds the minimum duration, the HMM state transitions no longer influence the turn length, which is solely governed by acoustics [Anguera *et al.*, 2006d]. Note that this strategy in the transition probabilities of the HMM leads to a non-standard HMM topology, since



(a) (b)

Figure 6.3: *The UPC tracking system scheme submitted to Albayzin 2006 Evaluation. (a) An ergodic HMM models the acoustic data with 6 states, each one of them composed of S sub-states which share same GMM speaker model. (b) The algorithm employed to perform speaker tracking in the Broadcast News domain based on iterative MAP adaptation of speaker models. The main differences with respect to speaker diarization presented in previous chapter are highlighted.*

$\alpha + \beta \neq 1.0$.

During Viterbi segmentation, some speaker verification techniques were applied. The emission probabilities at each state were normalized using Log-Likelihood Ratio (LLR) [Bimbot *et al.*, 2004] at score-level. Such a normalization was computed by dividing the likelihoods performed on the target speakers by the likelihoods computed on the GM_1 model. A likelihood threshold tuned with development data was also imposed as a condition for target speakers detection. Furthermore, in order to compute the probabilities of each state, only the 50% of the Gaussians that performed highest frame score by the GM_1 model were used. This incomplete computation of the likelihoods did not affect significantly the system performance, but decreased the time factor of the whole system.

Once speaker models were computed, it followed a Viterbi decoding of the audio stream in order to obtain a speaker turn sequence, see figure 6.3 (b), that is, a clustering of the audio stream is obtained. These newly segmented speaker data was combined with the enrollment set and new speaker models were estimated. It was done through MAP mean adaptation, but using a smaller adaptation factor due to the low confidence on the automatically detected speech. The MAP adaptation performed was iterative, i.e. the adaptation of speaker

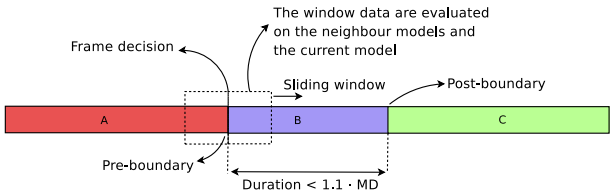


Figure 6.4: *Post-processing segment boundaries. A sliding window is applied on the shortest segments in order to split the middle segment data among the adjacent segments.*

distributions were recomputed a number of times n_{MAP}. The optimal value of n_{MAP} was fixed by experiments results obtained in the development data. This iterative training/segmentation process was repeated until no variation in the clustering output was noticed.

Next the iterative segmentation process, the resulting boundaries were post-processed. Those segments with duration smaller than 1.1 times the minimum duration (MD) were shared out among the adjacent segments. Thus the short segment was assigned to previous or next segment depending on the maximum likelihood obtained by evaluating the short segment data on the two adjacent segments, see figure 6.4. The idea behind this post-processing relies on the hypothesis that the shortest segments are usually associated to false alarms in BN data. Frequently in radio and TV broadcast news, the speaker turns are managed by the director of the show and speakers interventions follow a speak-by-turn script with no room for improvisation and avoiding speaker interruptions or overlapping in order to improve intelligibility. In contrast, in speech meeting data the spontaneity an multiparty conditions results on a high number of speaker interruptions and overlaps.

Once the iterative segmentation process is completed and the boundaries post-processed, the final hypothesis is obtained. Summarizing, different techniques were developed in order to perform speaker tracking. They make a substantial difference with respect to the system developed for speaker diarization. Such a techniques are:

- Computation of initial speaker models based on MAP adaptation.
- Loop on Viterbi segmentation: It applies the Viterbi decoding of the speaker sequence until no variation in the clustering is obtained.
- Post processing of cluster boundaries.
- Loop (n_{MAP}) on MAP adaptation, number of MAP adaptation n_{MAP} is proportional to the number of loop iterations in Viterbi imposing higher adaptation values as more Viterbi iterations has been completed.
- Frame purification, in where the 5% (fixed in development data) of highest energy frames are employed to compute the model likelihood.
- Gaussian selection to likelihood computation, by selecting 50% of the Gaussians which produces highest likelihood scores.
- Log-likelihood ratio normalization by the log-likelihood performed on the General Model (GM_1).

6.1.4 Experiments

Both EHU and UPC audio approaches rely on a common speech analysis. In both cases, the audio was analyzed employing a Hamming window, of 25 ms. in EHU approach and 30 ms. in UPC approach, at 10 ms. rate. Then, a 512-point FFT was computed and FFT-amplitudes were averaged in 24 overlapped triangular filters at EHU, 13 filters were used at UPC, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform was finally applied to the logarithm of the filter amplitudes, obtaining 13 Mel-Frequency Cepstral Coefficients (MFCC). However, the first coefficient, representing the energy, was not used to estimate the acoustic models. It was only used in the UPC system to prune low energy frames associated with silences in both enrollment and testing data. Experiments were conducted using recordings from Spanish radio broadcasts.

To measure the performance of the submitted approaches, the NIST evaluation software for speaker diarization was used (in particular, the newest version included in the Spring 2007 Rich Transcription Meeting Recognition Evaluation Plan [Fiscus and et al., 2007a]). This metric, called *Diarization Error Rate* (DER), is computed by first finding an optimal one-to-one mapping between the reference labels and the system labels, and then obtaining the error as the percentage of time that system labels after mapping are wrong, see section 2.3.3. Consider the example depicted in the figure 6.5, where not only segmentation errors but also clustering errors are illustrated. Note, for instance, that the last segment is erroneously assigned to a third speaker. After the alignment is done, the label *s01* is considered equivalent to *mm* and the label *s02* equivalent to *ft*. Finally, it is found that speakers have been erroneously identified during 10 seconds out of 25 (the shaded regions in figure 6.5) (a), which means a 40% speaker diarization error.

DER takes the system labels as if they were *blind*, but what we produce in this work are not blind but *informed* labels, i.e. *actual* speaker labels. The speaker identification error must be measured by comparing the system and reference labels on a frame-by-frame basis. Consider the example depicted in figure 6.5 (b). The system provides segment boundaries and segment labels. Labels refer to the speaker identity. If a given segment does not match any known speaker, it is assigned the label *Unknown*, which works as an additional speaker. The speaker identification error is computed as the fraction of time system labels do not match reference labels. In the example in the figure 6.5 (b) speakers are erroneously identified during 15 seconds out of 25, which means a 60% speaker identification error.

To get the speaker identification error related to speaker tracking, the NIST software was slightly modified, and two new metrics were defined: *Diarization Error Rate Modified* (DERM) and *Speaker Identification Error Rate* (SIER). DERM computes as errors those frames whose reference label r does not match the label mapped to r, $m(r)$. SIER computes as errors those frames whose reference label r does not match the system label s (before mapping). DER is an optimistic metric, since it evaluates the best segmentation in the sense of minimum DER, whereas SIER is a pure frame-by-frame speaker identification rate (the time system labels do not match reference labels divided by the total audio time). On the other hand, DERM attempts to measure the segmentation quality but forcing the correct mapping of speaker labels, becoming a hard metric, since all the segments whose reference label r does not match the mapped one $m(r)$ are taken as errors.

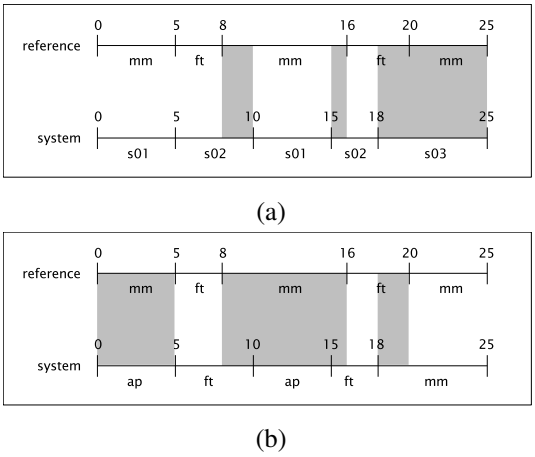


Figure 6.5: *(a) An example of speaker diarization. The system provides a sequence of segments with* blind *speaker labels. First, the system and reference segmentation are aligned. Then, among those labels assigned by the system to any given speaker, that appearing most times is taken as the system choice and considered equivalent to the reference label. Finally, the speaker identification error is computed as the fraction of time speakers are erroneously identified (shaded regions). (b) An example of speaker tracking. The system provides a sequence of segments with labels of* known *speakers. The speaker identification error is computed as the fraction of time speakers are erroneously identified (shaded regions). Figure courtesy of Professor Luis Javier Rodriguez Fuentes.*

In the table 6.1 it is displayed the official evaluation results for the different error metrics explained above. The results are displayed for both the step-by-step and the integrated approaches. Two different systems were submitted from UPC that just differ in the post-processing of the short-segment boundaries. The UPC-contrastive system does not make use of the post-processing technique in contrast to the UPC-primary submission. Furthermore, the development error rate for both UPC submission systems are also displayed in the last column of the same table. It is worth to mention that the results reported in the table 6.1 are computed without collar, see 2.3.3, i.e, the metric counts any inexactitude w.r.t the speaker references. Therefore the system's outputs should to match exactly the manual references. Such a evaluation protocol is a really challenge assessment.

SIER metric was the metric proposed after the evaluation. It compares directly the system output with the reference without any alignment or mapping, i.e, a frame-by-frame identification rate. In both approaches, the speaker identification error (SIER) is slightly higher than 17% and not statistically significant difference is noticed. This, comparable to other results reported in the literature [Dunn *et al.*, 2000], is specially relevant because: (1) speaker models are estimated from a few utterances taken from radio broadcasts, many of them (80%) intentionally distorted; and (2) parameters are tuned almost blindly, using just one track of 4 minutes including only two target speakers, (3) no collar is applied to compute the identification error, imposing the system to exactly match the reference.

Nonetheless, the results related to DER and DERM metrics are substantially different for both approaches. In

	Evaluation			Development
Metric → / System ↓	DER (NIST RT)	DERM	SIER	DER (NIST RT)
EHU	15.20 %	19.50 %	17.25 %	–
UPC-primary	17.44 %	17.44 %	17.44 %	6.54 %
UPC-contrastive	18.03 %	18.03 %	18.03 %	6.82 %

Table 6.1: *Official results reported by EHU and UPC in the 2006 Speaker Tracking Challenge Albayzin. The development results are displayed in the first column. The difference between the two UPC submission systems is just the final post-processing of short segments boundaries.*

the case of the NIST DER metric, the BIC-based system outperforms the results of the AHC approach whereas in the case of the modified DERM metric we see the opposed result. Such a variance in the error rate of the step-by-step approach is likely due to a high confusion among speakers in this approach than in the AHC system. The NIST DER metric corrects the error by aligning the output of the system into the best possible segmentation. However, it does not occur whether DERM metric is applied since DERM does not align the system hypothesis. Note that, independently the error metric employed, the corresponding error rates of the integrated approach remain constant. It illustrates the differences among the segmentation based on both BIC criterion and local audio information (depending on the analysis window) and the HMM segmentation which employs the whole audio recording to estimate the speaker segmentation.

Finally, the post-processing technique which was included in the UPC-contrastive submission, obtained promising results during development experiments, see last column in the table 6.1. It was confirmed by the official evaluation results in where the recognition was improved by 0.5% absolute.

Next Albayzin Evaluation some experiments were performed in order to compare the improvement on the speaker tracking error by the inclusion of the different speaker identification techniques. A wide development set was employed aiming to avoid over-training of the algorithm. The original Albayzin evaluation data was randomly splitted into two sets of the same size. A development set around 40 minutes of duration was used for tuning of parameters and a testing set of the same duration was employed to assess the AHC/GMM-UBM algorithm. Table 6.2 shows the impact of the different techniques on the degradation of the DER results. The first row corresponds to the UPC-primary system submitted to Albayzin evaluation. Several techniques are eliminated the more we get lower at the column. As can be seen in the table 6.2, the iterative Viterbi segmentation and the LLR normalization techniques represent most of the improvement from the UPC-primary system in both datasets. In the first case, it reaches 11% of absolute improvement in development and 7% in the "blind" data.

The figure 6.6 depicts the DERM error evolution in both development and test sets. The DERM is depicted per each training/segmentation iteration showing that most of the recognition error is mainly corrected at the two initial iterations.

The improvement by LLR normalization is lowest but still significant: 3.5% and 1.6% respectively. Other techniques as Gaussian selection and the iterative MAP-mean adaptation also obtained a enhanced performance

	Development	Test
System	DERM % (diff %)	DERM % (diff %)
UPC-primary	13.36%	18.13%
Loop Viterbi Segmentation	24.55% (+11.19%)	25.22% (+7.09%)
Post Processing	24.55% (+0%)	25.22% (+0%)
Loop MAP	22.50% (−2.05%)	27.38% (+2.16%)
Frame Purification	21.67% (−0.83%)	25.89% (−1.49%)
Gaussian Selection	21.67% (+0%)	26.01% (+0.12%)
LLR Norm	25.20% (+3.53%)	27.60% (+1.59%)

Table 6.2: *Diarization Error Rates (DERM modified) of the complete system. The different techniques are eliminated from the UPC-primary system as we lowered in the column. Looping in the Viterbi segmentation and the LLR normalization are the techniques that showed higher improvement in both datasets.*

but not at the same time in both development and test sets. It is worth to mention that Gaussian selection and frame purification reduced the time consumption of the tracking algorithm. In the case of Gaussian selection, it occurs as well as without any drop of performance despite of it uses just 50% of the Gaussians to compute the likelihoods.

Finally, in contrast to previous Albayzin evaluation results, in this post evaluation experiments the final processing of segment boundaries did not seem to aid the diarization system.

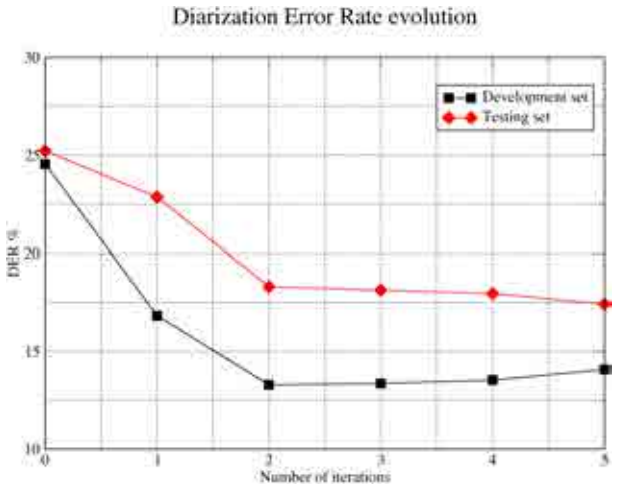


Figure 6.6: *DERM evolution with respect to the number of training/segmentation iterations. The depicted system corresponds to the UPC-primary whose results are also reported in table 6.2. The curves depicte the DER reduction achieved by the iterative segmentation and verification strategies at each iteration in both datasets.*

6.1.5 Conclusions

No significant differences in performance were found between the two approaches presented in this chapter. However, the methodologies are quite different. The former makes use of two independent and decoupled modules for audio segmentation and speaker identification. Moreover, all the acoustic models are single Gaussians, which can hardly model the spectral variability of speakers and segments, but at the same time provide robust estimates, even when not many training data are available, and allow real-time operation of the speaker tracking system. System time consumption was reported by EHU, claiming that in the experiments reported speaker tracking was completed in less than 0.1 real-time. The surprisingly good performance of this approach can be due to the low number of target speakers in the proposed evaluation. In other words, single Gaussians might be the most suitable models, in terms of accuracy and robustness of parameter estimates, when dealing with only 6 categories, the five target speakers and the *Unknown* category, and with few available training material.

The latter integrates audio segmentation and speaker identification in an iterative segmentation/identification loop in where the speaker acoustic models were continuously adapted maximum a posteriori. Additionally, it relies on a Gaussian of mixtures model to model speaker voices, each of them composed of 32 active components. Both submitted systems yielded around 17% speaker identification error time in the speaker tracking evaluation Albayzin. It is a promising performance due to the fact that the evaluation testing data were excerpts from radio broadcast recordings without any manual processing. Furthermore, training data was artificially corrupted with noise aiming to simulate mismatch conditions among training and testing data which together with error metric proposed drove to a hard evaluation conditions.

The integrated approach, in spite of their higher computational cost compared to the step-by-step system, should also show a good performance in a wide range of situations. The use of MAP adapted speaker models and an specific model for the *Unknown* class make suitable the algorithm to apply in different conditions data without modification while avoiding, e.g., the manual tuning of a verification threshold as ϵ at the step-by-step approach. Nonetheless, it results in a poor estimate of model parameters in the case of the Albayzin Evaluation which is composed of a small set of samples, thus limiting theoretically the potential performance of the integrated approach. The step-by-step approach should be a good choice to implement in the case of on-line processing requirements due its implementation simplicity and good performance. However, real-time tracking could be faced by applying minor modifications in the iterative system. Among others, the initial steps forward the adaptation of the speaker tracking system to real time processing conditions should be: Complexity reduction of the GMM models, adjusting of the number of training/segmentation iterations and the application of an analysis window instead of the whole recording in order to compute the Viterbi path.

6.2 Speaker Diarization on Catalan Television Broadcast News. Tecnoparla Project

Among the human language processing techniques is *audio segmentation*. It aims to segment the audio signal according to the predefined acoustic classes (music, speech,silence etc.), and typically, it is requested when processing broadcast news. Some recent approaches [Neto *et al.*, 2008] took advantage of the information about the structure of shows such as typical sounds which indicated audio changes. Several works [Scheirer and Slaney, 1997; Kim *et al.*, 2007] have investigated various features and feature sets for the problem of music/speech discrimination. Audio segmentation has been employed for instance in the context of robust speech recognition [Choi *et al.*, 2007]. Unlike previous works [Neto *et al.*, 2008], our aim is to provide not only a working audio segmentation system but also a more general solution that can fit in other broadcast news scenarios. Another technique applicable in the context of human language processing is *speaker diarization*. It can be seen as a particular task of the audio segmentation issue [Kemp and M. Schmidt, 2000] and consists in segmenting and labeling an unknown set of speakers in a continuous audio stream. Speaker diarization is usually described as the task of deciding *who spoke when* and it involves a large variety of applications.
In this chapter we seek to employ audio segmentation jointly with speaker diarization. We consider detection of audio segments relevant to speaker diarization task such as speech, music, speech over music, telephone speech, telephone speech over music, and silence. The main objective is to evaluate and compare the performance of our diarization system exploiting the audio segmentation information in different ways. On the one hand, it can be used beforehand to extract speech or more condition-specific segments which are then fed to the diarization system. On the other hand, the audio segmentation hypothesis can be used with the diarization labeling to perform some kind of time masking. Furthermore, it will be show that information about channel and background conditions might improve the diarization performance by handling independently such conditions.

6.2.1 Tecnoparla Corpora

The corpus consists of Àgora TV shows. The Àgora show airs on Monday night in Television of Catalunya channel. It is a debate show highly moderated and with a high variation in topics and invited speakers. The total audio time of the database is about 42 hours divided in 34 shows (approximately 1 hour and 20 minute each show), each one corresponding to an airing day. Each show has been split in two halves to delete the commercials present during the airing. Thus, 68 files corresponding to 34 shows are the total amount of audio of this database. The definition of acoustic conditions depends on three variables. First, *mode* denotes the type of speech. It can be planned, spontaneous or void if it is not specified. Second, *background* describes which conditions co-exist with the signal. We can distinguish between music, speech and void if nothing is specified. Final variable is the *channel*. All signals are broadcasted from the studio, but some of them can be band limited in the source (e.g. telephone speech), although they can be overlapped with whole band background conditions (e.g. telephone speech with music background). Two channels are defined for the shows: studio and telephone.

Transcriptions

The manual transcriptions available for the database are mainly intended for speech recognition experiments. This means that there is nearly no information about silences. Only long silences at the beginning or the end of the file (half a show) are transcribed. Long silences inside the body of the transcription are not transcribed directly or indirectly. Furthermore, the temporal positions of various acoustic events (e.g. *laugh*, *cough*, etc.) that appear in the shows are often referred to the words being affected instead of exact timestamps. Another characteristic is that speaker segmentation has only one boundary. This means that one speaker is supposed to start when the previous speaker ends, even if there is a long silence, or music, between them.
In this work, for audio segmentation, we considered detection of next classes: *speech, music, speech over music, telephone speech, telephone speech over music* and *silence*. These classes can be seen as a combination of a background condition (silence or music) with a foreground condition (speech, speech over telephonic channel or silence).
In general the transcription cannot be trusted for the task of audio segmentation or speaker diarization. For this reason, we adopted an algorithm of transcription correction to introduce more accurate temporal marks for the audio classes considered. In order to refine the transcriptions, a background detector was designed. It enabled us to refine the *speech* class by detecting either (*music* or *silence*, in our case) in the background. The following design constraints were introduced:

- Conservative behavior. Missing the detection of *silence* or *music* in the presence of *speech* is preferable to false alarms which cause the error that is not recoverable by posterior processes.
- Minimum duration of a background to detect is considered 500 ms.

The system which modified initial transcriptions was based on GMM. For each acoustic condition considered, i.e. speech over background music and speech, a 4 Gaussian GMM was trained on the log-short-time-energy computed in a 30 ms. frame at 10 ms. rate using the EM algorithm. After the model was built the Gaussian that corresponds to the lowest mean was used to generate a Gaussian probability density function for the background. The rest of the Gaussians were used to generate a GMM model for the foreground speech as it is represented in the figure 6.8. We assigned one Gaussian to the background condition as proposed in [Anguera *et al.*, 2006a], but we observed empirically that three Gaussians modeled foreground data more accurately.
The algorithm was applied separately to each of the considered acoustic conditions. Each frame was then classified according to maximum likelihood criterion either to one model or the other. Finally, a smoothing based in median filter is applied to meet the 500 ms. minimum duration restriction.

6.2.2 Audio Segmentation

Audio segmentation, sometimes referred to as acoustic change detection, consists of exploring an audio file to find acoustically homogeneous segments, or, in other words, detecting any change of speaker, background or channel conditions. It is a pattern recognition problem, since it strives to find the most likely categorization

of a sequence of acoustic observations, yielding the boundaries between segments as a by-product. Audio segmentation becomes useful as a preprocessing step in order to transcribe the speech content in broadcast news and meetings, because regions of different nature can be handled in a different way.
The SVM-based system that was previously used for detection and classification of acoustic events in meeting room scenario [Temko and Nadeu, 2006; Temko *et al.*, 2008] was employed in this work for a general task of audio segmentation. The sound signal was down-sampled to 16 kHz, and framed (frame length/shift was 30/10ms, a Hamming window is used). For each frame, a set of frame-level features was extracted. In [Temko and Nadeu, 2006] the best performance was obtained with a combination of features used in automatic speech recognition and other perceptual features. Hence in this work the set of features consisted of the concatenation of two types of features:

- 16 frequency-filtered (FF) log filter-bank energies, along with first and second time derivatives.
- The subset of features composed by: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth.

The first type of features has been already used for classification of acoustic events in a previous work [Temko and Nadeu, 2006], where the frequency-filtered log filter-bank energies [Nadeu *et al.*, 2001] shown a better performance than the usual mel-frequency cepstral coefficients which are used in speech recognition. The contribution of each feature from the second type is not so well-established and the spectral features included in the second type were chosen from a much larger pool of features after taking into account their individual importance and degree of interaction [Temko *et al.*, 2008].
In total, a vector of 60 components was built to represent each frame. In both training and testing processes, the mean and the standard deviation of the features were computed over all frames in a 0.5 second window thus forming one feature vector of 120 elements. The resulting vectors of statistical parameters were fed to the SVM classifier and the decision was made each 100 ms. as it is shown in the figure 6.7. The post-processing based on median filter was applied to eliminate uncertain decisions and minimum duration threshold was also imposed to avoid too frequent changes of audio classes. Concerning SVM training, the standard operations were undertaken: anisotropic data normalization with the normalization templates that were applied afterwards to test data, and 5-fold cross-validation to obtain optimal values of both the Gaussian kernel parameter and the C parameter.
Due to the huge amount of data, the dataset reduction technique developed and described in [Temko *et al.*, 2007] were used in order to decrease the number of training vectors and make model searching and final training feasible. Briefly, the process consists in dividing all the data into chunks of 1000 vectors per chunk and training a proximal SVM on each chunk performing 5-fold cross-validation (CV) to obtain the optimal kernel parameter and the C parameter that controls the training error. To select a pre-defined number of chunks an appropriate threshold was applied to the CV accuracies of all chunks. With that approach the data that corresponds to the most/least separable data in chunks can be chosen depending on the difficulty of the discrimination between classes.

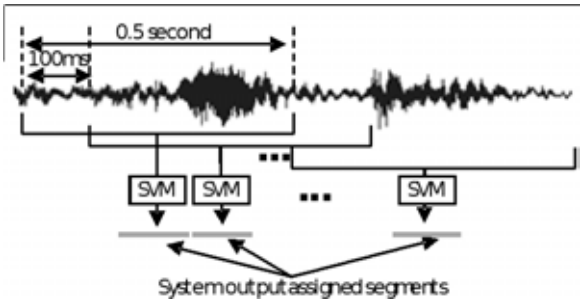


Figure 6.7: *Window analysis strategy performed by the SVM-based audio segmentation module.*

After amount of training data was decreased an SVM classifier was trained for each pair of classes. The 1 vs. 1 directed acyclic graphs (DAG) multi-class strategy [Platt *et al.*, 2000] was chosen to classify among 6 audio classes. It is worth to note that, in terms of training data, speech class highly dominates over the other classes considered. Usually, a discriminative classifier tends to give more priority to the prevailing class. In our case it implied that speech decisions will be produced more frequently than others. However, as it is known that a speaker diarization system suffers more from speech deletions than from speech insertions as unlike the formers the latter can still be corrected further on inside the speaker diarization system, no way to deal with dataset imbalance problem was considered in this work. As a result, the performance of the audio segmentation system for every class was not expected to be uniform, but that system still was preferable for the speaker diarization as a final application.

6.2.3 Speaker Diarization

The speaker diarization algorithm applied was mainly the same than the system described for meeting domain data. In contrast, the speech data was firstly pre-processed with a high-pass filter at minimum frequency of 100 Hz and analyzed in windows of 30 milliseconds at intervals of 10 milliseconds, i.e., at a rate of 16 kHz. Finally audio was parameterized using 30 frequency-filtered (FF) log filter-bank energies. The FF features were used because they outperformed the standard used MFCCs in our experiments. The Speech Activity Detector is changed by the acoustic event detector explained above, see box *B* in figure 6.8, resulting in the scheme (c) proposed in the figure 6.9. The non-speech frames were ignored for further processing in this situation. As in NIST RT Evaluations, the missed speech and the false alarm speech detected in this stage remained throughout the agglomerative process, therefore the accuracy and detection performance of the speech was, once more time, a key point for the performance of the global system. We will see that the non-speech frames masking can also be shifted to the output of the diarization algorithm, as showed in figure 6.9 (d), without a significant degradation on terms of DER.

Once the speech features were extracted and the speech segments detected, a uniform segmentation was applied to initialize the classes/states of the HMM model. The selection of the number of classes or states (K_{init}) was performed in an automatic way by fixing the CCR constant to 7 seconds of data per Gaussian

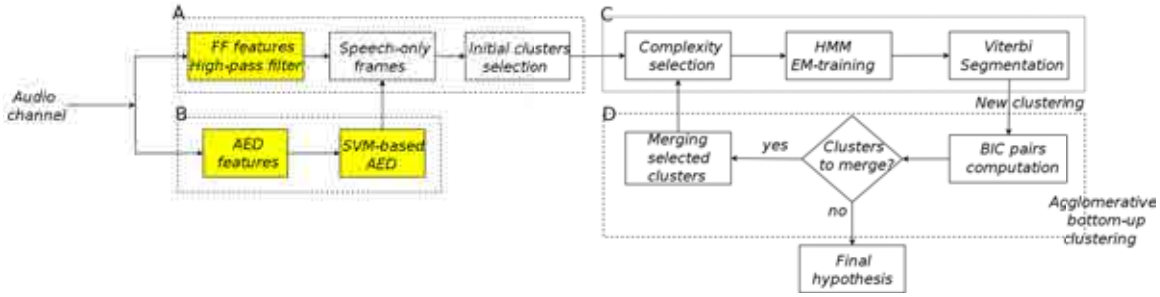


Figure 6.8: *Brief scheme of the diarization system implementation. The system employed for speaker diarization in BN are the same system employed in diarization for meeting domain. In contrast, Frequency Filtering parameters and a high-pass filter were employed for signal parametrization and the speech detector is changed by an acoustic event detection module.*

and the initial number of Gaussians per model (G_{init}) to 5. Considering the mean duration of Àgora shows, which is over 40 minutes, the initial number of cluster ranges between 40 and 50 clusters. A time constraint of 3 seconds was set to ensure minimum duration of the speaker segments. In addition, multiple merging of segments at each iteration was allowed and automatic complexity was also employed during the iterative training/segmentation steps.

6.2.4 Experiments

The main goal of the experiments which were conducted on the Àgora corpus was to assess the performance of both audio segmentation and speaker diarization systems, evaluate their joint operation and their integration. Prior to the speaker diarization experiments the performance of standalone audio segmentation system is presented. The initial diarization experiments should determine the performance interval in which the error rate of the system using audio segmentation hypothesis will be positioned.

In general, the segmentation hypothesis can be applied in two ways. First, it can be used beforehand to extract from the audio stream data of interest and those are consequently fed into the diarization system. Second, all data can be used in the diarization but the output labeling is masked according to segments of interest in the audio segmentation hypothesis.

The figure 6.9 depicts a schema for the different strategies employed for assessing influence of acoustic event detection w.r.t overall diarization performance. The figure (a) stands for the raw strategy which relies on feeding the diarization system with the whole audio data, by contrast, in figure (b) diarization is performed over speech segments extracted according an oracle event detector based on the reference transcriptions. In figure (c) the system relies on prior detection of speech thus the diarization is carried on over speech segments extracted according to the audio segmentation hypothesis. In figure (d) the diarization still depends upon acoustic event detection hypothesis but employing to mask the output of the diarization instead of masking the audio at the beginning. Finally in figure (e) and, as conclusions of the results from previous strategies, we proposed the use of two independent systems which separately performed diarization for telephone and non-telephone channel speech, combining both diarization outputs. The evaluation metric employed to assess

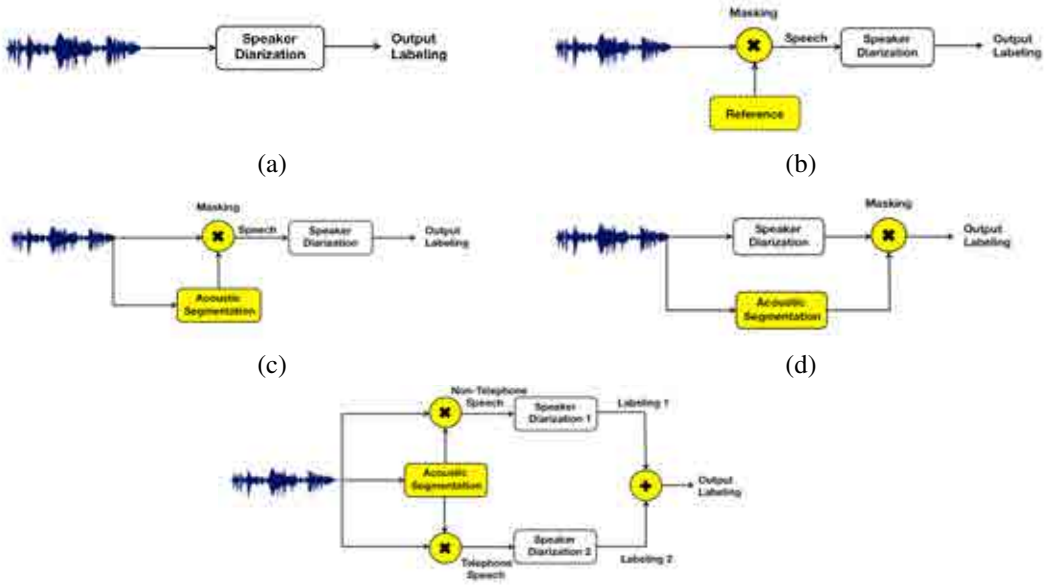


Figure 6.9: *Experiment strategies: (a) all audio is fed to the diarization system; (b) diarization over speech segments extracted according to the reference transcriptions; (c) diarization over speech segments extracted according to audio segmentation hypothesis; (d) diarization output speech-masked with audio segmentation hypothesis; (e) Independent diarization for telephone and non-telephone channel speech and combination of the two clustering outputs.*

the performance of the algorithm was the diarization error rate (DER). It involves three types of errors: Missed speech which refers to the speech segments that were not included in diarization labeling, false alarms which refer to non-speech segments that were falsely labeled as a speaker, and finally, speech for which an incorrect speaker tag was selected that is denoted as speaker error (SER). It is worth to note that, in the scenarios when the diarization system was using the segmentation hypothesis as audio masking, the first two error kinds can be directly linked to the latter system.

Audio Segmentation Performance

Given that the different acoustic conditions that are intended to be detected span in time much longer than the ones in [Temko and Nadeu, 2006] the segment duration in which mean and variance across the frames were computed was one of the parameters under study. Durations of 500 ms, 1 second and 2 seconds were put under test to evaluate their performance.

The training set consisted in 19.4 hours of audio corresponding to 15 Àgora shows described in section 6.2.1. These audio recordings were labeled using the references modified according to section 6.2.1 and provided an unbalanced amount of data for the different acoustic conditions. In average each show contains 3.9%

Selection	500 ms	1000 ms	2000 ms
Best	0.3 / 2.4 / 4.5 / 7.20	0.3 / 2.4 / 4.5 / 7.19	0.3 / 2.4 / 4.4 / **7.13**
Middle	0.1 / 2.5 / 4.5 / 7.18	0.1 / 2.5 / 4.6 / 7.24	0.2 / 2.5 / 4.6 / 7.31
Worst	0.1 / 2.6 / 4.8 / 7.47	0.1 / 2.6 / 4.9 / 7.56	0.1 / 2.6 / 4.7 / 7.39

Table 6.3: *Segmentation error rates - organized as Deletions / Insertions / Substitutions / Overall. The silence hypothesis for 1000 ms. and 2000 ms. is extracted from the 500ms column.*

of silence, 1.7% of telephonic speech over music, 5.6% of speech over music, 86.8% of speech, 1.4% of music and 0.6% of telephonic speech. For the development set 8.6 hours of audio was used, and the modified references showed that the acoustic condition distribution in this set is 3.4% of silence, 1.6% of telephonic speech over music, 5.3% of speech over music, 87% of speech, 1.5% of music and 1.3% of telephonic speech. We observe that there were not enough segments of silence with duration greater than 1 second. Therefore, it is not possible to train models with silence information for 1 second and 2 seconds. This problem is bypassed by detecting silence using 500 ms. and merging these silences in the final decisions taken without taking into account silence information. The table 6.3 shows the segmentation error rates for different silence durations 500 ms. 1 second and 2 second organized as Deletions / Insertions / Substitutions / Overall. In order to select those data chunks which will be used to train the SVMs, we sort them according to their cross validation accuracies and select either the best chunks, the worst chunks, or the chunks in the middle respectively. The segmentation errors obtained suggested us to employ the "Best" set of chunks in cross-validation and perform the segmentation using durations of 2 seconds.

Furthermore, taking into account the importance of silence detection for speaker diarization, we dealt with several approaches aiming to adapt so much as possible to the silence statistics from the data:

- Treating silence as another regular acoustic class.
- Treating silence as a regular class but taking precedence in the DAG architecture.
- Training a discriminative model of silence against non-silence.

Selection	SIvsNO	SIvsCLASS	SIvsBEST
Best	0.1 / 2.8 / 4.7 / 7.58	0.6 / 2.2 / 4.4 / 7.30	0.3 / 2.4 / 4.5 / 7.20
Middle	0.3 / 2.5 / 4.6 / 7.40	0.4 / 2.4 / 4.4 / 7.28	0.1 / 2.5 / 4.5 / **7.18**
Worst	0.3 / 2.4 / 5.0 / 7.70	0.4 / 2.4 / 4.8 / 7.63	0.1 / 2.6 / 4.8 / 7.47

Table 6.4: *Segmentation error rates - organized as Deletions / Insertions / Substitutions / Overall. SIvsNO represents the silence versus non-silence architecture, SIvsCLASS represents the architecture where silence is treated as a regular class and SIvsBEST represents the architecture where silence is treated as a regular class but taking precedence in the DAG topology. All these values have been computed using a segment of 500 ms.*

In the table 6.4 it is evaluated which of the three previously methods performs better for silence detection. In this case, the results corresponds to event detection in 500 ms. segments. The *SIvsBEST* method performed better than the others in our experiments hence we decided to follow such a strategy for silence detection in the following experiments.

Speaker Diarization Performance Bounds

The results for speaker diarization experiments were computed in two data sets: development and evaluation. Both of them was composed of 13 Àgora show recordings and they were defined to have similar acoustic characteristics. Development set comprised 8h 38' and evaluation set 7h 56' of audio data respectively. In the first experiment we performed speaker diarization without audio segmentation information. Thus all the audio data, silences and music included, was considered for speaker labeling. The idea was to define the lower performance bound of the system.

On the contrary, we can also defined the upper performance bound by using a perfect audio segmentation. A perfect segmentation was achieved by extracting speech segments according to the reference transcription. Here, the entire DER was caused either by incorrect speaker clustering or by missed speech due to overlapped speech of multiple speakers, since our system was assigning only one label for a segment. The difference between the upper and lower limit corresponds to the impact of applying an audio segmentation for speaker diarization.

The experimental setup schemas for these two experiments are depicted in figure 6.9 (a) and (b). From the numbers in the table 6.5 (columns "No AS" and "Oracle AS ") it is obvious that in the perfect case it might reach around 2% and 3% absolute DER improvements for development and evaluation sets, respectively. It is worth to note that the speaker error (SER), which is reported inside brackets, is also degraded due to corrupted clusters produced by silences, music and other events. As expected, it is lower in the case of "Oracle AS " system being more significative in evaluation set than in development, around 1.5% which represents half of the total DER.

Same conclusions are extracted from columns "AS Input" and "AS Output ". The use of audio segmentation as input for speaker diarization performs better than the output masking strategy. Speaker error is degraded over 0.4% by arranging the AED masking after diarization. It encourages the idea that audio segmentation applied at the input allows more "pure" speaker cluster resulting in a better diarization performance.

DER (SER) [%]	No AS (A)	AS Oracle (B)	AS Input (C)	AS Ouput(D)
Development	15,77 (10,9)	13,80 (10,7)	14,57 (11,0)	14,49 (10,9)
Evaluation	13,46 (8,7)	10,48 (7,4)	12,11 (8,3)	12.41 (8.7)

Table 6.5: *Speaker diarization experiment results: (No AS) without any audio segmentation; (Perfect AS) speech extracted according to the reference; (SpeechExt) speech extracted according to segmentation hypothesis; (OutMask) all audio data is given to diarization and speaker labeling is masked with speech segments; (Tel + Non-Tel). Speaker ERror (SER) is also reported between brackets.*

Speaker Diarization using Audio Segmentation

The speaker diarization algorithm will try to look for speaker changes and finally will assign cluster labels to any data which is given as input. The audio segmentation hypothesis assists the diarization process by localizing applicable data in order to prevent labeling of non-speech segments. One approach is to extract speech only before providing the data to diarization (figure 6.9 (c)) and the other is to perform a post-processing of the speaker transcription so that non-applicable time segments are discarded (figure 6.9 (d)). In the context of this work we refer to the latter case as output masking. It needs to be emphasized that the diarization labeling which is masked is obtained for the whole audio stream (including e.g. silences). The comparison of these two approaches unveils the influence of cluster purity on the performance. The difference in DER for evaluation set, as can be seen from last 2 columns in the table 6.5, is not more than 2.30% relatively and for the development set the masking approach is even slightly better than the extraction approach. The difference is not significant and, despite of input masking seems a better strategy, the speaker diarization system performed robustly to cope with cluster purity issues.

After taking a closer look on the erroneous points, it was discovered that for telephone channel speakers the diarization system usually creates just one cluster joining all speakers in such recording portions. In order to cope with this problem, we decided to apply a more tailored diarization for the telephone channel audio.

Again, audio segmentation information was used, but here also to distinguish between telephone and non-telephone speech. The structure of the TV shows guarantees that the identity of speakers in studio is different to those who are calling by telephone. Otherwise an additional recognition mechanism would be required for linking speakers speaking among both channels avoiding to introduce an artificial speaker error.

This diarization strategy is schematically illustrated in figure 6.9 (e). The speech data was split into two sets and separate diarization was performed for both of them in parallel. The output labeling for telephone and other speech is disjunctive, so it is easy to merge them into the final speaker labeling. A different feature extraction was selected for telephone speech diarization. Speech data was band-pass filtered (100 – 4000 Hz) and only 20 FF coefficients were used. Additionally, instead of using an automatic initial cluster selection, the number of initial clusters was fixed to 10, due to the prior knowledge about the Àgora shows structure.

The performance of this approach is presented in the table 6.6. For the development set, it can be observed an improvement over 4.6% relative compared to "AS Input" scenario, see the third column in the table 6.5. The result is actually slightly worse for the evaluation set although the difference is not very significant. In spite of the special treatment of telephonic channel speech, the error portions are very unbalanced. Speaker diarization

DER (SER)[%]	Tel(F.nt)	No-Tel(F.t)	Tel +No-Tel(F.unio)
Development	21,09 (12,6)	12,56 (8,8)	12,78 (9,2)
Evaluation	24,65 (15,6)	10,89 (7,0)	11,30 (7,5)

Table 6.6: *Individual diarization of telephone speech: (Tel) telephone speech only; (Non-Tel) non-telephone speech; (Tel + Non-Tel) tel. and non-tel. labeling merged.*

	Experiment	A	A.mfcc	B	C	D	D.mfcc	F.nt	F.t	F.merged
	Parametres	FF 30	MFCC 30	FF 30	FF 30	FF 30	MFCC 30	FF 30	FF 20	
Development	**DER [%]**	**15,77%**	**16,47%**	**13,80%**	**14,57%**	**14,49%**	**15,19%**	**12,56%**	**21,09%**	**12,78%**
	MS [%]	3,1%	3,1%	3,1%	3,2%	3,2%	3,2%	3,3%	8,5%	3,2%
	FA [%]	1,8%	1,8%	0,0%	0,4%	0,4%	0,4%	0,4%	0,0%	0,4%
	SER [%]	10,9%	11,6%	10,7%	11,0%	10,9%	11,6%	8,8%	12,6%	9,2%
Evaluation	**DER [%]**	**13,46%**	**14,52%**	**10,48%**	**12,11%**	**12,41%**	**13,45%**	**10,89%**	**24,65%**	**11,30%**
	MS [%]	3,1%	3,1%	3,1%	3,4%	3,4%	3,4%	3,5%	9,1%	3,4%
	FA [%]	1,6%	1,6%	0,0%	0,4%	0,4%	0,4%	0,4%	0,0%	0,4%
	SER [%]	8,7%	9,8%	7,4%	8,3%	8,7%	9,7%	7,0%	15,6%	7,5%

Table 6.7: *Results summary for the different strategies depicted in the figure 6.9. (A) without any audio segmentation; (B) speech extracted according to the reference; (C) speech extracted according to segmentation hypothesis; (D) all audio data is given to diarization and speaker labeling is masked with speech segments; (F.nt) diarization in non-telephonic channel audio; (F.t) diarization adapted to telephonic channel audio ;(F.merged) parallel diarization based on telephonic channel detection. In addition a comparison between mel cepstrum-based (MFCC) features and frequency filtering (FF) features is reported.*

in studio channel reported half of the DER compared to diarization in telephonic channel. The conclusion extracted by comparing the first and the second column in the table 6.5 is that telephone speech speaker errors account for the majority of the DER. Therefore, previous DER results support the idea of treating telephone speech separately is correct, but a more intense focus needs to be put on adjusting of the diarization system for the particularities of the telephone channel. It should be take in special account for those recordings in which low-band speech represents an important percentage of the total speech.

The table 6.7 summarizes the diarization results in terms of speaker errors, misses and false alarm errors for the different strategies joining audio segmentation and speaker diarization. In addition, a comparison between MFCC and FF features is also reported in which FF clearly outperforms MFCC features. The three last column in the table 6.7 report the DER obtained by the diarization based on channel detection. It is worth to mention that the DER reported, e.g., in the telephonic (F.t) system, was computed with respect to the total telephonic time. It represents the 0.6% and 1.3% total time in development and evaluation sets respectively. The (F.merged) column stands for the combination of the two hypothesis provided from diarization in both studio and telephonic channels. Miss speech (MS) and speaker error (SER) account for the degradation of DER comparing the diarization in studio (F.nt) and in telephonic (F.t) conditions. On the one hand the percentage of miss speech error increment suggest a not well adapted speech/non-speech segmentation. On the other hand the growing of speaker error points to a higher difficulty of the diarization system to discriminate among speakers in low-band conditions. Nonetheless, the specific diarization carried out improved the non-adapted diarization as it can be seen by comparing between (C) and (F.merge) columns. The speaker error was decreased by using the “ad hoc“ diarization which confirms the best adaptation of this approach to telephonic conditions. Summarizing, the results obtained support the idea of specific speaker diarization adapted to channel and background conditions as a strategy to take into account in broadcast news data.

	Experiment →	Reference	A.ff	A.mfcc	B	C	D.ff	D.mfcc	F.merge
	Show ↓	#Speakers							
Development	agora_2007_01_15_a	9	10 (+1)	10 (+1)	8 (-1)	9	10 (+1)	10 (+1)	12 (+3)
	agora_2007_01_29_b	10	10	10	8 (-2)	9 (-1)	10	10	14 (+4)
	agora_2007_02_26_a	13	10 (-3)	8 (-5)	9 (-4)	8 (-5)	10 (-3)	8 (-5)	14 (+1)
	agora_2007_03_12_a	8	8	7 (-1)	7 (-1)	7 (-1)	8	7 (-1)	10 (+2)
	agora_2007_03_26_a	13	11 (-2)	10 (-3)	9 (-4)	10 (-3)	10 (-3)	10 (-3)	14 (+1)
	agora_2007_04_02_a	11	8 (-3)	8 (-3)	7 (-4)	7 (-4)	8 (-3)	8 (-3)	13 (+2)
	agora_2007_05_07_b	11	10 (-1)	9 (-2)	8 (-3)	8 (-3)	10 (-1)	9 (-2)	12 (+1)
	agora_2007_05_14_a	10	9 (-1)	9 (-1)	8 (-2)	7 (-3)	9 (-1)	9 (-1)	12 (+2)
	agora_2007_05_14_b	11	9 (-2)	9 (-2)	8 (-3)	8 (-3)	9 (-2)	9 (-2)	12 (+1)
	agora_2007_06_11_b	11	10 (-1)	9 (-2)	8 (-3)	9 (-2)	10 (-1)	9 (-2)	14 (+3)
	agora_2007_07_02_a	14	12 (-2)	9 (-5)	10 (-4)	10 (-4)	12 (-2)	9 (-5)	17 (+3)
	agora_2007_10_01_b	14	12 (-2)	14	12 (-2)	12 (-2)	12 (-2)	14	16 (+2)
	agora_2007_11_12_b	13	11 (-2)	13	9 (-4)	11 (-2)	11 (-2)	13	13
	Total (Detected / Errors)	148	**130(20)**	125 (25)	111 (37)	115 (33)	129 (21)	127 (25)	173 (25)
Evaluation	agora_2007_01_08_b	6	9 (+3)	8 (+2)	7 (+1)	8 (+2)	9 (+3)	8 (+2)	9 (+3)
	agora_2007_02_12_a	11	9 (-2)	8 (-3)	7 (-4)	7 (-4)	9 (-2)	8 (-3)	12 (+1)
	agora_2007_04_16_a	11	9 (-2)	11	7 (-4)	7 (-4)	9 (-2)	11	11
	agora_2007_06_11_a	11	9 (-2)	9 (-2)	7 (-4)	8 (-3)	9 (-2)	9 (-2)	12 (+1)
	agora_2007_06_18_b	10	8 (-2)	8 (-2)	8 (-2)	8 (-2)	8 (-2)	8 (-2)	12 (+2)
	agora_2007_07_02_b	12	11 (-1)	10 (-2)	9 (-3)	9 (-3)	11 (-1)	10 (-2)	14 (+2)
	agora_2007_07_09_b	11	9 (-2)	10 (-1)	7 (-4)	7 (-4)	9 (-2)	10 (-1)	11
	agora_2007_10_01_a	10	10	9 (-1)	8 (-2)	9 (-1)	10	9 (-1)	11 (+1)
	agora_2007_10_08_a	15	12 (-3)	12 (-3)	9 (-6)	10 (-5)	12 (-3)	12 (-3)	12 (-3)
	agora_2007_10_15_b	12	9 (-3)	8 (-4)	8 (-4)	8 (-4)	9 (-3)	8 (-4)	12
	agora_2007_10_22_a	11	11	11	10 (-1)	10 (-1)	11 (-1)	11 (-1)	15 (+3)
	agora_2007_11_05_b	12	11 (-1)	9 (-3)	9 (-3)	9 (-3)	11 (-1)	9 (-3)	13 (+1)
	agora_2007_11_12_a	18	12 (-6)	11 (-7)	10 (-8)	10 (-8)	12 (-6)	11 (-7)	16 (-2)
	Total (Detected / Errors)	150	129(27)	116 (30)	106 (46)	110 (44)	129 (28)	124 (31)	**160 (19)**

Table 6.8: *Number of speakers detected for both show recording and speaker diarization strategy. The number of speakers is reported for development and evaluation test set respectively. First column labeled "Reference" shows the number of speakers present in the transcriptions. The number of speaker detected by the system. Between brackets: the error with respect to reference, (+) means overclustering and (-) underclustering. Furthermore, the total detected speakers and the associated speaker error count per system is also reported.*

Finally, the table 6.8 reports the number of speakers detected for each recording and by the different speaker diarization strategies. It is worth to note that a right number of detected speakers might not be directly linked to lower DER errors. It can be also noticed by comparing columns (C) corresponding to "AS Input" system with column (B) corresponding to "AS Output " system. The former tends to undercluster the data more than

the "AS Output " approach does: 33 compared to 21 in development and 44 to 28 in evaluation. Nonetheless, the experiments reported in the table 6.7 show a similar performance in terms of DER for both systems.
The last column (F.merge) corresponds to the number of speakers detected by the parallel diarization system. As can be seen such a system overclusters in mean the data. For example, it gets 10 more speakers in evaluation than those present in the reference. However it obtained the lower number of speaker detected errors in the same evaluation data, 19 in total. The undercluster suffered by most of the systems is corrected in this approach but it falls out in an overclustered hypothesis though with higher precision ability to discovering speakers.

6.2.5 Conclusions

The results obtained, with DER percentages around 11%, showed an acceptable adaptation of the audio segmentation and of the speaker diarization to the Àgora database. The DER error is a time-based metric which means that during the 11% of the total time of one Àgora show it occurs a speaker error label due to either speech/non-speech errors or not well recognized speakers. Such a results are comparable to other results reported in the literature [Fiscus and et al., 2007b].
The proposed strategy which combines audio segmentation with speaker diarization based on channel detection has reported improved results. Speaker diarization by means several diarization systems adapted to each background condition is a promising approach to deal with mismatch conditions likely to occur in broadcast news data as, e.g., changes between studio and outside or telephonic (low-band) channel.
The DER error computed was mainly composed of speaker errors (SER) around 7% and segmentation errors (MS+FA) around 4%. The experiments reported point towards including verification techniques in order to still reduce more the DER error by reducing SER. For example, the inclusion of speaker overlapping detection might reduce speaker overlap errors. It represents half of the total SER error in this case. Speaker verification techniques as normalization of adaptation might also improve clustering by reducing errors due to linking and verification errors, which represents the other half of the SER. Furthermore, different proposals for the stopping criterion should be take into account aiming to improve the number of speakers detected and indeed the quality of the speaker diarization hypothesis.
The combination and the integration between audio segmentation and speaker diarization should also be other path for the improvement as it has been reported in this chapter. The finding of useful information and possible synergies between these two different audio tasks will lead to improved results in both technologies.

Chapter 7

Multimodal Person Tracking

In this chapter we mainly describe a multi-information approach to deal with the task of person tracking into a smart-room. The use of several cues of information from multiple microphones and other information sources, such as video images, were combined to improve stand-alone audio and video person tracking. On the one hand, taking benefit of multiple microphones information might aid to speaker recognition as it has been reported in previous chapters, e.g., in the diarization task by joining speaker location information and classical spectral features or in person identification by combining speech and face identification. In addition, the combination or the fusion with other modalities as video person identification and video tracking drives to person identification algorithms which performs robustly in most of the situations. For example, it allows the automatic system constantly identify people inside a room, specially when it is the case of not existing speech and the video is the only source of information. The same occurs whenever no images were available into the room due occlusions and just audio information can be accessed.

Detecting the location and identity of users is a first step in creating context-aware applications for technologically-endowed environments. The spatio-temporal localization and recognition of people through various sensors poses problems of great theoretical and practical interest, in particular for home environments and smart rooms. In these scenarios, context-awareness is based on technologies like gesture and motion segmentation, unsupervised learning of human actions, determination of the focus of attention or intelligent allocation of computational resources to different modalities.

This work is the result of the collaboration between the speech processing group (VEU), the image and video processing group (GPI) both of them from the Signal Theory and Communications Department (TSC) at UPC; and the Signals and Images Research Group, Centre for Mathematics and Computer Science (CWI) from Amsterdam. The different groups met at Boğaziçi University, Instanbul, during the SIMILAR NoE Summer Workshop on Multimodal Interfaces, eNTERFACE'07 [enterface, 2007] which laid the foundations for this collaboration. Furthermore, the work performed during the summer school was collected in the Proceedings of the eNTERFACE 2007 and a journal publication [Salah *et al.*, 2008b] was also accepted in the Journal on Multimodal User Interfaces.

7.1 Audio and Video Modalities for Person Tracking

In this work we aimed at putting together different algorithms for detection, tracking and identification, working in a completely automatic way. Through the observation and subsequent processing of the data captured using a large number of sensors from multiple modalities, we tried to determine the identity and the spatial positions of people in the room. Obtaining this knowledge is the first step in developing more elaborate smart applications. The main contribution of this effort was an intuitive way of connecting different tracking and recognition methods to perform multimodal tracking and identification in the smart environment.
We propose a system that makes use of motion detection, video tracking, face identification, image-based identification, audio-based localization, and audio-based identification modules, fusing information to obtain robust localization and identification.
The smart room sensors and setup employed in this work correspond to the UPC smartroom laboratory explained in the section 3.1. The data streams are processed with the help of the generic client-server middleware SmartFlow, resulting in a flexible architecture that runs across different platforms.

7.1.1 Experimental Setup

Audiovisual recordings of interactive small working-group seminars were used. These recordings were collected at the UPC smart room, in accordance with the "CHIL Room Setup" specifications [Casas and Stiefelhagen, 2005]. Recordings were performed at different dates (several months apart) to ensure proper variability (face, hair, etc.) of the participants.
The training data employed for algorithm development was composed of one recording lasting 5 minutes. Target speakers and different classes of acoustic events were also present in the recording. Background noise, overlapping speech, door slam, laugh, steps are some examples of these events. Silences of a speaker in between talk segments were not labeled if they were smaller than 500 ms. Additionally, 1 minute of speech per speaker was used to train their models and 1 minute of different events were recorded to train an event model. Finally, an audio sequence around 7 minutes was employed to benchmark the performance of the approach (TEST).
In both training and test recordings, four people enter an empty room, one by one. Once inside, they move around a central table, always in standing position, talk to each other, walk around the room from time to time, and finally leave the room one by one. The length of each recording is approximately five minutes. For some algorithms a relatively large amount of training data needed to be available. One of the recordings was intended for training in those cases, and the second one was used for testing without any further change or parameter adjustment. The second recording is more difficult in terms of tracking, as all subjects wear clothes of similar colors. For additional training we used a set of similar recordings from the CLEAR evaluation campaign [Mostefa and et al., 2006].
The UPC's smart-room has an entrance door, a big window and a table in the middle, see the figure A.1 in appendix A. The window was closed during recordings to avoid illumination changes. However, small

illumination changes can occur when the door was opened. The room is monitored using six cameras: four fixed cameras at the corners of the room (labeled Cam1 to Cam4 in the figure A.1), one zenithal fish-eye camera at the ceiling (Cam5) and one active camera (PTZ) aimed and zoomed at the entry door to capture the faces of the incoming people at high resolution. Video is interlaced, recorded in compressed JPEG format, at 25 fps and 768×576 resolution.
The audio sensor setup is composed by one NIST Mark III 64-channel microphone array, three T-shaped four-channel microphone clusters and eight tabletop microphones. Audio is recorded in separate channels at 44.100 kHz sampling frequency and 2 bytes per sample. Far-field conditions have been used for both audio and video modalities. All data flows are timestamped and the computers used to record the signals are synchronized using the network time protocol (NTP). This makes possible to synchronize audio and video data. There is no manual segmentation of the data. Each technology is supposed to automatically segment the recorded signal.

7.1.2 The Middleware

The problem of interconnecting several algorithms that work on data streams coming from a high number of sensors from different modalities is far from trivial. Synchronization of the different data flows, distributed computing and the interconnection of the algorithms are issues that need to be addressed.
To allow efficient communication of sensor data and distributed computation, it is useful to have a middleware that provides infrastructure services. We propose to use the NIST SmartFlow system that allows the transportation of large amounts of data from sensors to recognition algorithms running on distributed, networked nodes [smartflow, 2002; darpa, 1998]. The working installations of SmartFlow is reportedly able to support hundreds of sensors [Stanford *et al.*, 2003]. In the present version of our system, the integration was not completed, as some modules are implemented with MATLAB, and data exchange of modules was simulated. However, the architecture is set up in a modular fashion to allow complete implementation under SmartFlow.
Smartflow offers a great deal of data encapsulation for the processing blocks, which are called “clients”. Each client can output one or more flows for the benefit of other clients. The communication over TCP/IP sockets is transparent to the user, and handled by the middleware. The design of a working system is realized through a graphical user interface, where clients are depicted as blocks and flows as connections. The user can drag and drop client blocks onto a map, connect the clients via flows, and activate these processing blocks.
The synchronization of the clients is achieved by synchronizing the time for each driving computer, and time stamping the flows. The network time protocol (NTP) is used to synchronize the clients with the server time, and this functionality is provided by SmartFlow. A separate client is used to start the processing clients simultaneously. The video streams are not completely in one-to-one correspondence, as clients sometimes drop frames.

7.1.3 The Information Flow of the System

The figure 7.1 depicts the information flow within the system: When people enter the room, the face detection module detects the face on the PTZ camera, and marks the face area, which is then identified by the face identification module. This provides a reliable identification, which is used to trigger the feature-based identification (FBI) module. The FBI receives moving blobs that are detected and tracked by the tracking module, and builds a feature model for the identified person on-the-fly. The FBI module is thus responsible for the continuity of tracked frames, and serves as a weak biometrics system that can identify users in cases where tracking fails, or stronger biometric information is not available. Motion detection and tracking within the room are performed using the data from the four corner cameras and the ceiling camera.
The audio modules track people in the room via sound localization, and identify them based on their speech characteristics. Since sound and speech are not constantly available, this modality is mainly used for making the decisions of the system more robust. The acoustic identification can help the FBI module to re-assign true IDs to the detected blobs in case of tracking failures due occlusions or color similarity. Similarly, acoustic localization is used to assign the ID of the speaker to one of the tracked blobs.
Localization and identification are both controlled from a central logic, which performs quality-based fusion of information. It corresponds to the labeled ”Multimodal Identification“ box in the figure 7.1.

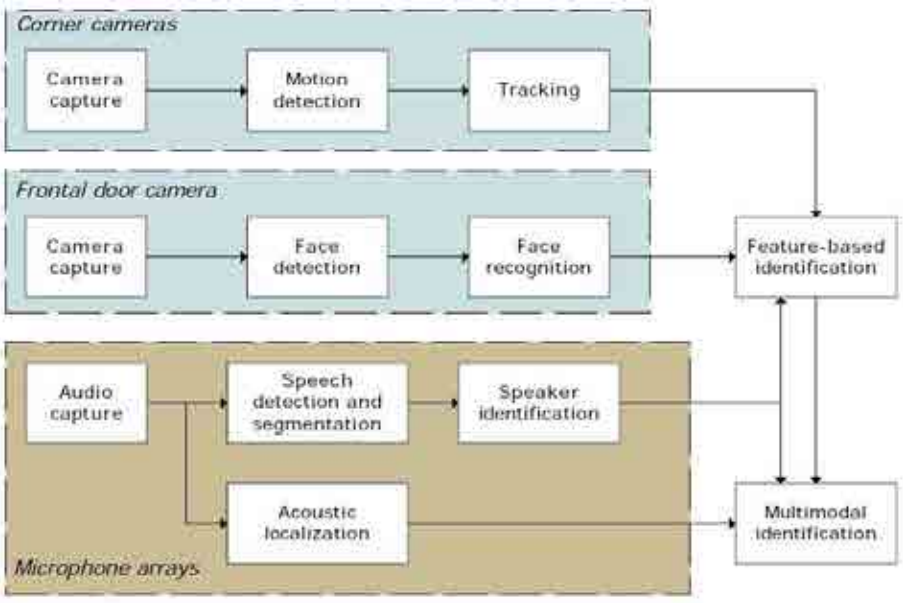


Figure 7.1: *The multimodal flow of information within the smart room architecture. Blue boxes are related to video and image technologies, that is, motion detection and tracking and face recognition; whereas brown box encloses acoustic technologies as speaker tracking and speaker localization. The feature based identification module combines multimodal information in order to provide a unique robust identity and along with the multimodal identification module perform speaker tracking into the room.*

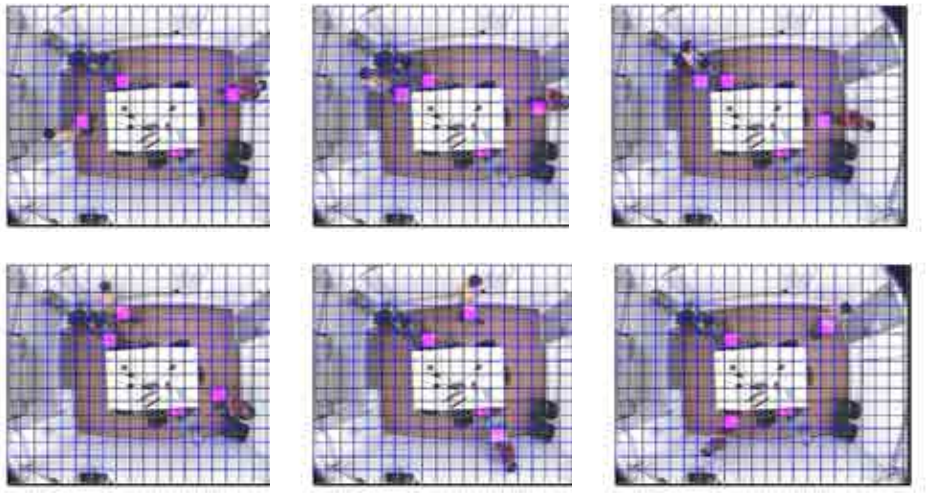


Figure 7.2: *The stand-alone operation of the localization module. The size of the occupancy grid spacing is roughly 40 centimeters.*

7.2 Visual Processing

The cameras in the system are responsible from motion detection, tracking, face detection and identification, and feature-based identification. Following we briefly describe the visual algorithms, for further details see the work in [Salah *et al.*, 2008b].

The motion detection module, see figure 7.1 attempts to separate the foreground from the background for its operation. The method used was based on detecting moving objects under the assumption that images of a scene without moving objects show regularities, which can be modeled using statistical methods. The training set is constructed with a short sequence of off-line recording taken from the empty room and adapted to illumination changes by on-line adding new samples to the training set. One significant advantage of employed technique is, as the new values are allowed to be part of the model, the old model was not completely discarded. If the new values become stabilized over time, the weighting changes accordingly, and new values tend to have more weight as older values become less important.

In order to track person motion a probabilistic occupancy map (POM) approach, related to the algorithm proposed in [Fleuret *et al.*, 2008], but simplified to deal with indoor environments, was employed. In the POM approach, the discrete occupancy map was used to back-project the stub image of a person (a simple rectangle) to each camera view. The overlaps between the stubs and the detected motion images across multiple cameras indicate the presence of a person at a given location.

The figure 7.2 illustrates the stand-alone operation of the algorithm on several consecutive frames. Only the four corner cameras were used in computing the occupancy, the ceiling camera was just used for ground truth annotation and visualization. The accuracy of this algorithm in terms of correct occupancy detection on

Figure 7.3: *Examples demonstrating variability in faces, acquired with the door camera. Variations due rotation, expression change and motion blur are apparent.*

our database was 96.3 per cent, allowing at most a single grid square deviation from the ground truth. The false detection rate is 5.5 per cent. Of the true detections, 58.7% were exact matches with the ground truth. The grid size was set to roughly 40 cm., which was slightly denser than [Fleuret *et al.*, 2008] that used 50 cm. For continuous detection and identification, the output of this module was combined with other types of information at a later stage.

Face detection was needed both for face identification and feature based identification modules. In this module, the face of each person present in the scene must be detected roughly (i.e. a bounding box around a face will be the output of this module). We used the OpenCV face detection module that relies on the adaboosted cascade of Haar features, i.e. the Viola-Jones algorithm [Viola and Jones, 2001]. The client that performs face detection receives a video flow from a client that in its turn directly receives its input from one of the cameras, and outputs a flow that contains the bounding box of the detected face. The face images captured by ceiling cameras are too small for reliable detection or identification. Consequently, only the door camera is used in face detection and recognition.

Our face recognition module is semi-automatic, in that it takes motion tracking and face detection for granted. This module therefore subscribes to the face detection flow that indicates face locations, and to the video flow to analyze the visual input to a camera. The same technique as in section 3.2.2 for face recognition in smart environments was employed [Vilaplana *et al.*, 2006; Luque *et al.*, 2006b]. The technique takes advantage of the continuous monitoring of the environment and combines the information of several images to perform the recognition. Models for all individuals in the database were created off-line using sequences of images collected at a different date than the training and testing recordings.

We evaluated the face identification module for the PTZ camera, which records people entering the room with a high resolution. Its positioning allows the capture of an head-and-shoulders image with a resolution of 768×576 pixels. An examples of images are depicted in figure 7.3. As previously mentioned, the gallery models were created with images of each person taken from a different recording. We used 20 training images

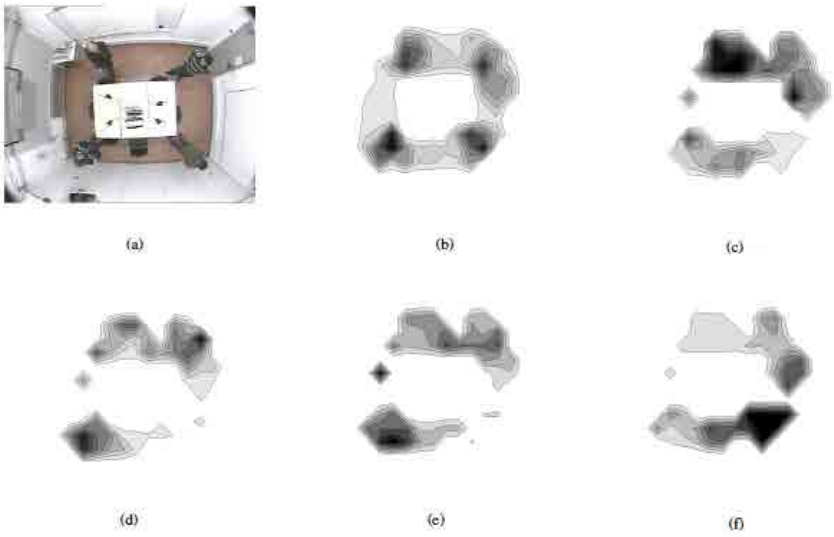


Figure 7.4: *The operation of POM and FBI modules. (a) A sample frame from the second session. (b) The POM is computed through motion detection in four corner cameras. Darker colors indicate higher probability. (c)-(f) The posterior probabilities under models of individuals in the room, computed by the FBI module. The order of maps reflects the relative positions of the persons in the room. While the persons in (c), (e) and (f) are unambiguously identified, the model for the top-right person produces high posterior probabilities for two persons as showed in (d). The assignment of identities to occupied locations that jointly maximizes the posterior probability is able to identify all persons correctly. Picture courtesy of Prof. Albert Ali Salah*

per individual, and the corpora was collected with four people participating the recordings. The identification module was able to give correct results in all the correctly detected groups of faces. In the case of false positives of the face detection module, the face identification module correctly classified all the cases in the *Unknown/No face* class. For the training recording, the face detection module outputs 172 groups of faces. Of these 172 groups, only 111 (65%) correspond to good detections of frontal faces. The rest of the detections correspond to non-frontal faces and false positives (35%).

In order to ensure robust identification in the room, we relied on weak or unanticipated sensor correlations and patterns for sensing. For this purpose, we propose to use features that are easy to capture, present most of the time, and helpful when the strong modalities (e.g. face or speech) were unreliable. This is particularly important in the smart-room scenario, where the ceiling cameras usually cannot capture discriminative face images. Most proposed systems for this type of application rely on continuous tracking of identified people in the room. Important approaches that are previously proposed and successfully used are Kalman filters [Katsarakis *et al.*, 2007] and particle filters [Nickel *et al.*, 2005a]. However, tracking multiple people for a long period is difficult, there is little possibility of recovering from mistakes.

The proposed feature based identification (FBI) module, see the figure 7.1, aimed at identifying persons in

the room when the tracking or speaker identification results are not available, or not discriminatory. The primary assumption behind the operation of this module was that the variability in a user's appearance for a single camera was relatively low for a single session (this case is termed intra session in [Vilaplana *et al.*, 2006]), and a user model created on-the-fly can provide us with useful information [Tangelder and Schouten, 2006]. We used the following general procedure for this purpose: Whenever a person was reliably identified (i.e. when the face identification or speaker identification modules return a result with high confidence[1]), the tracking cameras forward the detected motion blobs to the FBI module. Then, the pixel intensities within the motion blobs were modeled statistically, and this statistical model was used to produce posterior probabilities for identification purposes, see figure 7.4. Typically, the system created one model per person per camera, immediately after the person entered the room, as this was the point where the door camera acquires the face image. Using a separate module for each camera made a color-based calibration across cameras unnecessary.

Visual Person Tracking

The identification started at the door, where the PTZ camera identified the person. Once the person was identified, one FBI model per camera was created on the fly. The POM module provided the motion blobs used for FBI modeling. The motion blobs associated with a single location by the POM module were passed on to the FBI modules, which compared the whole collection under each person model. Note that there were four FBI modules working in parallel, one for each camera. Each of them used the most discriminative 20 per cent of the pixels received from the POM module.

We decided to make use of particle filter (PF) algorithm [Gordon *et al.*, 1993; Carpenter *et al.*, 1997] in order to implement visual tracking, see following section 7.4.1 for further details. For that purpose, the visual likelihood evaluation function must be defined. We used a combination of POM and FBI by multiplying both terms. POM gives a probability of occupancy for each grid location and FBI gives the posterior probabilities for all participants in each occupied grid position. In order to estimate the 2D position and the identity of a person $X_t = \{(x, y)_t, ID\}$ at time t, taking as observation the combined results from POM and FBI modules up to time k, denoted as $\mathbf{z}_{1:k}$.

The function $p(\mathbf{z}_k|\mathbf{x}_k)$ was defined as the likelihood of a particle belonging to the position corresponding to a person. For a given particle j occupying a room position, its likelihood may be formulated by multiplying the probabilities from POM and FBI. This will give the weight for the particle.

$$w_k^i = (POM * FBI) \tag{7.1}$$

Our implementation made use of a set of decoupled PFs, each per target, and define an interaction model to ensure track coherence. The interaction between filters was modeled by two steps: if a particle fell into the space occupied by another filter tracking a person, the weight of this particle was set to zero. If a particle changed randomly its identity and takes the identity of another filter with a higher scoring, the weight of this

[1] The confidence depends on the modality. For face identification, we required that the average distance to the best-matching template be smaller than a threshold. This ensured that the detected area contained a face.

particle was also set to zero. The particle filters were initialized with $10,000$ particles distributed randomly across the room. Each particle was characterized by its position and its identity. The number of particles, as well as the other parameters of the PF were optimized using a exhaustive search approach using the development recording. The resulting parameters were used blindly on the test set.

7.3 Audio Processing

A total of 20 microphones were used to track speakers into the room, 12 of which were omni-directional microphones and they were placed on the walls in three T-shaped groups of four microphones each. In addition to these T-shape arrays, four directional and four omni-directional microphones were placed on the table. All the data was collected at a rate of 44.100 kHz and 2 bytes per sample. The data was downsampled to 16.000 kHz as a pre-processing step.

7.3.1 Acoustic Localization Module

Many approaches to the task of acoustic source localization in smart environments have been proposed in the literature. Their main distinguishing characteristic is the way they gather spatial clues from the acoustic signals, and how this information is processed to obtain a reliable 3D position in the room space. Spatial features, like the Time Difference of Arrival (TDOA) between a pair of microphones [Rabinkin, 1995] or the Direction of Arrival (DOA) of sound to a microphone array can be obtained on the basis of cross-correlation techniques [Omologo and Svaizer, 1997], High Resolution Spectral Estimation techniques [Potamitis *et al.*, 2003] or by source-to-microphone impulse response estimation [Chen *et al.*, 2004]. Conventional acoustic localization systems also include a tracking stage that smooths the raw position measurements to increase precision according to a motion model. It is worth to mention that most of these techniques need several synchronized high-quality microphones.

The acoustic localization system used for this work was based on the SRP-PHAT localization method, which is known to perform robustly in most scenarios. The SRP-PHAT algorithm (also known as Global Coherence Field [DiBiase *et al.*, 2001]) tackles the task of acoustic localization in a robust and efficient way. In general, the basic operation of localization techniques based on steered response power (SRP) is to search the room space for a maximum in the power of the received sound source signal using a delay-and-sum or a filter-and-sum beamformer. In the simplest case, the output of the delay-and-sum beamformer is the sum of the signals of each microphone with the adequate steering delays for the position that is explored. The SRP-PHAT algorithm consists of exploring the 3D space while searching for the maximum of the contribution of the PHAT-weighted cross-correlations between all the microphone pairs. The SRP-PHAT algorithm performs very robustly due the the PHAT weighting, keeping the simplicity of the steered beamformer approach. The figure 7.5 shows an example of TDOA trajectories which identify an acoustic source in their intersections.

Consider a smart-room provided with a set of N microphones from which we choose M microphone pairs. Let $\mathbf{x}$ denote a $\mathbf{R}^3$ position in space. Then the time delay of arrival $TDOA_{i,j}$ of an hypothetical acoustic

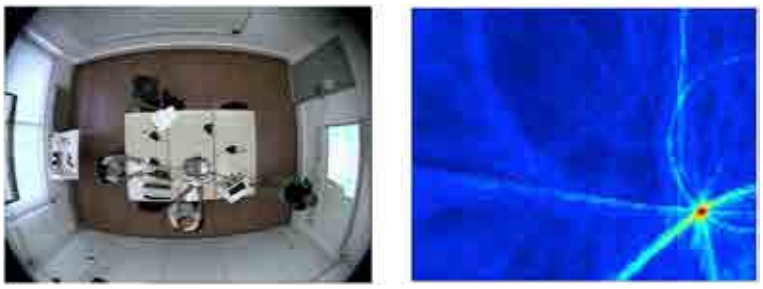


Figure 7.5: *An example of TDOA trajectories in a 2D space related to the real space into the smart-room. The intersection of trajectories identify an acoustic source, in this case, a lecturer close to a blackboard. The color is related to the contribution of cross-correlation of all microphone pairs in that position. The red point corresponds to the maximization of such a cross-correlation function. Right picture courtesy of Carlos Segura Perales.*

source located at $\mathbf{x}$ between two microphones i, j with position $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$TDOA_{i,j} = \frac{\| \mathbf{x} - \mathbf{m}_i \| - \| \mathbf{x} - \mathbf{m}_j \|}{s}, \tag{7.2}$$

where s is the speed of sound. The 3D room space is then quantized into a set of positions with typical separations of $5 - 10$ cm. The theoretical TDOA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair are pre-calculated and stored.

PHAT-weighted cross-correlations of each microphone pair are estimated for each analysis frame [Omologo and Svaizer, 1997]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($G_{m_1 m_2}(f)$) as follows:

$$R_{m_i m_j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{m_i m_j}(f)}{|G_{m_i m_j}(f)|} e^{j2\pi f \tau} df, \tag{7.3}$$

The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \sum_{i,j \in \mathbb{S}} R_{m_i m_j}(\tau_{\mathbf{x},i,j}), \tag{7.4}$$

where $\mathbb{S}$ is the set of microphone pairs. The sum of the contributions of each microphone pair cross-correlation gives a value of confidence of the estimated position, which can be used in conjunction with a threshold to detect acoustic activity and to filter out noise. In our work, we used a threshold of 0.5 per cluster of 4 microphones. It is important to note that in the case of concurrent speakers or acoustic events, this technique will only provide an estimation for the dominant acoustic source at each iteration. The experimental results obtained with the localization module are given in section 7.4, as they are used jointly with the speaker identification module.

7.3.2 Speaker Segmentation and Identification Module

Speaker tracking aims at detecting regions uttered by a given speaker, for which a speaker model is trained beforehand. Most of the proposed applications in the literature analyze the tracking problem in three parts:

1. Detecting the segments that contain speech.
2. Detecting the speaker turn.
3. Identifying the speaker.

Our speaker identification system was based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) relaying on the maximum a posteriori (MAP) adaptation for training. The HMM/GMM approach were used to model speaker features and to encode the temporal evolution of the speech segments. Mainly, the algorithm employed in this work is identical to the one described in chapter 6, which was employed to perform speaker tracking in broadcast news, but with some minor modifications. Among them: the inclusion of an acoustic model for detection of noisily events like chair moving, door opening, steps, laugh and so forth ; or the adaptation to the multiple audio channels available.
The speaker tracking algorithm segmented the audio signal into several clusters using a minimum duration of the speaker turn parameter, and assigned the most likely identity from the closed database of target speakers to each segment. We collected a small database composed of four target speakers. Just one minute of speech per speaker was collected into the room in order to estimate beforehand initial client models.
The input signals from each microphone channel were first Wiener-filtered using the implementation of the QIO front-end system [Adami *et al.*, 2002]. Then, these channels were combined in order to create a enhanced version based on beamforming techniques [Anguera, 2005]. The beamformed output channel was analyzed by the *Speech Activity Detector* (SAD) module, described in previous chapters, aiming to detect speech segments and to discard the non-speech portions. Followed, the enhanced speech data were parameterized by means 19 Mel Frequency Cepstral Coefficients (MFCC) features. A brief scheme of the audio processing is given in the figure 7.6. We will treat each of these stages for completeness' sake.

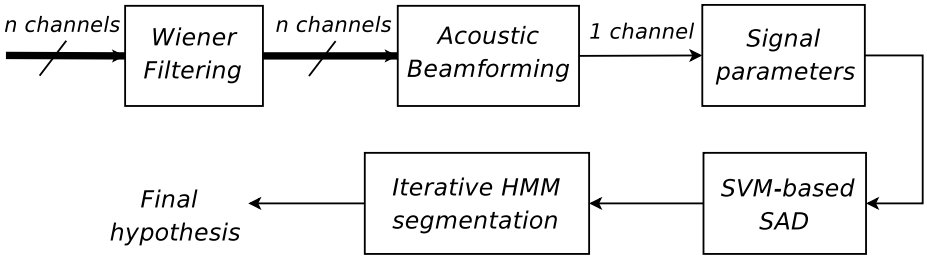
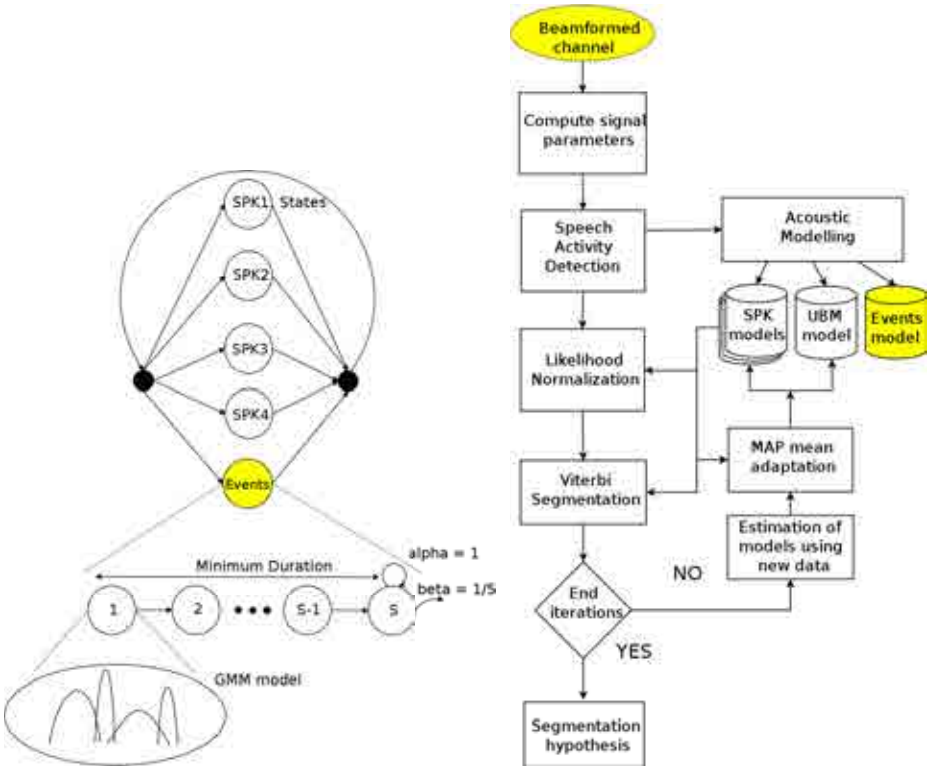


Figure 7.6: *Brief scheme of the speaker tacking system. A single enhanced channel is employed to perform diarization as in MDM condition for meeting scenario. Same speech activity detector based on SVM modeling was also employed.*



(a) (b)

Figure 7.7: *(a) An ergodic HMM models the acoustic data with* 5 *states, each composed of* S *sub-states sharing the same GMM model. (b) Steps of the iterative segmentation and identification algorithm. The tracking is performed on the beamformed channel and an events model was also included in the HMM topology.*

The speech parametrization was based on a short-term estimation of the spectrum energy in several sub-bands. The beamformed channel was analyzed in frames of 30 milliseconds at intervals of 10 milliseconds and 16 kHz of sampling frequency. A Hamming window was applied to each frame and a FFT was computed. The FFT amplitudes were then averaged in 19 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. The scheme we present follows the classical procedure used to obtain the Mel-Frequency Cepstral Coefficients (MFCC) [Davis and Mermelstein, 1980].

The tracking algorithm, as previous approaches, made use of an ergodic Hidden Markov Model (HMM) composed of 5 states. Four of them was employed per each target speaker model and the last one for the acoustic events model. The HMM topology is shown in the figure 7.7 (a). Each state contained a set of S sub-states which imposed a minimum duration constraint of the speaker turn duration. Each sub-state had an output probability density function modeled with a GMM model, which was the same for all the sub-states of a given state. For the particular dataset in consideration the minimum duration was set to 3 seconds. Such a value was changed to 1.5 second at the last segmentation/training iteration aiming to detect small speaker

turns.

For each speaker that the system had to track, the probability density function of the parameter vectors from the training speech was modeled. Gaussian Mixture Models (GMM) with diagonal covariance matrices were employed, and the number of components per mixture was set to 32. The large amount of components ensured that the statistical learning was robust, and its use was further justified by the availability of a large number of training samples. The parameters of the model were estimated from speech samples of the speakers using the iterative Expectation-Maximisation (EM) algorithm. The sensitivity of EM in cases with few training data is well known thus just 15 iterations were demonstrably enough for parameter convergence avoiding data overfitting. Such a parameter was retained for both prior training of models and for following MAP adaptations. Furthermore, a Universal Background Model (UBM) and a acoustic event model were also estimated. The former was employed to compute likelihood normalization and target speaker model adaptations. The UPC speaker database, see appendix B, was employed to compute the UBM's parameters. The second was used to model, in a explicit way, different noises as steps, chair moving, cough, laugh and so on. For that purpose a small database of events was collected. Note that in this work we used an integrated event detector by incorporating an acoustic event model into the HMM topology, at the same level than the target speaker models.

The figure 7.7 (b) depicts the algorithm step by step once we had enhanced the channel by the beamforming technique. After the segmentation and the Viterbi decoding, an initial clustering of each class was obtained. We decided to adapt the target speakers and the event models by means a MAP-mean adaptation through the UBM at each iteration. Thus the corresponding detected segments were merged with the initial enrollment data and the models were trained again. The stopping criterion for this iterative segmentation and training procedure was based on the likelihood returned by the Viterbi segmentation. The final segmentation was reached once the sequence likelihood did not decrease with respect to the previous iteration.

During the Viterbi segmentation, a normalization of likelihoods were also implemented. The emission probabilities of each state were normalized with the frame score computed from the UBM model as in the speaker verification task. That is, the Log-Likelihood Ratio (LLR) method at the score-level [Bimbot *et al.*, 2004]:

$$\hat{L}_\lambda(X) = L_\lambda(X) - L_{UBM}(X) \tag{7.5}$$

where X stands for a frame parameter vector, $L_\lambda(X)$ denotes the log-likelihood under class model λ and $L_{UBM}(X)$ is the log-likelihood under the UBM, trained with the complete training sequence.

7.4 Multimodal Processing

The purpose of the system was to identify the participants once they enter into the room and then track them during the complete session. To have robust tracking and identification of all persons at the same time, we used a multimodal approach. In this section we will describe this approach.

Our multimodal approach was based on a multi-hypothesis tracker that approximates the filtered posterior distribution by a set of weighted particles. The standard particle filter weights particles based on a likelihood score, and then propagates these weighted particles according to a motion model.

7.4.1 Particle Filter

Particle Filters (PF) have proved to be a very useful technique for tracking and estimation tasks when the variables involved do not hold Gaussianity uncertainty models and linear dynamics [Isard and Blake, 1998]. They have been successfully used for video object tracking and for audio source localization. Information of audio and video sources has also been effectively combined employing PF strategies for active speaker tracking [Nickel *et al.*, 2005b] or audiovisual multi-person tracking [Gatica-Perez *et al.*, 2007].
The optimal solution to multi-target tracking using PF is the joint PF presented in [Khan *et al.*, 2003]. However, its computational load increases dramatically with the number of targets to track, since every particle estimates the location of all targets in the scene simultaneously. The proposed solution makes use of a set of decoupled PFs, each per target, and define an interaction model to ensure track coherence. This approach has a performance advantage over the use of a joint particle filter while keeping a similar quality of the tracks generated [Khan *et al.*, 2003].
The estimation $\mathbf{x}_k$ of the position of a person at an instant k given a set of observations $\mathbf{z}_{1:k}$ can be written in the context of a state space estimate problem described by the following state process equation.

$$\mathbf{x}_k = \mathbf{f}\left(\mathbf{x}_{k-1}, \mathbf{v}_k\right), \tag{7.6}$$

and the observation equation:

$$\mathbf{z}_k = \mathbf{h}\left(\mathbf{x}_k, \mathbf{n}_k\right), \tag{7.7}$$

where $\mathbf{f}$ is a function describing the state propagation and $\mathbf{h}$ an observation function modeling the relation between the the hidden state $\mathbf{x}_k$ and its observable counterpart $\mathbf{z}_k$. The functions $\mathbf{f}$ and $\mathbf{h}$ are possibly non-linear and the noise components $\mathbf{v}_k$ and $\mathbf{n}_k$ are assumed to be independent stochastic processes with a given distribution.
From a Bayesian perspective, the tracking problem is to recursively estimate a certain degree of belief in the state variable $\mathbf{x}_k$ at time k, given the observations $\mathbf{z}_{1:k}$ up to time k. Thus, it is required to calculate the *pdf* $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, and this can be done recursively in two steps, namely prediction and update. The prediction step uses the process equation 7.6 to obtain the prior *pdf* by means of the Chapman-Kolmogorov integral:

$$p\left(\mathbf{x}_k|\mathbf{z}_{1:k-1}\right) = \int p\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right) p\left(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}\right) d\mathbf{x}_{k-1}, \tag{7.8}$$

with $p\left(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}\right)$ known from the previous iteration and $p\left(\mathbf{x}_k|\mathbf{x}_{k-1}\right)$ determined by in equation 7.6. When

a measurement $\mathbf{z}_k$ becomes available, it may be used to update the prior *pdf* via Bayes' rule:

$$p\left(\mathbf{x}_k|\mathbf{z}_{1:k}\right) = \frac{p\left(\mathbf{z}_k|\mathbf{x}_k\right) p\left(\mathbf{x}_k|\mathbf{z}_{1:k-1}\right)}{\int p\left(\mathbf{z}_k|\mathbf{x}_k\right) p\left(\mathbf{x}_k|\mathbf{z}_{1:k-1}\right) d\mathbf{x}_k)}, \tag{7.9}$$

being $p\left(\mathbf{z}_k|\mathbf{x}_k\right)$ the likelihood statistics derived from equation 7.7. However, the posterior *pdf* $p\left(\mathbf{x}_k|\mathbf{z}_{1:k}\right)$ in equation 7.9 can not be computed analytically unless linear-Gaussian models are adopted, in which case the Kalman filter provides the optimal solution.

Particle Filtering is a technique for implementing a recursive Bayesian filter by Monte Carlo (MC) simulations. The posterior density function $p\left(\mathbf{x}_k|\mathbf{z}_{1:k}\right)$ is represented by a set of random samples (particles) with associated weights:

$$p\left(\mathbf{x}_k|\mathbf{z}_{1:k}\right) \approx \sum_{j=1}^{N_s} w_k^j \delta(\mathbf{x}_k - \mathbf{x}_k{}^j), \tag{7.10}$$

where w_k^j are the weights associated to the particles fulfilling $\sum_{j=1}^{N_s} w_k^j = 1$. As the number of samples increases, the characterization of posterior *pdf* improves and the PF approaches the optimal Bayes estimate. The principal steps in the PF algorithm include:

1. *Resample*: the particles are resampled according to their score. This operation results in the same number of particles, but very likely particles are duplicated while unlikely ones are dropped.

2. *Apply motion model*: predict the new set of particles by propagating the resampled set according to a model of the target's motion.

3. *Score*: Form the likelihood function $p\left(\mathbf{z}_k|\mathbf{x}_k\right)$ and weight the new particles according to the likelihood function $w_k^i = p\left(\mathbf{z}_k|\mathbf{x}_k^i\right)$ and normalize so that $\sum_i w_k^i = 1$

4. *Average*: the location of the target is estimated as the weighted sum of all the particles. $E\mathbf{x}_k = \sum_{i=1}^{N} w_k^i * \mathbf{x}_k^i$

7.4.2 Fusion using Particle Filter

The state we wanted to estimate was the position and identity of each person. The propagation model for the particles was modeled by adding white noise to the positions and by randomly changing a given percentage of the identities. Likelihood functions used for scoring were depend on the modality. The proposed solution makes use of a set of decoupled PFs, each per target, and define an interaction model to ensure track coherence as visual tracking. For a given filter, the best state at time k was obtained by computing an histogram of the identities of the particles and selecting the maximum to decide the identity. The mean and the variance for the particle positions were estimated, and depending on the variance, a decision was taken on whether this filter was tracking or not a person.

As video modality, the function $p(\mathbf{z}_k|\mathbf{x}_k)$ was defined as the likelihood of a particle belonging to the position corresponding to a person. For a given particle j occupying a room position, its likelihood may be formulated

by combining the probabilities from POM, FBI, acoustic localization and speaker identification. That is, the acoustic information was added to the visual information by the combination of modalities by means of a weighted sum rule, with weights obtained experimentally on the development recording. In this case the weight for each particle was obtained as:

$$w_k^i = w_1 * (POM * FBI) + w_2 * (AcLoc * SpkId) \tag{7.11}$$

where POM, FBI, AcLoc and SpkId represent the likelihoods of POM, FBI, acoustic localization and speaker identification respectively. As in video identification, the same number of particles were used. That is, the particle filters were initialized with $10,000$ particles distributed randomly across the room each of them characterized by its position and its identity. The number of particles, as well as the other parameters of the PFs, were optimized using a exhaustive search approach using the development recording and the resulting parameters were used blindly on the test set.

7.5 Experiments

The following section is devoted to provide identification and localization results for the different modalities described previously. It is worth to mention that the results from audio and video modality are not directly comparable due to the fact that audio modality was not always available. Nevertheless, the aim of this experiments was to assess the improvement by joining audio and video modalities rather than compare individual performances between them.

7.5.1 Acoustic Identification

The table 7.1 reports the speech detection results in both development (TRAIN) and evaluation data (TEST) in terms of missed speech and false alarms. The recording employed as development data lasted 281.19 seconds, 201.39 of which was speech. The non-detected speech is around 9.65 seconds which represents 4.79 per cent of the total speech whereas false alarm speech is lower, 1.82 seconds which just represent a 0.9 per cent. The total error associated to the speech activity detector (SAD), sum of previous kind of mistakes, is 11.47 seconds, that is, 5.69 per cent of the total speech in the recording. Such a percentages are quite similar in the evaluation set with a total speech/non-speech detection error around 4.71 per cent. These results show the well adaptation of the speech activity detector to the smart-room conditions with percentages of accuracy and detection comparable to those reported in the literature [2].

The metric employed to evaluate the performance of the acoustic identification system was the Diarization Error Rate (DER), see section 2.3.3. It is worth to remember that the SAD is in charge to choose speech frames and to discard non-speech therefore the SAD errors impact directly on the DER. That is, the speaker tracking system will not be able to recover false alarms errors or the missed speech from this stage. The results given in

[2] As example, see speech activity detection performance for the different submitted systems in the NIST RT evaluations [Fiscus and et al., 2007a; Fiscus and et al., 2009a].

Data sets	EVAL TIME	EVAL SPEECH	MISSED SPEECH	FA SPEECH	MS + FA
TRAIN	281.19 s	201.39 s	9.65 s (4.79%)	1.82 s (0.9%)	11.47 s (5.69%)
TEST	422.93 s	371.42 s	10.68 s (2.87%)	6.82 s (1.84%)	17.50 s (4.71%)

Table 7.1: *Speech activity detection performance in both development (TRAIN) and evaluation (TEST) datasets.*

Data sets	SCORED SPK TIME	MISSED SPK TIME	FA SPK TIME	SPK ERROR TIME	DER
TRAIN	201.39 s	15.73 s (7.81%)	1.82 s (0.9%)	22.98 s (11.41%)	20.12%
TEST	371.42 s	14.23 s (3.82%)	6.82 s (1.84%)	95.54 s (25.72%)	31.39%

Table 7.2: *Speaker Identification performance in terms of DER. In addition, the miss speaker time, the false alarm speaker time and the speaker error time are also reported.*

the table 7.2 correspond to the time-weighted DER averages for both datasets decomposed in miss speech, false alarms and speaker error. The difference in the missed speaker time, first column in the table, and the miss speech from the SAD is due to overlap among speakers. It is not significative since the total speaker overlapped time in the database is just some seconds but it yields to an overall increment of the DER over 3 per cent and 1 per cent in development and evaluation respectively. In both datasets, the speaker error time dominates the DER.

The speaker error in development is around 11.41 per cent, over 50 per cent of the total DER, whereas it is around 25 per cent in the evaluation set, over 81 per cent of the total DER. This difference can be mainly attributed to the small size of the development set which drives to an over-fitting of the tuning parameters.

The DER variation w.r.t. the number of segmentation/training iterations is given in the figure 7.8. The curves show an abrupt decrement in DER at the iteration 5 and 3 and they remain unaltered up to reach the stopping criterion. The final decreased of the DER at the last iteration is due to the change (1.5 seconds instead of 3 seconds) in the minimum duration of the speaker turn applied in the last Viterbi segmentation.

7.5.2 Multimodal Identification and Tracking

To have robust tracking and identification of all persons at the same time, we have used a multi-tiered, multimodal approach. In this section we describe this approach. The most reliable component of the system was identified as the face detection and recognition module. For this purpose, the first identification started at the door, where the PTZ camera identified the person. Typically, a single identification result is enough to bootstrap the system, although multiple subsequent identifications can be used in conjunction to bolster the confidence of the system. With the presented setup, we did not see any need for further fusion of results, as we had perfect identification at this tier.

Once the person was identified, one FBI model per camera was created on the fly. We used 3-component

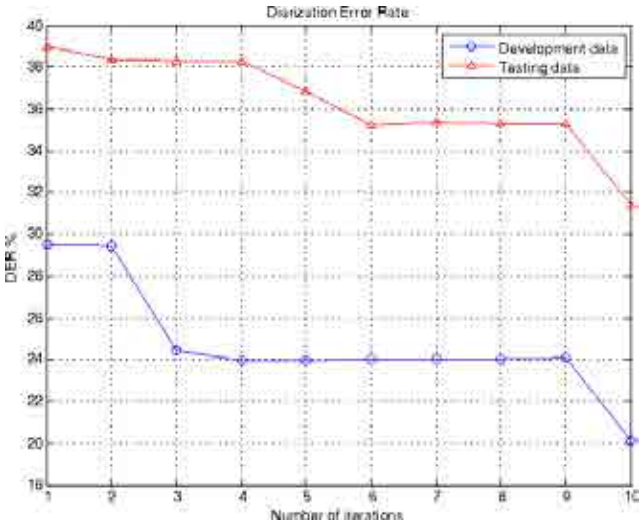


Figure 7.8: *Diarization Error Rate (DER) for training and test datasets. The final decreased of the DER at the last iteration is due to the change (1.5 seconds instead of 3 seconds) in the Minimum Duration of the speaker turn.*

Gaussian mixtures on the HSV color space as FBI models for each person. The POM module provided the motion blobs used for FBI modeling.

The second tier of the approach consists of a joint tracking and identification by the POM and FBI modules. The motion blobs associated with a single location by the POM module are passed on to the FBI modules, which compare the whole collection under each person model. The table 7.3 summarizes the results for visual and multimodal identification. For multimodal identification scoring must take into account the position of the identified target. An identification hypothesis was considered a correct match if the identity corresponded with the one in the ground truth and the detected position of the identified target felt below a given threshold (50 cm). Number of ground truth targets minus correct matches are misses. The number of hypotheses minus the number of correct matches are false positives.

The development set was used to optimize the PF parameters and the weights between visual and acoustic modalities in the multimodal approach. This optimization was performed using a exhaustive search. Once

	Visual			Visual+Acoustic		
Data sets	Matches	Misses	FP	Matches	Misses	FP
TRAIN	81.4%	18.6%	0.0%	83.8%	16.2%	4.2%
TEST	49.3%	50.7%	3.0%	64.5%	35.5%	1.0%

Table 7.3: *Visual identification performance compared to multimodal visual+acoustic identification. For identification scoring must take into account the position of the identified target.*

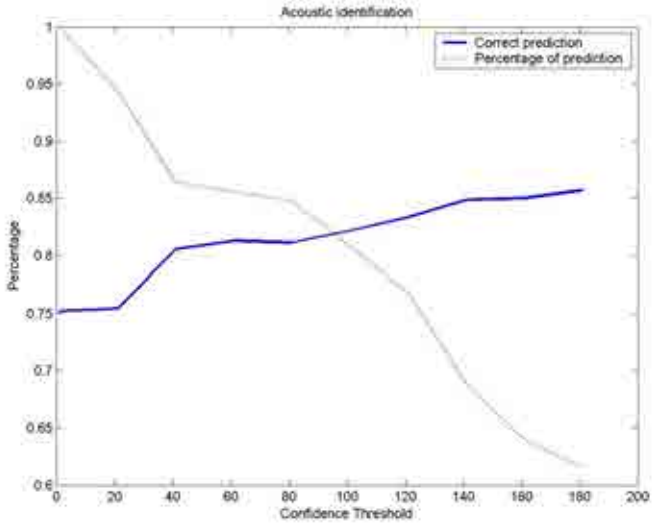


Figure 7.9: *The trade-off between prediction accuracy and prediction frequency for speaker identification.*

these weights were obtained they were used without modification for the test recording. The accuracy of the combined system at the second tier was 81.4 per cent of correct identification, allowing deviations of at most one ground square (approximately 50 cm.) in the localization. The false identification rate is 18.6 per cent. The system sometimes locates occupancy on several adjacent squares, especially if the person is poorly aligned with the ground grid.

The third tier was the addition of the acoustic information. This information was only considered when it was available. The acoustic localization of the sound source was jointly considered with speaker identification output, and used to correct the visual identification results, if it was at conflict. The confidence in acoustic identification was determined by the likelihood ratio of the most likely class to the second most likely class. A threshold on this confidence provided us with a trade-off between prediction accuracy and prediction frequency. The figures 7.9 demonstrates the effect of the threshold. Without the threshold, all acoustic localizations were associated with an identification, and the accuracy of identification was 75.2 per cent. When the threshold was increased, so that the identification decision was given for only 75 per cent of the time, the accuracy increased to 83.3 per cent. In any case, these results are not directly comparable to those from visual identification due to the fact that speech could be unavailable some times during the recording.

The introduction of the third tier did not affect the accuracy for a reasonable confidence threshold, but when the confidence threshold was too low, the accuracy decreased up to one per cent (80.4% instead of the previous 81.4%). The reason is that the visual identification was more accurate for the database we used. Obviously the acoustic information would be much more useful for datasets where the visual information is not as discriminative. The tuning of such a threshold yields to an improvement on the multimodal approach as can be

	Visual		Visual+Acoustic	
Data sets	MOTA	MOTP	MOTA	MOTP
TRAIN	65.4%	337 mm	67.2%	307 mm
TEST	23.0%	291 mm	40.9%	256 mm

Table 7.4: *Visual tracking performance compared to multimodal visual+acoustic tracking in terms of MOTA and MOTP metrics.*

seen in the table 7.3 for both development (TRAIN) and evaluation (TEST) recordings. The huge difference between matches and misses, around 30 per cent, in the visual identification is mainly due to the differences between the two datasets. The FBI module, based on color-feature identification, was highly stressed in the evaluation data in where people wore similar colors of clothes in contrast to development data. Nevertheless, the fusion of visual and acoustic identification outperforms the visual modality in both cases.

7.5.2.1 Multimodal Tracking Assessment

For tracking evaluation, the metrics proposed in [Bernardin *et al.*, 2006] have been used. The two metrics employed are: the Multiple Object Tracking Precision (MOTP), which shows ability of the tracker to estimate precise object positions, and the Multiple Object Tracking Accuracy (MOTA), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories. MOTP scores the average metric error when estimating multiple target centroids, while MOTA evaluates the percentage of frames where targets have been missed, wrongly detected or mismatched. Low MOTP and high MOTA scores are preferred indicating low metric error when estimating multiple target positions and high tracking performance.
Results in the table 7.4 show a very good performance in the development recording. The evaluation recording is much more difficult and a degradation on the performance can be observed. Even though video tracking is good, the audio modalities result into an improvement of the performance and precision of the visual tracking. For the test recording the results show a performance degradation. This recording is much more complex from the video point of view as explained above. The gain using acoustic information is much more important here because the FBI module has problems discriminating the individuals due to the fact that they wear similar clothing. For such cases, adding an additional visual identification modality, such as face recognition, would be of great interest.
In overall, results are similar to the ones obtained for identification, with very good performance on the development set and good performance for the test set. When evaluating the performance we must have in mind that we are presenting a completely automatic detection/tracking/identification system, able to track up to 4 persons at the same time into a room.
While our results are not comparable directly with the ones obtained at the CLEAR evaluation [CLEAR, 2007] because of the different data sets, they are very promising.

7.6 Conclusions

In this chapter we have evaluated several methods for monitoring a room for the purposes of locating and identifying a set of individuals. We worked with different modalities from multiple sensors to observe a single environment and to generate multimodal data streams. These streams were processed with the help of a generic client-server middleware called SmartFlow and signal processing modules. The system was designed to operate in a completely automatic fashion, there was no manual segmentation or user intervention.

Modules for visual motion detection, visual face tracking, visual face identification, visual feature-based identification, audio-based localization, and audio-based identification were implemented. Our proposed system was based on multi-tier information processing, where available modalities were fused into a final decision.

The tracking and identification system that was based on the combination by means a particle filter of the probabilities of the occupancy map, the on-the-fly color feature modeling, the acoustic localization and acoustic identification works effectively. Furthermore, it avoids some of the classical pitfalls faced by continuous tracker systems. The collected database was relatively small, but it represents a realistic application scenario. Both the visual and acoustic channels result in high identification rates and tracking accuracy which are outperformed by the fusion of audio and visual modalities. Such a fusion produce a notable increasing of the overall accuracy in the case where the visual tracking, based on color features, is not well conditioned to the data.

Future research within this topic involve the incorporation of additional modalities. For instance, face identification from the corner cameras would be helpful in maintaining the identity of every moving object in the cases that FBI is not discriminative enough. Real-time implementations of the presented algorithm are under study as several of the components are not computationally intensive. One obvious future direction should be to work with more wide and difficult data (e.g. the CLEAR 2006 and 2007 datasets) in order to strengthen our previous conclusions and to compare the approach with results reported in CLEAR evaluations.

Part IV

Conclusions

Chapter 8

Conclusions and Future Work

8.1 PhD Thesis Final Overview

This PhD thesis presents the work done by the author in the area of Speaker Recognition (SR) focusing on: 1) speaker identification and speaker verification, using techniques to deal with the difficult problem of signal mismatch present in multiple type mics or between meeting-room setups; 2) speaker segmentation techniques in the broadcast news and meeting domain for speaker diarization and tracking tasks and 3) multimodal integration, combination and fusion of multiple cues of information in both previous fields. In overall, successful speaker recognition (SR) applications has been developed and implemented during this thesis work.

In a first line of research, we focused on classical speaker identification and verification tasks. It attempts to deal with the problem of classifying speakers in a room in multi far-field microphone conditions. The GMM classifier is chosen as the basic detection technique in speaker identification and verification tasks. It is worth to mention that two UPC acoustic systems were submitted to both CLEAR 2006 and 2007 evaluations. Furthermore, the acoustic person identification results obtained in CLEAR evaluation 2007 [Luque and Hernando, 2008a] ranked the best among the different approaches submitted by other participants [Stiefelhagen *et al.*, 2007]. When trying to deal with the problem of acoustic person recognition in the framework of the CHIL project [chil, 2006], we soon noticed that GMM based systems performed quite good in same room setup whereas high degradation of correct identification rates is noticed among different mics, sites or speaker accents [Luque *et al.*, 2006b]. The UPC acoustic person identification system developed in the framework of the CHIL project suited the evaluation requirements reaching identification rates over $90\ \%$ in most of the conditions of the CLEAR database, which included the UPC smart-room among others sites. Such a result should be cautiously taken due to the fact of the few number of target speakers (around 30), few sites (5), the manually segmented recordings which ensured a unique speaker per test segment, the absence of long silences or overlapped speech.

Such a findings leaded the research to seek for techniques to deal with mismatch conditions and variability

both in channel and room setups. Despite of Gaussian mixture models and frequency filtering coefficients showed a good performance in meeting room environment, that is in the stand-alone identification evaluation conducted in CLEAR, they have proved insufficient in a more challenging conditions. The development and comparison of state-of-the-art techniques and algorithms in the framework of NIST Speaker Recognition Evaluation (SRE) 2010 [Martin, 2010] is a contribution of this thesis. The work was partially done during a stay in L^{2}F laboratory (from Lisbon, Portugal). Several speaker verification systems were submitted as a joint submission of L^2F and UPC labs. The main objective in participating in SRE evaluation was to introduce ourselves to explore the recently proposed methods and to learn as much as possible about speaker recognition methods in a really challenging evaluation. NIST SRE is the ideal scenario to benchmark speaker verification systems. The evaluation is composed of more than 5 thousand speakers and 1 million of trials just in the core condition. The first participation for both L^{2}F and UPC laboratories at NIST SRE was focused on the development and assessment of SR algorithms and methods. The submitted systems fused complementary information from speaker verification systems based on classifiers as: GMM, SVM, GMM push back, JFA and a newly published algorithm based on Neural Networks adaptation combined with phonetic information from automatic speech recognition. Techniques to deal with speaker and mic variability were also compared, among them: UBM adaptation, Z and T score normalization, NAP compensation or JFA adaptation. Without going too far, most of the ideas aiming to improve speaker tracking and diarization systems which were implemented in this thesis have their origin in classical speaker recognitions tasks. In our tests, the JFA-based techniques show a higher classification capability than both the GMM-based techniques or the SVM-based one, and the best results were consistently obtained with a wide set of speakers and conditions. We also reported that the newly proposed algorithm based on NN-adaptation obtained a comparable performance to SVM and better than the results obtained for the GMMUBM-based system. The collaboration between the two research groups from different countries was a nice achievement and it produced several workshop and international conferences publications [Abad and Luque, 2010; Abad *et al.*, 2010; Abad and Trancoso, 2010; Abad *et al.*, 2011] and we expect more fruitful collaborations in the future.

Other contribution of this thesis is related to speaker segmentation, specifically to speaker diarization in meeting domain. During this process a variety of strategies and algorithms were proposed and benchmarked, once more time, in the framework of national and international evaluations. The speaker diarization system is able to process a variable number of microphones spread around the meeting room and determine the optimum output without any prior knowledge of the number of speakers or their identities. The HMM-GMM classifier and Agglomerative Hierarchical Clustering (AHC) are chosen as the basic detection and clustering technique to perform off-line experiments in the framework of NIST Rich Transcription (RT) evaluations [Fiscus and et al., 2007a; Fiscus and et al., 2009a]. The AHC-HMM based diarization system, previously developed for Xavier Anguera [Anguera, 2006], was adopted as starting point in the research. As the previous work, the system makes use of multiple channel for channel beamforming and to retrieve speaker position information in order to perform diarization in a single channel. The performance of this approach was benchmarked in a wide database composed of 24 meeting recordings each of them lasting 25 minutes and with multiple speaker interactions,

that is, the NIST Rich Transcription (RT) 2006, 2007 and 2009 evaluations. Moreover, the development of the system was closely linked to participation in the speaker diarization evaluations in RT for meetings proposed by NIST in 2007 and 2009. In both submissions the systems proposed by UPC for conference room data, with various numbers of microphones, obtained consistently good results. In the experiments performed we identify the main sources of problems which mainly contribute to speaker errors. We focused our research in the Speech Activity Detection (SAD) influence on the diarization errors, in cluster initialization strategies and in the speaker overlap detection and handling. Such an issues are shown as key points in the improvement of the global diarization. We have also described several algorithms, we tune different algorithm's parameters and we assessed its performance in the NIST RT'06-09 conference databases, see appendix B. Furthermore, a couple of systems were submitted to both RT 2007 and 2009 evaluations [Luque *et al.*, 2008a; Fiscus and et al., 2009b] and we developed newly algorithms and strategies as initial clustering based on tracking of TDOA features [Luque *et al.*, 2008b], HMM-based overlap detection/handling [Zelenák *et al.*, 2011] and language modeling based on frequency and duration of speaker n-grams.

It is worth to highlight the newly proposed speaker diarization system based on spectral clustering theory. The system still relies on HMM segmentation but switches classical BIC pair computation by euclidean distances in a transformed space spawned by the means of the Gaussians estimated from clusters. The presented experiments showed a comparable performance with respect to the AHC HMM/GMM diarization based on BIC metric whereas reduce computational cost of the AHC approach. The algorithm was presented during the last Speaker and Language Recognition Workshop [Luque and Hernando, 2012].

The adaptation of the AHC HMM/GMM-based system to perform speaker tracking in Spanish radio broadcast news and to carry out speaker diarization in Catalan TV broadcast news was also reported in this work. New strategies and implementations were proposed to bridge the gap between speaker diarization and speaker tracking resulting in good performances in terms of DER per cent. The works were performed in the framework of the speaker tracking Evaluation Albayzin 2006 [Segundo, 2006] and the Catalan founded project Tecnoparla. The adaptation to tracking and broadcast news was accomplished into two main research lines: 1) by using techniques and algorithms coming from the field of speaker verification like as score normalization, UBM normalization, Gaussian pruning, maximum a posteriori adaptation (MAP) and so forth and 2) by combining audio segmentation with speaker diarization. In the first work, the tracking system based on AHC obtained similar results than the step-by-step approach based on BIC metric segmentation and speaker verification techniques. In the second work, the detection of channel and background conditions is reported as a potential cue to improve global diarization in broadcast news data. Speaker diarization adaptation to background condition as telephonic or studio channels has been shown as a possible way for improvement in such a scenario. The proposed strategy which combines audio segmentation with speaker diarization based on channel detection resulted in improved results parallel speaker diarization dependent on each background condition is a promising approach to deal with mismatch conditions likely to occur in broadcast news data. The results reported evidences a well adaptation of both the audio segmentation and the speaker diarization to the Àgora database (more than 24 hours) and comparable to other results reported in the literature [Fiscus and et al.,

2007b].

Most significant improvements in the speaker recognition systems presented in this work are reached by the fusion, integration and combination of different cues of information. It has been highlighted across the different chapter of this PhD thesis. The fusion with face identification information, when it is available, allows us to increase performance of the stand-alone acoustic identification as is reported in chapters 2 and 7. Same occurs in the case of simple decision strategies for fusing acoustic system working on different input channels or for fusing complementary systems as was reported in SRE evaluation. The combination of speaker verification with Automatic Speech Recognition (ASR) transcripts is given in chapter 4 by incorporating ASR information to the verification system based on NN adaptation. Moreover, it contributes to the global verification system complementing the other speaker verification subsystems. In chapter 6, audio segmentation is also combined with speaker diarization aiming to improve speaker detection in telephonic channel. Finally, in last chapter 7, we monitor the smart-room for the purposes of locating and identifying – speaker tracking – a set of individuals. A database was collected at the UPC smart-room for such a purpose, relatively small but representing a realistic application scenario. We worked with different modalities from multiple sensors to observe a single environment and to generate multimodal data streams. The system was designed to operate in a completely automatic fashion, there was no manual segmentation or user intervention. Modules for visual motion detection, visual face tracking, visual face identification, visual feature-based identification, audio-based localization, and audio-based identification were implemented. The proposed system was based on multi-tier information processing, where available modalities were fused into a final decision by means a particle filter algorithm. Both visual and acoustic channels result in high identification rates and high tracking accuracy which are outperformed by the fusion of audio and visual modalities [Salah *et al.*, 2008b].

Finally, this PhD thesis has pro actively contributed to provide the tools and resources that make the research possible, including: making recordings and labeling of databases, making available databases by participating in technology evaluations and providing support to students and colleagues into the smart-room. The real-time implementation of a speaker verification system was also carried out and it is currently running in real time in the UPC smart-room. The participation in technology demonstrations has also continuously been linked to this work, see appendix A. Besides, the developed real-time SR component contributed to various demonstrations of technologies and services developed within CHIL, SAPIRE and SARAI projects.

8.2 Future Research Lines

Future research within this topic involve research in several speaker recognition fields and into information fusion and multimodality topics. As it has been shown during this thesis work, speaker identification and verification techniques might help to increase the discriminate ability of acoustic based diarization and tracking systems in order to distinguish among speakers in a wide variety of situations and conditions. In the same way, the fusion and incorporation of additional information and modalities drastically improves acoustic recognition rates and allows for a continuous identification. For instance, the combination with speaker localization allows

to improve diarization by linking positions to speakers in meetings; speaker tracking along with acoustic segmentation enables the detection of different channel conditions in broadcast news, discarding music and commercials and giving the possibility to improve speaker detection by prior adaptation to such a conditions. Classification of sounds and acoustic segmentation has usually been carried out so far to segment digital audio streams using a limited number of categories, like music/speech/silence/environmental sound. Inclusion of a variety of event categories, i.e. laugh and cough detection, might benefit diarization performance by creating "pure" initial clusters in the same way the inclusion of an overlapped speech detector for speech exclusion reduce speaker errors. In the case of multimodality, face identification technology applied from corner cameras covering the whole space into the smart-room would be helpful in maintaining the identity of every moving object in the cases that acoustic identification is not discriminative enough or even not present. The inclusion of other technologies, like as reliable head orientation information, might also improve robustness, i.e. by assuming that an interrupting speaker will draw attention by some of the participants.

Nowadays, one of the main drawbacks of current speaker detection systems are the issues with robustness and lack of flakiness. They are problems present in most of the speaker diarization systems in the literature. The modeling of speaker variability and channel variability as in JFA or NAP techniques could improve discrimination among speakers into the same smart-room or between different sites by taking benefit of multi microphone setup. The experiments reported in the speaker verification task point towards including verification techniques in order to still reduce more the DER error by reducing SER. The speaker diarization and tracking systems might be significantly improved by increasing its robustness through these speaker modeling techniques. Speaker verification techniques as normalization of adaptation might also improve clustering by reducing errors due to linking and discrimination errors, which represents most of the SER. These strategies will allow easy adaptation to new domains and parameters should be robust and not flaky. Anyway, dealing with the *flakiness* problem should be among the priorities of future research.

Furthermore, different proposals for merging clusters should be take into account aiming to improve the number of speakers detected and indeed the quality of the speaker diarization hypothesis. This is the idea behind new proposed algorithms as spectral clustering. Several other strategies should be taken into account to improve speaker diarization and tracking systems in future works. Among them: The use of a gender and language detectors might improve the system by performing parallel analysis specifically adapted to characteristic of the speakers, the application of other speaker features as prosodic or pitch features and other speaker related characteristics and statistics related to phonemes, spoken words, pauses or silences.

In terms of easy adaptation, the system was developed using separate blocks but the core of the speaker detection module was kept very similar in the speaker tracking system. It allowed easy recombination of modules and quick adaptation to the requirements of each task. In this way, modules can be easily employed or discarded like as the use of the speech activity detection technology, the clustering initialization based on TDOAs dynamics or the application of the overlapped speech detection module.

Related to real-time implementations of presented algorithms, it is under study since several of the components are not computationally intensive. Initial steps forward the adaptation of the speaker tracking system to real

time processing requirements should be: Restrict the complexity of the GMM modeling or take advantage of faster scoring algorithms as SVM and JFA, adjust the number of training and segmentation iterations or to apply an analysis window with optimal size instead of the whole recording.

”¿Has acabado la tesis?”
– Bartolomé Luque

Appendices

Appendix A

Development into UPC's Smart-Room

Information technology has penetrated virtually all forms of communication an it has profoundly affected the way we live our lives. It has been observed that human-human interaction is what humans enjoy and prefer, and human-machine interaction is generally viewed as a chore and necessary evil to access the benefits that computing can bring. Projects working on the human-computer interactions set as theirs goals the development of systems that respond proactively to the needs of their human users, without requiring peoples constant attention. Such systems must be aware of their users' current activities, whether they are in a meeting, holding a seminar, or drafting a document, in order to provide information and services.
In this appendix the activities concerning Speaker Detection (SD) in the UPC's smart-room are described. It gives basic clues on the implementation of the speaker ID component into the smart-room. We briefly describe two different demonstrations in which speaker ID component allows computer awareness about people identity into the room and enables personalized services.

UPC's Smart Room

The Speech research group owns, jointly with the Image Processing Group (GPI) a smart multimodal room, that has been constructed and equipped in a room of the Department of Signal Theory and Communications, in relation to the integrated european project CHIL, in which both groups participate as a unique partner. The group is currently involved in several research activities in the area of speech and audio processing, and has used the room as a laboratory for testing techniques and collecting real data.
Based on the perception and understanding of human activities and social context, a new type of context aware and proactive services can be developed. Within the years of the CHIL project, four instantiations of such CHIL services have been implemented:

- *The connector*: This service attempts to connect people at the best time by the best media, whenever it is most opportune to connect them. In lieu of leaving streams of voice messages and playing phone tag, the Connector tracks and knows its masters activities, preoccupations and their relative social relationships

and mediates a proper connection at the right time between them.

- *The memory jog*: This is a personal assistant that helps its human user remember and retrieve needed facts about the world and people around him/her. By recognizing people, spaces and activities around its master, the memory jog can retrieve names and affiliations of other members in a group. It provides past records of previous encounters and interactions, and retrieves information relevant to the meeting.

- *Socially supportive workspaces*: This service supports human gathering. It offers meeting assistants that track and summarize human interactions in lectures, meetings and office interactions, and provide automatic minutes and create browseable records of past events.

- *The attention cockpit*: This agent tracks the attention of an audience and provides feedback to a lecturer or speaker. CHIL represents a vision of the future - a new approach to more supportive and less burden some computing and communication services. The research consortium includes 15 leading research laboratories from 9 countries representing todays state of the art in multimodal and perceptual user interface technologies in European Union and the US. The team sets out to study the technical, social and ethical questions that will enable this next generation of computing in a responsible manner.

As commented previously, for the purposes of person identification (PID) a database of target speakers was designed and collected at the UPC and at other smart-rooms from CHIL partners, which were publicly disseminated by the European Language Resources Association. This database has been used as a training material and as a testing material to evaluate algorithm performance of the PID component and its online implementation as a smartflow component. Algorithm development was assessed on data from the CLEAR evaluation campaigns [Mostefa and et al., 2006] and performance results can be checked in chapter 3.
Apart from the CLEAR database and aiming to develop an speaker recognition application, working on real time conditions, a new database, composed of members of UPC and TALP Research Center, was recorded during several years. Therefore, algorithm adaptation to specifically application conditions, such as target speakers from UPC, was accomplished by recording individually each UPC participant at different dates. The data was collected at the UPC smart room, in same accordance with the "CHIL Room Setup" specifications [Casas and Stiefelhagen, 2005]. Recordings were performed at different dates (several months apart) to ensure proper variability of the participants. A total of 43 person voices were recorded during three different sessions. The details of the recording protocol as well as a the session information can be found in Appendix A. Recorded data has been used to estimate UBM-GMM target speaker models from UPC smart-room's users as well as for T-norm score normalization.
The figure A.1 depicts a brief description of the UPC smart-room sensors and space conditions. The room has an entrance door, a big window and a table in the middle. The window was closed during recordings to avoid illumination changes. However, small illumination changes can occur when the door is opened. The room is monitored using six cameras: four fixed cameras at the corners of the room (labeled Cam1 to Cam4 in figure A.1), one zenithal fish-eye camera at the ceiling (Cam5) and one active camera (PTZ) aimed and zoomed at

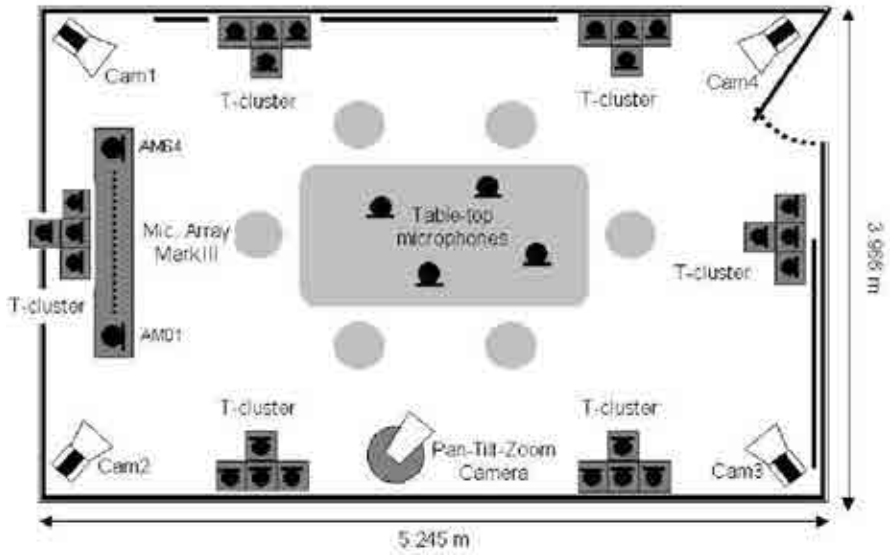


Figure A.1: *The UPC smart room setup. A variety of sensors from audio and video modalities can be found:* 6 *T-shape omni-directional wall mounted microphones (a total of* 24 *mics),* 4 *directional table-top and* 4 *omni-directional table-top mics, a* 64 *mic MarkIII array,* 4 *high resolution cameras at each room corner and a pan-tilt-zoom camera. During each recorded seminar, participants also wore a lavalier microphone which was mainly used for annotation and synchronization purposes.*

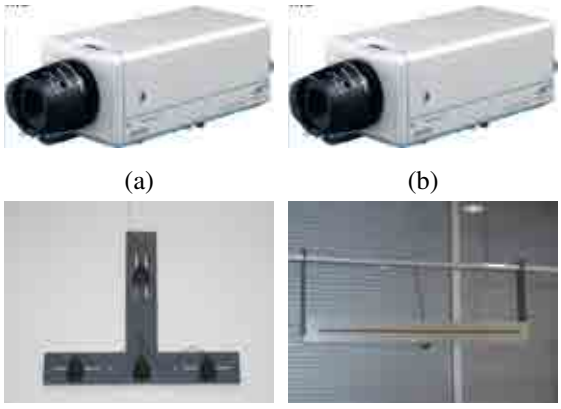
the entry door to capture the faces of the incoming people at high resolution. Video is interlaced, recorded in compressed JPEG format, at 25 fps and 768×576 resolution.

The audio sensor setup is composed by one NIST Mark III 64-channel microphone array, six T-shaped four-channel microphone clusters and eight tabletop microphones. Audio is recorded in separate channels in *wav* format, at 44.100 sampling frequency. Far-field conditions have been used for both audio and video modalities. All data flows are timestamped and the computers used to record the signals are synchronized using the network time protocol (NTP). This makes possible to synchronize audio and video data. There is no manual segmentation of the data. Each technology is supposed to automatically segment the recorded signal. The figure A.3 shows a sample set of recordings from the room setup.

The Middleware

The problem of interconnecting several algorithms that work on data streams coming from a high number of sensors from different modalities is far from trivial. Synchronization of the different data flows, distributed computing and the interconnection of the algorithms are issues that need to be addressed.

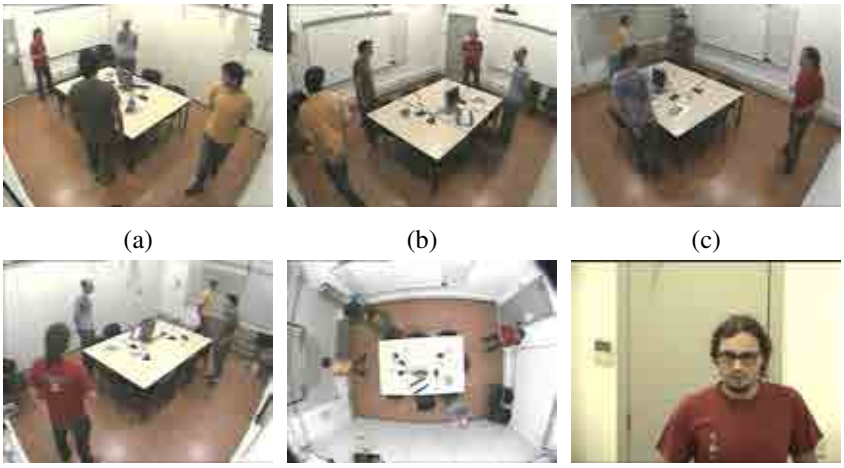
To allow efficient communication of sensor data and distributed computation, it is useful to have a middleware that provides infrastructure services. We propose to use the NIST Smartflow system that allows the transportation of large amounts of data from sensors to recognition algorithms running on distributed, networked nodes [smartflow, 2002; darpa, 1998]. The working installations of Smartflow is reportedly able to support hundreds



(a) (b)

(d) (e)

Figure A.2: *Sample sensor devices: video and audio.(a) A corner ceiling camera and (b) A Pan Tilt Zoom (PTZ) mobile camera. (c) T-shape wall-mounted microphone array. (d) The NIST MarkIII array.*



(a) (b) (c)

(d) (e) (f)

Figure A.3: *Sample camera recordings.(a)-(d) Four corner ceiling cameras. (e) Center ceiling camera. (f) Door camera.*

of sensors [Stanford *et al.*, 2003].
Smartflow offers a great deal of data encapsulation for the processing blocks, which are called "clients". Sensor data are captured by clients, cast into a standard format, and a flow for each data stream is initiated. The processing blocks are themselves Smartflow clients, and they can subscribe to one or more flows to receive the data required for their processing. Each client can output one or more flows for the benefit of other clients. The communication over TCP/IP sockets is transparent to the user, and handled by the middleware. The design of a working system is realized through a graphical user interface, where clients are depicted as blocks and flows as connections. The user can drag and drop client blocks onto a map, connect the clients via flows, and activate these processing blocks.
The synchronization of the clients is achieved by synchronizing the time for each driving computer, and timestamping the flows. The network time protocol (NTP) is used to synchronize clients with servers' time, and this functionality is provided by Smartflow.

Speaker Identification Component Implementation

Our system, which is described in the section 3.2, is written in C++ programming language and is a part of the smartAudio++ software package developed at UPC which includes other audio technology components (such as speech activity detection, acoustic source localization, and event detection) for the purpose of real-time speaker detection and observation in the smart-room environment.
The software architecture chosen in the UPC's smart-room is based on NIST smartflow system [smartflow, 2002] and a socket messaging system (know as KSC). The lower level of the software architecture consists of the video and audio sensors. The signal capture software is implemented as smartflow clients in the computers with the corresponding acquisition hardware. The resulting data streams are transferred as smartflow flows into other computers that can either pre-process the data streams or directly analyze the raw data streams (as in the case of the speech activity detection audio technology). Smartflow also provides a mechanism to dynamically decide on which computer in the local area network a specific technology should run. The KSC message server and the KSC client library allow sending results of data analysis asynchronously.

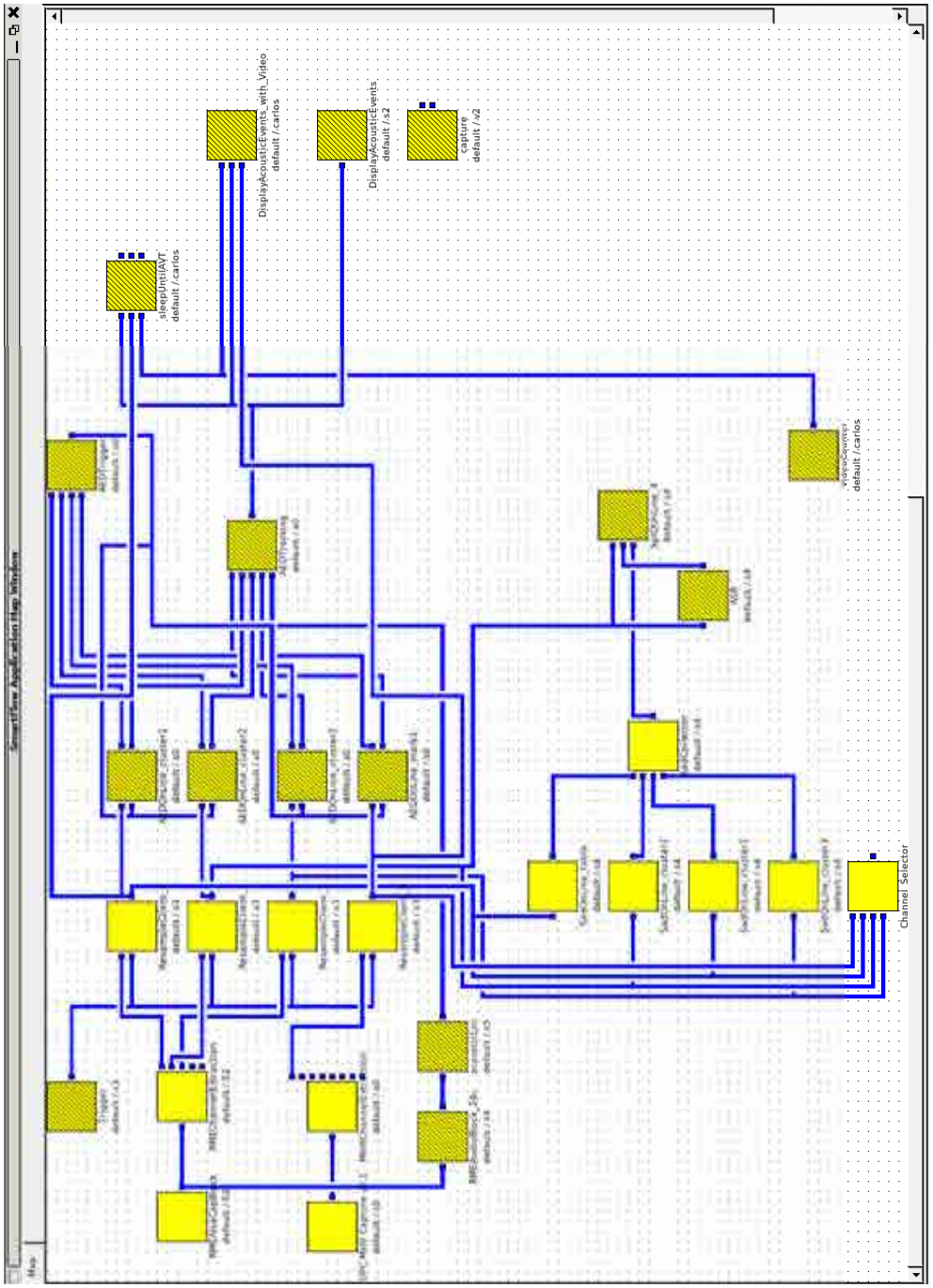


Figure A.4: *Smartflow audio map for demonstration of several acoustic technologies. Among audio technologies: Acoustic event detection (AED), acoustic source localization (SLOC), speech activity detection (SAD), automatic speech recognition (ASR) and speaker identification (SI). In the case of speaker identification, the SI inputs consist of three flows: speech data samples, dialogue state and speech/non-speech detection state. The ASR and SAD outputs are applied in order to segment speech. SI client just performs person identification whether speech exists and there is a dialogue with CHIL computer to request a service.*

The figure A.4 shows the smartAudio map that corresponds to the UPC system technologies already implemented at the UPC smart-room. In addition to speaker identification, other acoustic technologies has been developed during last years. Among them: Acoustic event detection (AED), acoustic source localization (SLOC), speech activity detection (SAD), automatic speech recognition (ASR), and speaker identification (SID). The map shows the various smartflow clients and the interconnections among them. Firstly, an audio signal from cluster microphones and the MarkIII microphone array are captured with data acquisition clients RMEAlsaCapBlock and UPCMarkIIICapture, respectively. Three cluster microphones (one from each cluster) and a channel of the MarkIII are extracted with the RMEChannelExtractors and the MarkIIIChannelExtractor, respectively. Such audio channels are used to feed AED, SAD, ASR and SID clients.
Following, Resample-Clients are in charge of downsampling signal to 16 kHz. In the case of SLOC client, all the T-shape wall-mounted microphones, extracted by the client RMEAudioBlock_24c, are used to estimate the location of the acoustic source. In the case of AED and SAD technologies, a fusion client was implemented to merge the decisions of 4 individuals algorithms working independently in each audio channel selected previously. For ASR and SID technologies just one audio channel is applied, though a channel selector based on SNR is in charge to select the channel with highest SNR value.
Depending on the demo, in which the SID perceptual component is involved, there can be components shared among several clients, for instance data acquisition or resample clients, and clients responsible for results visualization. In addition, SID client versions use several strategies to segment speech: a non-stop SID, which gives speakers labels every N seconds, a SID which performs identification on speech frames detected by SAD client and a ASR combination which tries to recognize people that are in a concrete dialogue state.

Speaker Identification in CHIL demonstration

Several demonstrations has been conducted in the UPC smart-room during the elaboration of this PhD. dissertation. Among them, it is worth to mention the integration of the SID component together with other technologies in the mockup demonstration [Casas and Neumann, 2007], within the framework of the CHIL project, and the effort conducted to integrate it in an intelligent manner with speech/non-speech detection, automatic speech recognition and localization technologies into a smart-room, within the framework of the Spanish founded project Sapire. Demonstrations into the UPC smart-room have become an efficient but subjective way for assessing quality and performance of the technologies involved as well as a never ending source of new troubles to solve which usually there not exists in controlled recordings from other databases. Real-time constrain, overlapping of events and speech from various speakers, speech style, speaker and session variability are some of them from a long list.
The aim of the mockup demonstration [Casas and Neumann, 2007] designed at UPC is to gain context awareness in the framework of the CHIL (Computers in the Human Interaction Loop, IP506909) memory jog assistant. It is achieved by means of detection of people, objects, events, and situations in the interaction scene. Some effort has been made in the last years in the design of communication systems that respond proactively to the needs of their human users, without requiring peoples constant, undivided attention. However, in order

Figure A.5: *Service example: The journalists service provide the front page edition of the UPC newspaper by joining the news from a field journalist and journalists in the office (smart-room).*

to bring the new communication technologies near to real world, it is still required the development of more intelligent systems with perceptive functionalities based on non-intrusive sensors providing visual and auditive capacity, that is, that they can detect and robustly adjust to new and varied environments, and accommodate and adapt to individual user preferences and requirements in a minimally disruptive manner. With this purpose, this project developed technologies that observe and model human activity and communication, and that use such models to provide a family of non-disruptive multi-modal user services, for human memory support and proactive telecommunication assistance. Substantial progress has been made in the component technologies required for the automatic perception of the activities, intentions, and needs of human subjects.

The information needed to build the relevant context awareness and computer's cognition stems from the analysis of the signals acquired in real-time from a collection of sensors. Specifically, the journalist service developed at UPC focuses at providing information to a group of newspaper journalists gathered together in the CHIL smart-room. Within ten minutes the front page of tomorrows edition of their newspaper has to be decoded. One of the most outstanding means of the journalist service to interact with the journalists is a talking head. It is depicted in figure A.5. Talking head responds to a few commands, looks to the speaker or where activity is detected and responds to a few events. In addition, talking head not only informs the journalists about available resources, and points out events such as the arrival of a latecomer or news being contributed by remote colleagues (by means text-to-speech technology), but also facilitates information requests from the journalists in a human-like interface based on automatic speech recognition technologies.

A real-time video stream coming from one of the cameras of the meeting room is used to show audience what is happening at the room. An automatic cameraman is choosing the optimal camera from five possible angles. This decision is based on the location of persons and the last speech or acoustic event in the room and it is smoothed by a hysteresis to avoid rapid camera changes. A person of interest (e.g. the latecomer) can be tracked and identified in the room. This location is used to direct the talking head and automatic cameraman to his current position. The real-time video streaming also displays annotations in the form of subtitles that explain the situation, e.g., "people enter", "Journalist Ramon detected", "interaction with ASR", "keyboard typewriting", "front page published", "The meeting has started", by means the position of all participants and

Figure A.6: *Journalist service: Screen-shot of the field journalist's laptop.*

person identification technologies, e.g. between the states people enter, meeting starts or coffee break. On the left side of the screen, a graphical user interface allows the field journalist to add a piece of news (a test and an image) to share with the journalists in the room.

The dialogue system allows a human-like verbal interaction with the computer's service. It is based on two components: a commercially available 2D animation of a talking head [Haptek, 2011] and an ASR based dialogue system that utilizes the ATK recognizer, an on-line API for the well known HTK speech recognition toolkit [Young *et al.*, 1993]. Interactive behavior of the talking head depends upon the latecomer detection or an acoustic event. Such interaction is expressed like an utterance Dont forget your keys! when key jingle is detected, by exclamation: Great! Well done! when applause is detected or by welcome the latecomer using his name.

In the service implemented at the UPC's smart-room, context awareness consists of knowledge about the number of persons in the room, their identification, position in the room and their orientation. Objects in the room and acoustic events also add information to the context awareness. It is worth to mention that when humans experience the computer-driven service, another subjective bias naturally arises: unexpected actions of the service triggered by a false-positive detection of one of the technologies turn out to be far more annoying than a service not provided due to false-negative detection.

The combination of video-based and audio-based systems allows computer to gain a basic understanding of what happens in the smart-room. Perceptual components are computing modules that analyze the signals provided by the network of sensors in order to detect and classify objects of interest, persons and events adding information to context awareness. In total 8 perceptual components which are based on more than 40 smartflow clients, are integrated into the single application called central logic which fuses the information

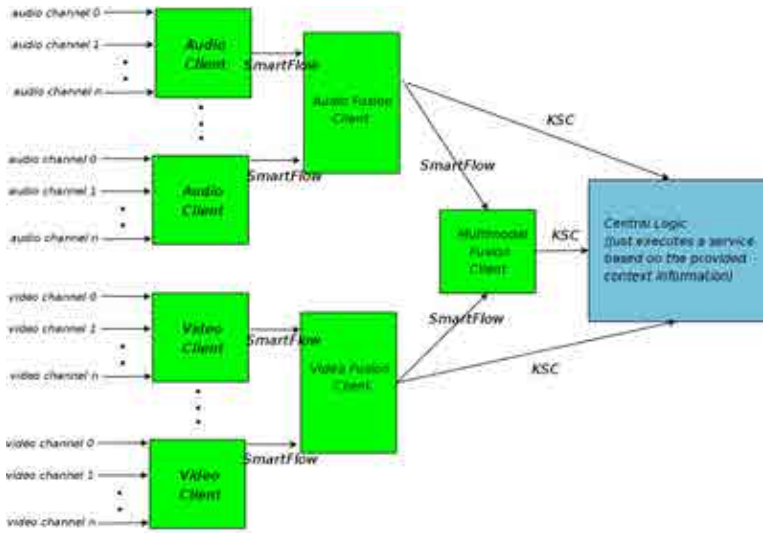


Figure A.7: *Sapire data flow schematic in smartflow.*

coming from the several technologies, creating a virtualization of the environment which brings cognition to computer and allows interaction with humans.

Speaker Identification in Sapire project

Sapire project (TEC2007-65470) seeks to work on a system of acoustic scene and human-to human communication analysis, both verbal and non-verbal, which shows a number of perceptual and cognitive functionalities, doing research in the speech and audio technologies that make them possible: speaker identification, speech recognition, acoustic source localization, sound detection and classification, head pose and emotion estimation, etc. The project hunted for continuing the development, started in CHIL, aiming of creating the technological foundation required to bring the multimodal technologies more near to the real world. In this way, multimodal perception and situation understanding technologies developed under CHIL were extended, so that they could:

- Robustly detect and adapt and adjust to new and varied environments

- Accommodate and adapt to individual user preferences and requirements in a minimally disruptive manner.

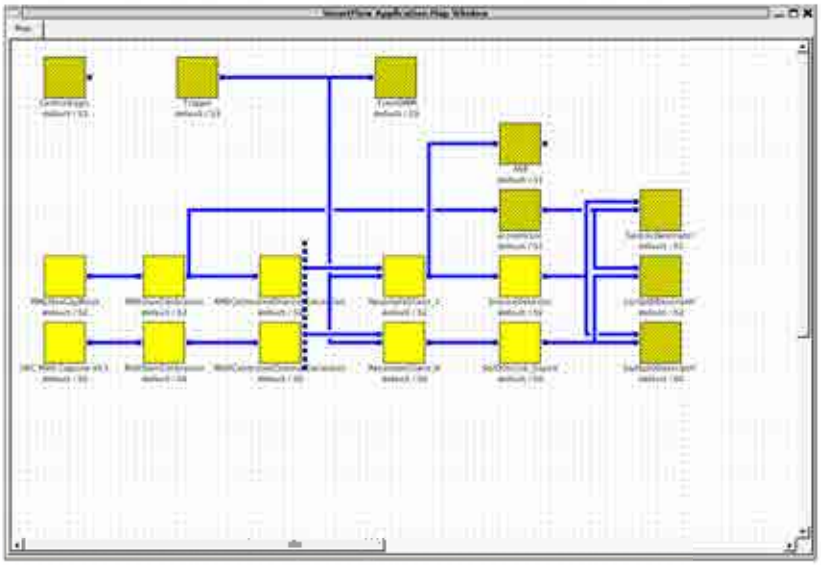


Figure A.8: *Audio map for demonstration in Spanish Sapire project (TEC2007-65470). Among audio technologies: Acoustic event detection (AED), acoustic source localization (SLOC), speech activity detection (SAD), automatic speech recognition (ASR) and speaker identification (SI). In the case of speaker identification, the SI inputs consist of three flows: speech data samples, dialogue state and speech/non-speech detection state. The ASR and SAD outputs are applied in order to segment speech. SI client just performs person identification whether speech exists and there is a dialogue with CHIL computer to request a service.*

Project Sapire wants to augment services with an ability to initiate brief, optimized direct interaction to learn progressively abstract information about objectives and preferences. Main objective includes the research on several technologies of speech and audio processing for the development of functionalities in the environment of the smart multimodal room of the UPC taking as framework, activities related to presentations, meetings or seminar-like courses. We seek to detect of the presence and position, either spacial and/or temporal, of the acoustic events that take place in a given environment. Once the acoustic event has been detected (Acoustic Event Detection), and localized (Acoustic Source Localization), this event has to be classified. A first level of classification specially important due to its direct use by the speech recognition system- is the distiction between presence or absence of speech (Speech Activity Detection). In a second level, the aim is determining the specific kind of sound (Acoustic Event Classification), or speaker (Speaker Identification), obtaining a temporal segmentation of each audio signal corresponding to the characteristics of each different source. At

the same time, the segments classified as speech have to be recognized (Speech Recognition). Moreover, it would be useful the detection of head-pose (Speaker Head Orientation) or the perception of affect (Emotion Recognition). The project aims at the three concrete objectives:

- The development of perceptive functionalities in the environment of a smart multimodal room, where teaching activities like presentations or seminars are carried out. The functionalities answer questions about what is happening in the room: Who is talking? What is he saying? Where is the speaker?, etc.
- Research on robust speech&audio processing technologies which allow functionalities such as detection, recognition, localization, separation, segmentation, etc.
- Coordination with the image&video processing activities, looking for the multimodal integration and fusion of audio and video technologies.

SAPIRE will strongly emphasize real world issues, articulating them in a number of subtopics: the notion of technology appropriation, which addresses how real users adopt the technologies, and adapt them to their needs; the balance between implicit and explicit interaction modes, the new challenges that adapting to groups and teams dynamics (groupalization) arise with respect to the better-studied individual-oriented adaptation (personalization); robustness and scalability issues; the supports for the development, integration and deployment of real world services, in the form of tools and software infrastructure.
The range of interaction scenarios considered in SAPIRE presents significant challenges to the development of technologies required for the envisioned services. We propose to achieve this goal by focusing on the following two themes: adaptability and multimodality of the designed components.
Adaptability: The perceptual and communication technologies should be able to characterize, remember, and react to the variability factors encountered: environmental, sensory, individual, group patterns.
Multimodality: The perceptual and communication technologies should be able to utilize all or an appropriate subset of the available modalities to accommodate the requirements of the envisioned services.

Appendix B

UPC-TALP Database of Speakers for Recognition in Smart-Room

Last update: May 12, 2011

DDBB 2006: CT mic = 020
DDBB 2007: CT mic = 021
DDBB 2009: no CT mic

(1) Without MarkIII recordings

Count	Person Unique ID	Gender	DDBB 2006	DDBB 2007	DDBB 2009	DDBB 2011
1	UPC_000	M	-	x	-	
2	UPC_001	M	-	x	-	
3	UPC_002	M	January 12 06	x	-	
4	UPC_003	M	-	x	-	
5	UPC_004	M	July 20	x	-	
6	UPC_005	M	June 1, July 20	x	-	
7	UPC_006	M	-	x	-	

Count	Person Unique ID	Gender	DDBB 2006	DDBB 2007	DDBB 2009	DDBB 2011
8	UPC_007	M	-	-	x	
9	UPC_008	M	-	-	x	
10	UPC_009	M	-	-	-	
11	UPC_010	M	January 12 06	x	x	
12	UPC_011	M	-	-	x	
13	UPC_012	F	-	-	x	
14	UPC_013	M	January 12 06	x	-	
15	UPC_014	M	June 1	x	-	
16	UPC_015	M	-	x	-	
17	UPC_016	M	-	x	-	
18	UPC_017	M	-	x	-	
19	UPC_018	M	July22, July 27	x	-	
20	UPC_019	M	-	x (1)	-	
21	UPC_020	M	-	x	-	
22	UPC_021	M	-	x	-	
23	UPC_022	M	-	x	-	
24	UPC_023	F	July22	x	-	
25	UPC_024	M	-	x	-	
26	UPC_025	F	-	x	-	
27	UPC_026	M	December 20	-	-	
28	UPC_027	F	December 20	-	-	
29	UPC_028	M	June 1, July 6	-	-	
30	UPC_029	F	July 6	-	-	
31	UPC_030	M	July 6	-	-	

Count	Person Unique ID	Gender	DDBB 2006	DDBB 2007	DDBB 2009	DDBB 2011
32	UPC_031	M	July 6, July 27	-	-	
33	UPC_032	M	July 6	-	-	
34	UPC_033	M	July 20	-	-	
35	UPC_034	F	July 20	-	-	
36	UPC_035	M	July 20	-	-	
37	UPC_036	F	July 22	-	-	
38	UPC_037	M	July 22	-	-	
39	UPC_038	M	July 22	-	-	
40	UPC_039	F	July 27	-	-	
41	UPC_40	M	July 27	-	-	
42	UPC_41	M	July 27	-	-	
43	UPC_42	M	January 12 06	-	-	

Appendix C

NIST Rich Transcription Database

The Rich transcription evaluations conducted by NIST started with the RT02s in 2002 until the latest one in 2009. According to NIST Rich Transcription Meeting Recognition Evaluation Plan, the Rich Transcription (RT) of a spoken document addresses the need for information other than the set of words that have seen said (extracted with a Speech-to-Text, STT, system). When obtaining a transcription of the words that have been spoken in a recording it is difficult to receive all the information that the speakers tried to convey. This is because spoken language is much more than just the spoken words; it contains information about the speakers, prosodic cues and intend, and much more. The goal of future RT systems is for transcripts to be created with all sorts of metadata to allow the user to fully understand the content of an audio recording without listening to it. In the recent RT evaluations NIST has focused on three core technologies that are important elements of the metadata content. These are Speech-to-Text (STT), Speaker Diarization (SPKR) and Speech Activity Detection (SAD)[1]. In the last years (RT05s, RT06s, RT07s and RT09s) evaluations have mainly been focusing on the meetings domain.
First three databases are similar in that two different subdomains were proposed, with different microphone configurations within each subdomain whereas RT09s was focused on the conference subdomain. All systems were allowed to run with unlimited runtime speed so that they could be comparable within the same metrics. The speed of each system was reported as part of the system description. In brief, the two proposed subdomains were:

- *Conference room meetings*: These are conducted around a meetings table with several participants involved in an active conversation among them. It contains various amounts of speaker overlap (depending on the nature of the meeting). These have been the focus of research of several projects including the European AMI project.

[1] Speech Activity Detection Evaluation was eliminated from RT07s and RT09s.

- *Lecture room meetings*: These are conducted in a lecture setting where a lecturer gives a presentation in front of an audience, which normally interrupts with questions during the talk. In these meetings the lecturer normally speaks for most of the time during the talk, and it becomes more balanced during question and answer sections. It has been the focus of research of the European CHIL project.

In each one of the meeting rooms there are multiple microphones available which record the signal synchronously. In some settings there are also cameras, but these fall outside of the scope of the speaker diarization evaluation. The microphones are clustered in different groups to determine different conditions/evaluation subtasks. The following list points out the terminology used for each of the possible groups and whether it is used in the speaker diarization evaluation and in which domain:

- *SDM (Single Distant Microphone)*: This is defined as one of the centrally located microphones in the room, located on the meetings table. This microphone is always part of the bigger MDM group. Both lecture and conference room subdomains run this task.

- *MDM (Multiple Distant Microphones)*: These are a set of microphones situated on the meeting table. All participants in the conference room subdomain sit around the table as well as participants on the lecture room subdomain except for the lecturer. This task also exists in both subdomains.

- *MM3A (Multiple Mark III Microphone Arrays)*: The lecture meetings contain one or two of these arrays, which were built by NIST and contain 64 microphones setup linearly. Diarization could be run on either 64 channels or a beamformed version of it distributed by Karlsruhe University for RT06s.

- *MSLA (Multiple Source Localization Microphone Arrays)*: These are four groups of four microphones positioned into a “T“ shape array which were originally defined for speaker localization. They are only found in the lecture subdomain.

- *ADM (All Distant Microphones)*: In lecture room recordings this task allows the system to use all possible microphones previously explained (all except for the IHM microphones). The conference room subdomain does not usually define this task as all distant microphones are of MDM type.

- *IHM (Individual Headphone Microphone Arrays)*: Although not evaluated in the diarization evaluations, these microphones are worn by some of the participants in the meetings. They are a task in the STT evaluation and are also used when creating the forced-alignment reference segmentations for speaker diarization.

The test datasets used in both RT05s and RT06s evaluations were composed of conference and lecture type data. The conference data is composed of ten and nine meeting excerpts of 12 minutes each. One meeting was eliminated from RT06s after the evaluation finished for technical issues. These datasets have been used in this thesis to evaluate the different proposed techniques and are covered in mode detail in the experiments chapter and in appendix B. The lecture room data for test was composed of excerpts of different sizes contributed

by the different partners in the CHIL project and corresponding to different instants in a lecture meeting. In particular:

- RT05s test data was composed of 29 excerpts, all recorded at Karlsruhe University. Up to three excerpts were selected from each meeting, but systems were not expected to process the data from each meeting together. The majority of data corresponded to the lecturer, resulting in many excerpts where only one person was speaking. The shortest excerpt was 69 seconds and the longest 468 seconds.

- RT06s test data was composed of 38 excerpts of five minutes each, recorded in 5 different CHIL meeting rooms: 4 at AIT, 4 at IBM, 2 at ITC, 24 at Karlsruhe and 4 at UPC. This year the excerpts were chosen to contain a bigger variety of speakers and situations. After the evaluation finished, the set was reduced to 28 excerpts for technical reasons [2]. The development data used in these evaluations was usually a compilation of the data sets from previous evaluation campaigns. The used sets for conference room data were from RT02s and RT04s evaluations for RT05s, and a subset of RT02s through RT05s for the RT06s evaluation. For the lecture room evaluations, as this subdomain was first included in the evaluation in RT05s, there was no prior datasets available and therefore NIST distributed a set of transcribed lecture recordings similar to those in RT05s. For RT06s development was done using a subset of the original development set plus the RT05s evaluation set.

- RT07s evaluation data was divided into three portions according to meeting genre: conference meetings ,lecture meetings, and coffee breaks, the latter being a more interactive variant of the lecture room setup. The conference data consisted of excerpts from 8 meetings recorded at 4 sites in the U.S. and Europe (CMU, Edinburgh, NIST, and Virginia Tech), totaling 3 hours in duration. This part has been used in the experiments reported during this thesis work. The lecture data was collected at 5 different CHIL-consortium sites (AIT, IBM, ITC, UKA, and UPC), and comprised 32 lecture excerpts totaling 2.7 hours. Coffee break data originated from the same 5 sites and added up to 0.7 hours.

- RT09s evaluation data not includes lecture room or coffee break data. Only conference meetings were included in the evaluation. However, visual data and a new video input condition was presented. The conference data consisted of excerpts from 7 meeting recorded at 3 sites: 2 meeting from Edinburgh, 3 from NIST and 2 meeting from IDIAP, lasting 1 hour per each site, that is, a total of 3 hours of speech data.

Although the diarization system does not use any training data, the speech/non-speech detector based on the SVM classifier that has been used in the experiments of this PhD. was trained witn RT05s data, among other databases such as the CHIL meetings and the Speecon Database. All the diarization experiments performed in RT data were conducted on the conference subdomain. Following a brief description of each of the shows, in that subdomain, is provided in terms of SNR, number of speakers, speech time and overlap time, see table C.1.

[2] TNO show was discarded for evaluation purposes during RT06s but it was recover for experiments in this PhD. thanks to the corrections provided by NIST.

RT data	Show Name	SNR (dB)	#Speakers	Eval time	Speech time	Speaker time	Overlap time
Conference RT06s	CMU_20050912-0900	18.75	4	1070.9	1033.93	1495.96	387.13
	CMU_20050914-0900	13.50	4	1078.18	1036.76	1472.16	361.24
	EDI_20050216-1051	16.75	4	1079.29	964.11	1206.61	210.1
	EDI_20061114-1500	18.50	4	1356.91	964.71	1022.48	56.4
	NIST_20051024-0930	18.25	9	1088.24	1065.71	1642.07	423.84
	NIST_20051102-1323	17.25	8	1085.78	1031.82	1447.91	324.27
	TNO_20041103-1130	19.75	4	1079.53	971.42	1172.6	176.16
	VT_20050408-1500	27.25	5	1344.14	1023.84	1044.97	20.58
	VT_20050425-1000	27.50	4	1356.78	1031.3	1171.63	131.03
Conference RT07s	CMU_20061115-1030	21.75	4	1349.12	1100.53	1288.75	179.14
	CMU_20061115-1530	21.00	4	1353.16	1030.61	1130.69	96.52
	EDI_20050218-0900	19.25	4	1088.65	997.07	1272.04	211.91
	EDI_20061113-1500	20.00	4	1354.8	1094.86	1323.8	199.28
	NIST_20060216-1347	20.75	6	1347.86	1053.49	1131.2	70.75
	NIST_20051104-1515	20.00	4	1340.21	1054.88	1166.41	107.89
	VT_20050623-1400	35.50	5	1080.3	955.01	1306.95	276.48
	VT_20051027-1400	25.25	4	1064.91	878.16	982.78	96.77
Conference RT09s	EDI_20071128-1000	20.00	4	1761.77	1355.4	1475.91	114.37
	EDI_20071128-1500	21.75	4	1843.07	1266.92	1467.71	189.3
	IDI_20090128-1600	40.00	4	1805.96	1615.74	1808.6	177.35
	IDI_20090129-1000	23.00	4	1803.55	1366.91	1510.37	133.96
	NIST_20080201-1405	19.50	5	1218.64	1088.7	1569.24	385.61
	NIST_20080227-1501	24.50	6	1134.71	1021.36	1274.32	216.58
	NIST_20080307-0955	20.00	11	1278.59	1121.05	1286.71	140.81

Table C.1: *NIST Rich Transcription official conference evaluation data from RT06s, RT07s and RT09s. The "Show Name" column gives the site which provided the recording, next column stands for the number of speaker involved, "Eval Time" is the total time evaluated in seconds, "Speech Time" stands for the total time evaluated without non-speech in seconds, "Speaker Time" is the speech time counting the overlap speech between two speakers as twice or more depending of the numbers of overlapped speakers and "Overlap Time" is the speech time corresponding to any kind of speaker overlap. All time columns are expressed in seconds.*

<Recname>	<#Spks>	<Evaltime>	<Speech>	<Spktime>	<Overlap>	<2-Ovlp>	<3-Ovlp>	<4-Ovlp>
CMU_20050912-0900	4	1070.9	1033.93	1495.96	387.13	317.4	66.19	3.54
CMU_20050914-0900	4	1078.18	1036.76	1472.16	361.24	294.23	60.12	6.90
CMU_20061115-1030	4	1349.12	1100.53	1288.75	179.14	170	8.87	0.11
CMU_20061115-1530	4	1353.16	1030.61	1130.69	96.52	92.98	3.55	0.00
EDI_20050216-1051	4	1079.29	964.11	1206.61	210.1	180.84	26.86	2.41
EDI_20050218-0900	4	1088.65	997.07	1272.04	211.91	157.65	45.63	8.63
EDI_20061113-1500	4	1354.8	1094.86	1323.8	199.28	170.48	27.92	0.88
EDI_20061114-1500	4	1356.91	964.71	1022.48	56.4	55.04	1.36	0.00
EDI_20071128-1000	4	1761.77	1355.4	1475.91	114.37	108.23	6.14	0.00
EDI_20071128-1500	4	1843.07	1266.92	1467.71	189.3	178.02	11.05	0.22
IDI_20090128-1600	4	1805.96	1615.74	1808.6	177.35	163.05	13.37	0.93
IDI_20090129-1000	4	1803.55	1366.91	1510.37	133.96	124.75	8.91	0.30
NIST_20051024-0930	9	1088.24	1065.71	1642.07	423.84	309.99	81.32	20.96
NIST_20051102-1323	8	1085.78	1031.82	1447.91	324.27	256.15	54.9	12.19
NIST_20051104-1515	4	1340.21	1054.88	1166.41	107.89	104.48	3.17	0.24
NIST_20060216-1347	6	1347.86	1053.49	1131.2	70.75	64.34	5.9	0.51
NIST_20080201-1405	5	1218.64	1088.7	1569.24	385.61	302.6	76.18	6.83
NIST_20080227-1501	6	1134.71	1021.36	1274.32	216.58	183.87	30.09	2.62
NIST_20080307-0955	11	1278.59	1121.05	1286.71	140.81	119.35	18.96	2.49
TNO_20041103-1130	4	1079.53	971.42	1172.6	176.16	151.21	24.86	0.08
VT_20050408-1500	5	1344.14	1023.84	1044.97	20.58	20.03	0.55	0.00
VT_20050425-1000	4	1356.78	1031.3	1171.63	131.03	121.5	9.3	0.00
VT_20050623-1400	5	1080.3	955.01	1306.95	276.48	215.11	51.31	8.60
VT_20051027-1400	4	1064.91	878.16	982.78	96.77	89.61	6.47	0.69

Table C.2: *Statistics of recordings in NIST Rich Transcription official conference evaluation datasets RT06s, RT07s and RT09s. Number of speakers (#Spks); duration of recording (Evaltime); duration of speech (Speech); total speaker time including overlaps (Spktime); total overlap time (Overlap); total overlap time involving just two speakers (2-Ovlp); involving three speakers (3-Ovlp) and four-speaker overlap (4-Ovlp) in seconds.*

Own publications

Journal and Book Papers

- M. Zelenák, C. Segura , J. Luque , J. Hernando, ”Simultaneous Speech Detection With Spatial Features for Speaker Diarization,” IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no.2, pp.436-446, Feb. 2012
- B. Luque, L. Lacasa, F. Ballesteros and J. Luque “Horizontal visibility graphs: Exact results for random time series“, Physical Review E, PRE, vol. 80, Issue 4, id. 046103, 2009
- L. Lacasa, B. Luque, J. Luque, J. C. Nuño. “The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion“, Journal on Europhysics Letters, EPL, vol. 86, pp. 30001, 2009
- A. A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten and E. Pauwels, “Multimodal identification and localization of users in a smart environment“, Journal on Multimodal User Interfaces, Springer Berlin / Heidelberg, ISSN 1783-7677 (Print) 1783-8738 (Online), 2008
- L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J.C. Nuño, “From time series to complex networks: The visibility graph“, Proceedings of the National Academy of Science of the USA, PNAS, vol. 105, no. 13, pp. 4972-4975, 2008
- J. Luque, J. Hernando, ”Robust Speaker Identification for Meetings: UPC CLEAR’07 Meeting Room Evaluation System”, LNCS Springer-Verlag, vol. 4625, pp. 266-275, CLEAR 2007 and RT 2007, 2008
- J. Luque, X. Anguera, A. Temko, J. Hernando, ”Speaker Diarization for Conference Room: The UPC RT07s Evaluation System”, LNCS Springer-Verlag, vol. 4625, pp. 543-553, CLEAR 2007 and RT 2007, 2008
- J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, J. Hernando, “Audio, video and multimodal person identification in a smart room”, LNCS , Springer-Verlag, vol. 4122, pp. 258-269, CLEAR ’06 Classification of Events, Activities and Relationships, 2007

International Conferences

- J. Luque and J. Hernando, “On the use of Agglomerative and Spectral Clustering in Speaker Diarization of Meetings“, Proceedings Odyssey’12, The Speaker and Language Recognition Workshop 2012

- E. Bonet-Carne, T. Cobo, J. Luque, M. Martnez-Terrn, A. Perez-Moreno, M. Palacio, E. Gratacos, I. Amat-Roldan, “Consistent Association Between Image Features of Fetal Lungs from Different Ultrasound Equipments and Fetal Lung Maturity from Amniocentesi”, IEEE International Symposium on Biomedical Imaging (ISBI’12)

- A. Abad, J. Luque, I. Trancoso, “Parallel Transformation Network features for Speaker Recognition“, in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP’11), May 2011

- M. Sanz-Corts, F. Figueras, N. Padilla, N. Bargall, E. Bonet, J. Luque, I. Amat, E. Gratacs, ”Evaluation of a computer-based analysis of brain textures on fetal MRI to detect changes in small for gestational age (SGA) fetuses and to predict neurodevelopmental outcome“, ISUOG’11 World Congress, International Society of Ultrasound in Obstetrics and Gynecology.

- A. Abad and J. Luque, “Connectionist Transformation Network Features for Speaker Recognition“, Proceedings Odyssey’10, The Speaker and Language Recognition Workshop 2010

- A. Abad, J. Luque, I. Trancoso and J. Hernando, ”The L2F - UPC Speaker Recognition System for NIST SRE 2010“, Proceedings of the 2010 NIST Speaker Recognition evaluation (SRE’2010) Workshop, 2010

- J. Luque, C. Segura, J. Hernando,“Clustering initialization based on spatial information for speaker diarization of meetings“, Proceedings International Conference on Spoken Language Processing (ICSLP’08), pp. 383-386, 2008

- R. Morros, A.A. Salah, B. Schouten, C. Segura, J. Luque, O. Ambekar, C. Kayalar, L. Akaru,“Event Recognition for meaningful human-computer interaction in a smart environment”, Proccedings of the eNTERFACE’07 Workshop on Multimodal Interfaces, pp. 71-86, 2008

- J. Hernando, M. Farrús, P. Ejarque, A. Garde, J. Luque, “Person verification by fusion of prosodic, voice spectral and facial parameters”, International Conference on Security and Cryptography (SECRYPT’06), pp. 17-23, 2006

- M. Farrús, A. Garde, P. Ejarque, J. Luque, J. Hernando, “On the fusion of prosody, voice spectrum and face features for multimodal person verification”, 9th International Conference on Spoken Language Processing (ICSLP’06), pp. 2106-2109, 2006

- B. Luque, J. Luque, F. J. Ballesteros, J. C. Nuño, “From Temporal Series to Complex Networks”, IX Latin American WorkShop on nonlinear phenomena (LAWNP), Proceedings LAWNP, San Carlos de Bariloche, Argentina, 2005

National Conferences

- J. Luque, L. Lacasa, B. Luque, “Power Laws and Scaling in the Waiting Time Distribution of Speech ”, FisEs 2012, XVIII Congreso de Fsica Estadstica. 18-20 de octubre de 2012
- J. Luque, D. Ferrés, J. Hernando, J.B. Mariño, H. Rodríguez, “GeoVAQA: A Voice Activated geographical Question Answering system” IV Jornadas en Tecnología del Habla, pp. 309-314, 2006
- J. Luque, R. Morros, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, J. Hernando, “Multimodal person identification in a smart room, IV Jornadas en Tecnología del Habla”, pp. 327-331, 2006

Bibliography

[Abad and Luque, 2010] A. Abad and J. Luque. Connectionist transformation network features for speaker recognition. In *The Speaker and Language Recognition Workshop, Odyssey*, 2010.

[Abad and Trancoso, 2010] Alberto Abad and Isabel Trancoso. Speaker recognition experiments using connectionist transformation network features. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 378–381, 2010.

[Abad *et al.*, 2010] A. Abad, J. Luque, I. Trancoso, and J. Hernando. The l2f - upc speaker recognition system for nist sre 2010. In *The 2010 NIST Speaker Recognition evaluation (SRE'2010) Workshop*, 2010.

[Abad *et al.*, 2011] A. Abad, J. Luque, and I. Trancoso. Parallel transformation network features for speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 5300–5303, 2011.

[Abrash *et al.*, 1995] V. Abrash, H. Franco, A. Sankar, and M. Cohen. Connectionist speaker normalization and adaptation. In *European Conference on Speech Communication and Technology, Eurospeech*, pages 2183–2186, Madrid, 1995.

[Adami *et al.*, 2002] Andre Adami, Lukas Burget, Stephane Dupont, Hari Garudadri, Frantisek Grezl, Hynek Hermansky, Pratibha Jain, Sachin Kajarekar, Nelson Morgan, and Sunil Sivadas. Qualcomm-icsi-ogi features for asr. In *International Conference on Spoken Language Processing, ICSLP*, pages 4–7, 2002.

[Ajmera and et al., 2004] J. Ajmera and et al. Clustering ans segmenting speakers and their locations in meetings. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 605–608, 2004.

[Ajmera and Wooters, 2003] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, 2003.

[Ajmera *et al.*, 2002] J. Ajmera, IA McCowan, and H. Bourlard. Robust HMM-based speech/music segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, 2002.

[Anastasakos and et al., 1996] T. Anastasakos and et al. A compact model for speaker-adaptative training. In *International Conference on Spoken Language Processing, ICSLP*, 1996.

[Anguera and Hernando, 2004] X. Anguera and J. Hernando. Evolutive speaker segmentation using a repository system. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2004.

[Anguera *et al.*, 2006a] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando. Hybrid speech/non-speech detector applied to speaker diarization of meetings. In *The Speaker and Language Recognition Workshop, 2006. IEEE Odyssey*, pages 1 –6, june 2006.

[Anguera *et al.*, 2006b] X. Anguera, C. Wooters, and J. Hernando. Purity algorithms for speaker diarization of meeting data. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2006.

[Anguera *et al.*, 2006c] X. Anguera, C. Wooters, and J. Hernando. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *International Conference on Spoken Language Processing, ICSLP*, 2006.

[Anguera *et al.*, 2006d] Xavier Anguera, Chuck Wooters, and Javier Hernando. Automatic cluster complexity and quantity selection: towards robust speaker diarization. In *Proceedings of the Third international conference on Machine Learning for Multimodal Interaction*, MLMI'06, pages 248–256, Berlin, Heidelberg, 2006. Springer-Verlag.

[Anguera *et al.*, 2007a] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.

[Anguera *et al.*, 2007b] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.

[Anguera *et al.*, 2011] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[Anguera, 2005] X. Anguera. Beamformit: The robust acoustic beamforming toolkit. Website, http://www.icsi.berkeley.edu/~xanguera/beamformit, 2005. http://www.icsi.berkeley.edu/~xanguera/beamformit.

[Anguera, 2006] X. Anguera. *Robust Speaker Diarization for Meetings*. PhD thesis, Universitat Politecnica de Catalunya, 2006.

[Anliker and et al., 2006] U. Anliker and et al. Speaker separation and tracking system. *Journal on Applied Signal Processing, EURASIP*, 2006 No. 1, 2006.

[Appel and Brandt, 1982] U. Appel and A. Brandt. Adaptative sequential segmentation of piecewise stationary time series. In *Inf. Sci.*, volume 29, 1982.

[Araki *et al.*, 2008] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino. Speaker indexing and speech enhancement in real meetings/conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 1, pages 93–96, Las vegas, USA, 2008.

[Atkinson, 1968] K. Atkinson. Language identification from nonsegmental cues. *Journal of the Acoustical Society of America*, 44:378(A), 1968.

[Auckenthaler *et al.*, 2000] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, jan 2000.

[Bar-Shalom and Fortman, 1988] Y. Bar-Shalom and T.E. Fortman. Tracking and Data association. In *Academic Press*, 1988.

[Barra-Chicote *et al.*, 2011] R. Barra-Chicote, J.M. Pardo, J. Ferreiros, and J.M. Montero. Speaker diarization based on intensity channel contribution. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):754 –761, may 2011.

[Barras and et al., 2004] C. Barras and et al. Improving speaker diarization. In *Fall 2004 Rich Transcription Workshop (RT04)*, 2004.

[Barras *et al.*, 2006] C. Barras, Xuan Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1505–1512, sept. 2006.

[Beigi and et al., 1998] H. Beigi and et al. A distance measure between collections of distributions and its application to speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 1998.

[Ben and et al., 2004] M. Ben and et al. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2004.

[Bennani and Gallinari, 1991] Y. Bennani and P. Gallinari. On the use of tdnn-extracted features information in talker identification. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:385–388, 1991.

[Bernardin and Stiefelhagen, 2007] K. Bernardin and R. Stiefelhagen. Audio-visual multi-person tracking and identification for smart environments. In *MM Multimedia*, pages 661–670, 2007.

[Bernardin *et al.*, 2006] K. Bernardin, A. Elbs, and R. Stiefelhagen. Multiple object tracking performance metrics and evaluation in a smart room environment. *IEEE Int. Workshop on Vision Algorithms*, pages 53–68, 2006.

[Beyerlein and et al., 2002] P. Beyerlein and et al. Large vocabulary continous speech recognition of broadcast news. the Philips/RWTH approach. In *Speech Communcations*, volume 37, pages 109–131, 2002.

[Bimbot and Mathan, 1993] F. Bimbot and L. Mathan. Text-free speaker recognition using an arithmetic-harmonic spherecity measure. In *European Conference on Speech Communication and Technology, Eurospeech*, pages 169–172, 1993.

[Bimbot *et al.*, 2004] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D.A. Reynolds. A tutorial of text-independent speaker verification. *Journal on Applied Signal Processing, EURASIP*, 4:430–451, 2004.

[Boakye *et al.*, 2008a] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland. Overlapped speech detection for improved speaker diarization in multiparty meetings. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4353 –4356, 31 2008-april 4 2008.

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

[Boakye *et al.*, 2008c] K. Boakye, O. Vinyals, and G. Friedland. Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 32–35, Brisbane, Australia, 2008.

[Boakye *et al.*, 2008d] K. Boakye, O. Vinyals, and G. Friedland. Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In *Proc. Interspeech '08*, pages 32–35, Brisbane, Australia, 2008.

[Boll, 1979] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-27(2), April 1979.

[Bolle and et al., 2004] R.M. Bolle and et al. *Guide to Biometrics*. Springer, 2004.

[Boser *et al.*, 1992] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT*, pages 144–152, 1992.

[Bozonnet *et al.*, 2010a] S. Bozonnet, N.W.D. Evans, and C. Fredouille. The lia-eurecom rt'09 speaker diarization system: Enhancements in speaker modelling and cluster purification. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4958 –4961, march 2010.

[Bozonnet *et al.*, 2010b] Simon Bozonnet, Nicholas W. D. Evans, X Anguera, O. Vinyals, G. Friedland, and C. Fredouille. System output combination for improved speaker diarization. In *Annual Conference of the International Speech Communication Association, Interspeech*, 2010.

[Brandstein and Silverman, 1997] M.S. Brandstein and H.F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 375–378, Munich, Germany, 1997.

[Brümmer and Preez, 2006] N. Brümmer and J. Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20:230275, 2006.

[Brummer *et al.*, 2007] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2072 –2084, sept. 2007.

[Brummer, 2005] N. Brummer. Tools for fusion and calibration of automatic speaker detection systems, 2005.

[Burget *et al.*, 2008] Lukas Burget, Michal Fapso, Valiantsina Hubeika, Ondej Glembek, Martin Karafiat, Marcel Kockmann, Pavel Matjka, Petr Schwarz, and Jan ernocký. But system description: Nist sre 2008. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, pages 1–4. National Institute of Standards and Technology, 2008.

[Busso and et al., 2005] C. Busso and et al. Smart room: participant and speaker localization and identification. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 2, pages 1117–1120, 2005.

[Campbell *et al.*, 2004] J.P. Campbell, H. Nakasone, C. Cieri, K. Walker D. Miller, A.F. Martin, and M.A. Pryzbocki. The mmsr bilingual and crosschannel corpora for speaker recognition research and evaluation. In *The Speaker and Language Recognition Workshop, 2004. IEEE Odyssey*, page 2932, 2004.

[Campbell *et al.*, 2006a] W. M. Campbell, J.R. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[Campbell *et al.*, 2006b] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[Campbell, 1997] J.P. Campbell. Speaker recogntion: A tutorial. *Invited Paper of Proceedings of the IEEE*, 85 No. 9:1437–1462, 1997.

[Campbell, 2008] W. M. Campbell. A covariance kernel for svm language recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2008.

[Carpenter *et al.*, 1997] J. Carpenter, P. Clifford, and P. Fernhead. An improved particle filter for non-linear problems. Technical report, Department of Statistics, University of Oxford, 1997.

[Casas and Neumann, 2007] Josep R. Casas and Joachim Neumann. Context awareness triggered by multiple perceptual analyzers. In *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 371–383, Amsterdam, The Netherlands, The Netherlands, 2007. IOS Press.

[Casas and Stiefelhagen, 2005] J. Casas and R. Stiefelhagen. Multi-camera/multi-microphone system design for continuous room monitoring, 2005.

[Castaldo *et al.*, 2008] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair. Stream-based speaker segmentation using speaker factors and eigenvoices. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4133 –4136, 31 2008-april 4 2008.

[Cettolo and Federico, 2000] M. Cettolo and M. Federico. Model selection criteria for acoustic segmentation. In *ISCA ITRW ASR2000*, pages 221–227, 2000.

[Cettolo and Vescovi, 2003] M. Cettolo and M. Vescovi. Efficient audio segmentation algorithms based on the BIC. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2003.

[Chan *et al.*, 2006] W. Chan, T. Lee, N. Zheng, and H. Ouyang. Use of vocal source features in speaker segmentation. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2006.

[Chang and Lin, 2001] C.-C. Chang and C-J Lin. Libsvm - a library for support vector machines. Website, http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html, 2001.

[Chen and Gopalakrishnan, 1998] S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA BNTU Workshop*, 1998.

[Chen *et al.*, 1997] Ke Chen, Lan Wang, and Huisheng Chi. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, 11:417–445, 1997.

[Chen *et al.*, 2002] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, and P. Olsen. Automatic transcription of broadcast news. *Speech Communication*, 37:69–87, May 2002.

[Chen *et al.*, 2004] Jingdong Chen, Nteng Huang, and J. Benesty. An adaptive blind SIMO identification approach to joint multichannel time delay estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages iv–53–iv–56, 2004.

[Chen *et al.*, 2011] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y. Chang. Parallel spectral clustering in distributed systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):568–586, 2011.

[Chetty and Wagner, 2007] G. Chetty and M. Wagner. Audio visual speaker verification based on hybrid fusion of cross modal features. In *PReMI 2007, Lecture Notes on Computer Science, LNCS*, volume 4815, pages 469–478, 2007.

[chil, 2006] European union, 6^{th} framework integrated project CHIL, 2006.

[Choi *et al.*, 2007] Mu Yeol Choi, Hwa Jeon Song, and Hyung Soon Kim. Speech/music discrimination for robust speech recognition in robots. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 118 –121, aug. 2007.

[Chollet and Bimbot, 1995] Gérard Chollet and Frédéric Bimbot. Assessment of speaker verification systems. In *Spoken Language Resources and Assessment*. EAGLES Handbook, 0 1995.

[CLEAR, 2007] Classification of events, activities and relationships: Evaluation and workshop. Website,http://www.clear-evaluation.org, 2007. http://www.clear-evaluation.org.

[Cortes and Vapnik, 1995] Corinna Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[Cover and Thomas, 1991] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991.

[Cristianini and Shawe-Taylor, 2000] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, 2000.

[darpa, 1998] *DARPA/NIST Smart Spaces Workshop*, volume 3, 1998.

[Davis and Mermelstein, 1980] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions ASSP Magazine*, No. 28:357–366, 1980.

[de la Torre *et al.*, 2005] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, and A.J. Rubio. Histogram equalization of speech representation for robust speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13(3):355 – 366, may 2005.

[Dehak *et al.*, 2007] N. Dehak, P. Dumouchel, and P. Kenny. Modeling prosodic features with joint factor analysis for speaker verication. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2007.

[Delacourt and Kryze, 1999] P. Delacourt and D. Kryze. Detection of speaker changes in an audio document. In *European Conference on Speech Communication and Technology, Eurospeech*, 1999.

[Delacourt and Wellekens, 2000] P. Delacourt and C.J. Wellekens. DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication*, 32 No. 1-2:111–126, 2000.

[Dempster *et al.*, 1977] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–3, 1977.

[DiBiase *et al.*, 2001] J. DiBiase, H. Silverman, and M. Brandstein. *Microphone Arrays. Robust Localization in Reverberant Rooms*. Springer, 2001.

[Doddington *et al.*, 2000] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The nist speaker recognition evaluation overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225 – 254, 2000.

[Doddington, 1971] G. Doddington. A method for speaker verification. *Journal of the Acoustical Society of America*, 49:139, 1971.

[Duda and Hart, 1973] R. Duda and P. Hart. *Pattern classification and Scene analysis*. John Wiley and Sons, 1973.

[Dunn *et al.*, 2000] R.B. Dunn, D.A. Reynolds, and T.F. Quatieri. Approaches to speaker detection and tracking in conversational speech. In *Digital Signal Processing 10*, pages 93–112, 2000.

[Ellis and Liu, 2004] Daniel P. W. Ellis and Jerry C. Liu. Speaker turn segmentation based on between-channel differences. In *at NIST Meeting Recognition Workshop at ICASSP 2004*, 2004.

[enterface, 2007] The similar noe summer workshop on multimodal interfaces, enterface'07, 2007.

[Farrús *et al.*, 2006] M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando. On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *9th International Conference on Spoken Language Processing, ICSLP*, pages 2106–2109, 2006.

[Farrús *et al.*, 2007] Mireia Farrús, Pascual Ejarque, Andrey Temko, and Javier Hern. Histogram equalization in svm multimodal person verification. In *Lecture Notes in Computer Science, Proceedings of the IEEE International Conference on Biometrics*, pages 819–827. Springer, 2007.

[Ferrer *et al.*, 2008] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg. System combination using auxiliary information for speaker verification. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4853 –4856, 31 2008-april 4 2008.

[Ferrer *et al.*, 2010] L. Ferrer, N. Scheffer, and E. Shriberg. A comparison of approaches for modeling prosodic features in speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2010.

[Fiscus and et al., 2004] J. Fiscus and et al. Results of the fall 2004 stt and mde evaluation. In *Proceedings of the Fall 2004 Rich Transcription Workshop*, 2004.

[Fiscus and et al., 2007a] J. Fiscus and et al. The rich transcription 2007 meeting recognition evaluation. Website, http://www.nist.gov/speech/tests/rt/2009/, 2007.

[Fiscus and et al., 2007b] J. Fiscus and et al. The rich transcription spring 2003 evaluation (rt-03s). Website, http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/, 2007.

[Fiscus and et al., 2009a] J. Fiscus and et al. The rich transcription 2009 meeting recognition evaluation. Website, http://www.nist.gov/speech/tests/rt/2009/, 2009.

[Fiscus and et al., 2009b] J. Fiscus and et al. The rich transcription 2009 meeting recognition evaluation, workshop agenda and presentations. Website, http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-Agenda.htm, 2009.

[Flanagan *et al.*, 1985] J. Flanagan, J. Johnson, R. Kahn, and G. Elko. Computer-steered microphone arrays for sound transduction in large rooms. In *ASAJ*, volume 78, No. 5, pages 1508–1518, 1985.

[Flanagan *et al.*, 2008] James L. Flanagan, Jont B. Allen, and Mark A. Hasegawa-Johnson. *Speech analysis synthesis and perception*. Springer-Verlag, third edition, 2008.

[Fleuret *et al.*, 2008] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 2008.

[Fox and et al., 2003] N.A Fox and et al. Person Identification Using Automatic Integration of Speech, Lip and Face Experts. In *ACM SIGMM WBMA*, pages 25–32, 2003.

[Fox *et al.*, 2011] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A sticky hdp-hmm with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[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 *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4069–4072, Washington, DC, USA, 2009. IEEE Computer Society.

[Friedland *et al.*, 2009b] Gerald Friedland, Chuohao Yeo, and Hayley Hung. Visual speaker localization aided by acoustic models. In *ACM Multimedia*, 0 2009.

[Friedland *et al.*, 2011] G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. Knox, and O. Vinyals. The icsi rt-09 speaker diarization system. *Audio, Speech, and Language Processing, IEEE Transactions on*, PP(99):1, 2011.

[Fung and Mangasarian, 2001] G. Fung and O. Mangasarian. Proximal Support Vector Machine Classifiers. In *seventh ACM SIGKDD international booktitle on Knowledge discovery and data mining, KDD*, pages 77–86, 2001.

[Furui, 1981] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics Speech and Signal Processing*, 29:254–272, 1981.

[Furui, 1986] S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions ASSP Magazine*, No. 34:52–59, 1986.

[Furui, 1996] S. Furui. *An overview of speaker recognition technology In Automatic Speech and Speaker Recognition: Advanced Topics*. Academic Publishers, 1996.

[Furui, 2005] S. Furui. 50 years of progress in speech and speaker recognition. In *10th International Conference on Speech and Computer, SPECOM*, pages 1–9, 2005.

[G. Lathoud and I.A. McCowan, 2003] G. Lathoud and I.A. McCowan. Location based speaker segmentation. In *2003 International Conference on Multimedia and Expo (ICME'03)*, volume 3, pages III–621–4, Baltimore, USA, 2003.

[Gangadharaiah and et al., 2004] R. Gangadharaiah and et al. A novel method for two-speaker segmentation. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2004.

[Gatica-Perez *et al.*, 2007] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(2):601–616, 2007.

[Gauvain and et al., 2002] J.L. Gauvain and et al. The LIMSI broadcast news transcription system. In *Speech Communcations*, volume 37, pages 89–108, 2002.

[Gauvain and Lee, 1994] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations markov chains. *IEEE Transactions on Speech Audio Processing*, 2(2):291–298, April 1994.

[Gauvain *et al.*, 1998] J.L. Gauvain, L. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. In *International Conference on Spoken Language Processing, ICSLP*, pages 1335–1338, 1998.

[Ghahramani and Jordan, 1997] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Mach. Learn.*, 29:245–273, November 1997.

[Gish and Schmidt, 1994] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, 1994.

[Glembek *et al.*, 2009] O. Glembek, L. Burget, N. Dehak, N. Brmmer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 4057–4060, 2009.

[Glembek, 2009] Ondrej Glembek. Joint factor analysis matlab demo. WebSite, http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo, 2009.

[Godfrey *et al.*, 1992] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 1992.

[Gordon *et al.*, 1993] N.J. Gordon, N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings on Radar and Signal Processing*, 140(2):107–113, 1993.

[Gupta *et al.*, 2007] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel. Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations. *Signal Processing Letters, IEEE*, 14(12):1040 –1043, dec. 2007.

[Han and Narayanan, 2008] Kyu Jeong Han and Shrikanth S. Narayanan. A novel inter-cluster distance measure combining glr and icr for improved agglomerative hierarchical speaker clustering. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4373–4376, Las Vegas, Nevada, April 2008.

[Han *et al.*, 2008] K.J. Han, S. Kim, and S.S. Narayanan. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(8):1590 –1601, nov. 2008.

[Haptek, 2011] Haptek, 2011.

[Hecker *et al.*, 1968] M.H.L. Hecker, K. Stevens, G. von Bismark, and C. Williams. Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44:993–1001, 1968.

[Hermansky and Morgan, 1994] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, pages 578–589, 1994.

[Hernando *et al.*, 2006] J. Hernando, M. Farrús, P. Ejarque, A. Garde, and J. Luque. Person verification by fusion of prosodic, voice spectral and facial parameters. In *International Conference on Security and Cryptography, SECRYPT*, pages 17–23, 2006.

[Hernando, 1997] J. Hernando. Cdhmm speaker recognition by means of frequency ltering of lter-bank energies. In *European Conference on Speech Communication and Technology, Eurospeech*, 1997.

[Huijbregts *et al.*, 2009] M. Huijbregts, D. van Leeuwen, and F. de Jong. Speech Overlap Detection in a Two-Pass Speaker Diarization System. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 1063–1066, Brighton, UK, 2009.

[Imseng and Friedland, 2010] D. Imseng and G. Friedland. Tuning-robust initialization methods for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2028 –2037, nov. 2010.

[Indovina and et al., 2003] M. Indovina and et al. Multimodal Biometric Authentication Methods: A COTS Approach. In *MMUA*, pages 99–106, 2003.

[Isard and Blake, 1998] Michael Isard and Andrew Blake. CONDENSATION–Conditional density propagation for visual tracking. *International Journal of Computer Vision*, V29(1):5–28, August 1998.

[Iso, 2010] K. Iso. Speaker clustering using vector quantization and spectral clustering. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4986 –4989, march 2010.

[Johnson and Woodland, 1998] S. Johnson and P. Woodland. Speaker clustering using direct maximization of the MLLR-adapted likelihood. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 5, pages 1775–1779, 1998.

[Juang and Rabiner, 1985] B.H. Juang and L.R. Rabiner. A probabilistic distance measure for hidden markov models. In *AT&T Technical Journal*, volume 64 No. 2, pages 391–408, 1985.

[Kasabov, 1973] N. Kasabov. *Evolving connectionist systems: Methods and applications in bioinformatics*. Springer Verlag, 1973.

[Katsarakis *et al.*, 2007] Katsarakis, N., Souretis, G., Talantzis, F., Pnevmatikakis, A., and L. Polymenakos. 3D Audiovisual Person Tracking Using Kalman Filtering and Information Theory. In *Lecture Notes on Computer Science, LNCS*, volume 4122, page 45. Springer, 2007.

[Kemp and M. Schmidt, 2000] T. Kemp and A. Waibel M. Schmidt, M. Westphal. Strategies for automatic segmentation of audio data. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 1423–1426, 2000.

[Kenny and Castaldo, 2009] D. Reynolds P. Kenny and F. Castaldo. A study of new approaches to speaker diarization. In *Annual Conference of the International Speech Communication Association, Interspeech*, 2009.

[Kenny and Dumouchel, 2004] P. Kenny and P. Dumouchel. Experiments in speaker verification using factor analysis likelihood ratios. In *The Speaker and Languakge Recognition Workshop, Odyssey*, 2004.

[Kenny *et al.*, 2005] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Factor analysis simplified. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 637–640, 2005.

[Kenny *et al.*, 2007] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2072–2084, 2007.

[Kenny *et al.*, 2008a] P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel. The role of speaker factors in the nist extended data task. In *The Speaker and Language Recognition Workshop, Odyssey*, 2008.

[Kenny *et al.*, 2008b] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):980 –988, july 2008.

[Kenny *et al.*, 2010] P. Kenny, D. Reynolds, and F. Castaldo. Diarization of telephone conversations using factor analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 4(6):1059 –1070, dec. 2010.

[Kenny, 2005] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. Technical Report CRIM-0608-13, CRIM, 2005.

[Kenny, 2008] P. Kenny. Bayesian analysis of speaker diarization with eigenvoice priors, 2008.

[Kent and Read, 1992] R.D. Kent and C. Read. *The Acoustic Analysis of Speech*. Whurr Publishers - Singular Publishing Group, London - San Diego, 1992.

[Kersta, 1962] L.G. Kersta. Voiceprint identification. *Journal of the Acoustical Society of America*, 34:725, 1962.

[Keshet and Bengio, 2008] Joseph Keshet and Samy Bengio. *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*. John Wiley & Sons, 2008.

[Khan *et al.*, 2003] Z. Khan, T. Balch, and F. Dellaert. Efficient particle filter-based tracking of multiple interacting targets using an mrf-based motion model. International Conference on Intelligent Robots and Systems, 2003.

[Kim and et al., 2005] T. Kim and et al. Hybrid speaker-based segmentation system using model-level clustering. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2005.

[Kim *et al.*, 2007] Bong-Wan Kim, Dae-Lim Choi, and Yong-Ju Lee. Speech/music discrimination using mel-cepstrum modulation energy. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, pages 406–414, Berlin, Heidelberg, 2007. Springer-Verlag.

[Kirby and L.Sirovic, 1990] M. Kirby and L.Sirovic. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 12(1):103–108, January 1990.

[Knapp and Carter, 1976] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 24, No. 4:320–327, 1976.

[Koh and et al., 2008] E. C. W. Koh and et al. Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I^2R-NTU Submission for the NIST RT 2007 Evaluation. In *Lecture Notes on Computer Science, LNCS: CLEAR 2007 and RT 2007*, volume 4625. Springer-Verlag, 2008.

[Kohonen, 1990] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[Kotti *et al.*, 2008] M. Kotti, E. Benetos, and C. Kotropoulos. Computationally efficient and robust bic-based speaker segmentation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):920 –933, july 2008.

[Kryszczuk *et al.*, 2007] Krzysztof Kryszczuk, Jonas Richiardi, Plamen Prodanov, and Andrzej Drygajlo. Reliability-based decision fusion in multimodal biometric verification systems. *EURASIP J. Appl. Signal Process.*, 2007:74–74, January 2007.

[Kubala and et al., 1997] F. Kubala and et al. The 1996 BBN byblos HUB-4 transcription system. In *Speech Recognition Workshop*, volume 32 No. 1-2, pages 90–93, 1997.

[Kurematsu *et al.*, 2005] A. Kurematsu, M. Nakano-Miyatake, H. Perez-Meana, and E. Simancas-Acevedo. Performance analysis of gaussian mixture model speaker recognition systems with different speaker features. *Electronic Journal Technical Acoustics*, 14, 2005.

[Lapidot, 2003] I. Lapidot. SOM as likelihood estimator for speaker clustering. In *European Conference on Speech Communication and Technology, Eurospeech*, 2003.

[Laskowski and Schultz, 2006] K. Laskowski and T. Schultz. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *IEEE International Conference on Acoustics, Speech and Signal Processing,ICASSP*, volume I, pages 993–996, Toulouse, France, 2006.

[Laskowski and Shriberg, 2012] Kornel Laskowski and Elizabeth Shriberg. Corpus-independent history compression for stochastic turn-taking models. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*. IEEE, 2012.

[Laskowski *et al.*, 2004] K. Laskowski, Q. Jin, and T. Schultz. Crosscorrelation-based multispeaker speech activity detection. In *Eighth International Conference on Spoken Language Processing*, pages 973–976, Jeju Island, Korea, 2004.

[Lathoud and et al., 2002] G. Lathoud and et al. Unsupervised location-based sementation of multi-party speech. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP, NIST Meeting Recognition Workshop*, 2002.

[Lebrun *et al.*, 2004] G. Lebrun, C. Charrier, and H. Cardot. SVM Training Time Reduction using Vector Quantization. In *International Conference on Pattern Recognition ,ICPR*, pages 160–163, 2004.

[Levy and Lindenbaum, 2000] A. Levy and M. Lindenbaum. Sequential Karhunen-Loeve basis extraction and its application to images. *IEEE Transactions on Image Processing*, 9(8):1371–1374, 2000.

[Li and Schultz, 2009] Runxin Li and Tanja Schultz. Improving speaker segmentation via speaker identification and text segmentation. In *Annual Conference of the International Speech Communication Association, Interspeech*, 2009.

[Linguistic Data Consortium, LDC, 2002] Catalogue of speaker recognition corpora, 2002.

[Liu and Kubala, 1999] Daben Liu and Francis Kubala. Fast speaker change detection for broadcast news transcription and indexing. In *European Conference on Speech Communication and Technology, Eurospeech*, pages 1031–1034, 1999.

[Liu *et al.*, 1993] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, and Alejandro Acero. Efficient cepstral normalization for robust speech recognition. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 69–74, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.

[Lu and Zhang, 2002] L. Lu and H.J. Zhang. Speaker change detection and tracking in real-time news broadcasting analysis. In *ACM International Conference on Multimedia*, pages 602–610, 2002.

[Lucey and Chen, 2003] S. Lucey and T. Chen. Improved Audio-visual Speaker Recognition via the Use of a Hybrid Combination Strategy. In *The 4th International Conference on Audio- and Video- Based Biometric Person Authentication*, 2003.

[Luque and Hernando, 2008a] J. Luque and J. Hernando. Robust speaker identification for meetings: UPC clear-07 meeting room evaluation system. In *International CLEAR Evaluation and NIST Rich Transcription Workshop, Lecture Notes on Computer Science, LNCS*, volume 4625. Springer-Verlag, 2008.

[Luque and Hernando, 2008b] J. Luque and J. Hernando. Robust speaker identification for meetings: Upc clear-07 meeting room evaluation system. In *Lecture Notes on Computer Science, LNCS*, volume 4625. Springer-Verlag, 2008.

[Luque and Hernando, 2012] J. Luque and J. Hernando. On the use of agglomerative and spectral clustering in speaker diarization of meetings. In *The Speaker and Language Recognition Workshop, 2012. IEEE Odyssey*, 2012.

[Luque *et al.*, 2006a] J. Luque, R. Morros, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando. Multimodal person identification in a smart room. In *IV Jornadas en Tecnología del Habla*, pages 327–331, 2006.

[Luque *et al.*, 2006b] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando. Audio, video and multimodal person identification in a smart room. In *Lecture Notes on Computer Science, LNCS*, volume 4122. Springer Berlin/Heidelberg, 2006.

[Luque *et al.*, 2008a] J. Luque, X. Anguera, A. Temko, and J. Hernando. Speaker Diarization for Conference Room: The UPC RT07s Evaluation System. In *Lecture Notes on Computer Science, LNCS: Multimodal Technologies for Perception of Humans*, volume 4625, pages 543–553. Springer Berlin/Heidelberg, 2008.

[Luque *et al.*, 2008b] J. Luque, C. Segura, and J. Hernando. Clustering initialization based on spatial information for speaker diarization of meetings. In *International Conference on Spoken Language Processing, ICSLP*, pages 383–386, Brisbane, Australia, 2008.

[Luxburg *et al.*, 2007] Ulrike Von Luxburg, Mikhail Belkin, Olivier Bousquet, and Pertinence. A tutorial on spectral clustering. *Stat. Comput*, 2007.

[Macho and Nadeu, 1999] D. Macho and C. Nadeu. On the interaction between time and frequency filterinf of speech parameters for robust speech recognition. In *International Conference on Spoken Language Processing, ICSLP*, page paper 1137, 1999.

[Macho *et al.*, 2006] Dušan Macho, Climent Nadeu, and Andrey Temko. Robust speech activity detection in interactive smart-room environments. In Steve Renals, Samy Bengio, and Jonathan Fiscus, editors, *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 236–247. Springer Berlin / Heidelberg, 2006. 10.1007/11965152_21.

[Mackay, 2003] D. J. C. Mackay. *Information theory, inference, and learning algorithms*. Cambridge University Press New York, 2003.

[Makhoul and et al., 2000] J. Makhoul and et al. Speech and language technologies for audio indexing and retreival. In *Proceeding of IEEE*, volume 88, pages 1338–1353, 2000.

[Malegaonkar and et al., 2006] A. Malegaonkar and et al. Unsupervised speaker change detection using probabilistic pattern matching. *IEEE Signal Processing Letters*, 13(8):509–512, 2006.

[Markel *et al.*, 1977] J.D. Markel, B.T. Oshika, and A.H. Gray Jr. Long-term feature averaging for speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(4):330–337, 1977.

[Martin and et al., 2000] A. Martin and et al. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspectives. *Speech Communications*, 31:225–254, 2000.

[Martin and et al., 2008a] A. F. Martin and et al. The 2008 nist speaker recognition evaluation results. Website,http://www.itl.nist.gov/iad/mig/tests/sre/2008/official_results/index.html, 2008. http://www.itl.nist.gov/iad/mig/tests/sre/2008/official_results/index.html.

[Martin and et al., 2008b] A. F. Martin and et al. The 2010 nist speaker recognition evaluation results. Website,http://www.nist.gov/itl/iad/mig/sre10results.cfm, 2008. http://www.nist.gov/itl/iad/mig/sre10results.cfm.

[Martin and Przybocki, 2001] A. F. Martin and M. A. Przybocki. The nist speaker recognition evaluations: 1996-2001. In *The Speaker and Language Recognition Workshop, 2001. IEEE Odyssey*, pages 39–43, 2001.

[Martin, 2010] A.F. Martin. The 2010 nist speaker recognition evaluation plan (sre10). Website, http://www.itl.nist.gov/iad/mig//tests/sre/2010/, 2010.

[Matsoukas and et al., 1997] S. Matsoukas and et al. Practical implementations of speaker-adaptative training. In *DARPA Speech Recognition Workshop*, 1997.

[Matsoukas *et al.*, 2006] S. Matsoukas, J.-L. Gauvain, G. Adda, T. Colthurst, Chia-Lin Kao, O. Kimball, L. Lamel, F. Lefevre, J.Z. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and Bing Xiang. Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1541–1556, September 2006.

[Mccowan *et al.*, 2005] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

[Meignier and et al., 2001] S. Meignier and et al. E-HMM approach for learning and adapting sound models for speaker indexing. In *The Speaker and Language Recognition Workshop, Odyssey*, pages 175–180, 2001.

[Meignier *et al.*, 2006] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-Franois Bonastre, and Laurent Besacier. Step-by-step and integrated approaches in broadcast news speaker diarization. *Computer Speech and Language*, 20(2-3):303–330, 2006.

[Meinedo *et al.*, 2003] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso. Audimus.media: a broadcast news speech recognition system for the european portuguese language. In *PROPOR*, 2003.

[Messer and et al., 1999] K. Messer and et al. XM2VTSDB: The extended M2VTS Database. In *AVBPA*, 1999.

[Mirghafori and Wooters, 2006] N. Mirghafori and C. Wooters. Nuts and flakes: A study of data characteristics in speaker diarization. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processin, ICASSP*, pages 1017–1020, 2006.

[Moh *et al.*, 2003] Y. Moh, P. Nguyen, and J.C. Junqua. Toward domain independent speaker clustering. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 2, pages 85–88, April 2003.

[Moraru *et al.*, 2005] D. Moraru, M. Ben, and G. Gravier. Experiments on Speaker Tracking and Segmentation in Radio Broadcast News. In *Ninth Annual Conference of the International Speech Communication Association, Interspeech*. ISCA, 2005.

[Moreau *et al.*, 2008] Nicolas Moreau, Djamel Mostefa, Rainer Stiefelhagen, Susanne Burger, and Khalid Choukri. Data collection for the chil clear 2007 evaluation campaign. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Mori and Nakagawa, 2001] K. Mori and S. Nakagawa. Speaker change detection and speaker clustering using vq distortion for broadcast news speech recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 1, pages 413–416, 2001.

[Mostefa and et al., 2006] D. Mostefa and et al. Clear evaluation plan v1.1. Website,http://isl.ira.uka.de/~nickel/clear/downloads/chil-clear-v1.1-2006-02-21.pdf, 2006. http://isl.ira.uka.de/~nickel/clear/downloads/chil-clear-v1.1-2006-02-21.pdf.

[Mostefa and et al., 2007] D. Mostefa and et al. Clear evaluation plan 07 v0.1. Website, http://isl.ira.uka.de/clear07/?download=audio_id_2007_v0.1.pdf, 2007. http://isl.ira.uka.de/clear07/?download=audio_id_2007_v0.1.pdf.

[Nadeu *et al.*, 1995] C. Nadeu, J. Hernando, and M. Gorricho. On the Decorrelation of filter-Bank Energies in Speech Recognition. In *European Conference on Speech Communication and Technology, Eurospeech*, volume 20, page 417, 1995.

[Nadeu *et al.*, 1997] C. Nadeu, P. Paches-Leal, and B. H. Juang. Filtering the time sequence of spectral parameters for speech recognition. In *Speech Communication*, volume 22, pages 315–332, 1997.

[Nadeu *et al.*, 2001] C. Nadeu, D. Macho, and J. Hernando. Time and Frequency Filtering of Filter-Bank Energies for Robust Speech Recognition. In *Speech Communication*, volume 34, pages 93–114, 2001.

[National Institute of Standards and Technology, NIST, 1995] Website nist speaker recognition evaluation. Website, http://www.itl.nist.gov/iad/mig/tests/sre/, 1995.

[Neto *et al.*, 2008] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro. Broadcast news subtitling system in portuguese. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1561 –1564, 31 2008-april 4 2008.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856. MIT Press, 2001.

[Nickel *et al.*, 2005a] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *7th international booktitle on Multimodal interfaces*, pages 61–68. ACM Press New York, NY, USA, 2005.

[Nickel *et al.*, 2005b] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *Proc. IEEE Int. Conf. on Multimodal Interfaces (ICMI)*, pages 61–68, 2005.

[Ning *et al.*, 2006] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S. Huang. A spectral clustering approach to speaker diarization. In *INTERSPEECH'06*, pages –1–1, 2006.

[Ning *et al.*, 2010] Huazhong Ning, Wei Xu, Yun Chi, Yihong Gong, and Thomas S. Huang. Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recognition*, 43(1):113 – 127, 2010.

[Noulas and Krose, 2007] A.K. Noulas and B.J.A. Krose. On-line multi-modal speaker diarization. In *ICMI*, pages 350–357, 2007.

[Noulas *et al.*, 2011] A. Noulas, G. Englebienne, and B. Krose. Multimodal speaker diarization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1, 2011.

[Nwe *et al.*, 2008] T.L. Nwe, M. Dong, S.Z.K. Khine, and H. Li. Multi-speaker meeting audio segmentation. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 2522–2525, Brisbane, Australia, 2008.

[Nwe *et al.*, 2012] Tin Lay Nwe, Hanwu Sun, Bin Ma, and Haizhou Li. Speaker clustering and cluster purification methods for rt07 and rt09 evaluation meeting data. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):461 –473, feb. 2012.

[Oglesby and Mason, 1990] J. Oglesby and J.S. Mason. Optimisation of neural models for speaker identification. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 261 –264 vol.1, apr 1990.

[Omologo and Svaizer, 1997] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, 5(3):288–292, 1997.

[Otterson and Ostendorf, 2007] S. Otterson and M. Ostendorf. Efficient use of overlap information in speaker diarization. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, pages 683–686, Kyoto, Japan, 2007.

[Otterson, 2007] S. Otterson. Improved location features for meeting speaker diarization. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 1849–1852, Antwerp, Belgium, 2007.

[Ouellet *et al.*, 2004] P. Ouellet, G. Boulianne, and P. Kenny. Flavors of gaussian warping. In *9th European Conference on Speech Communication and Technology (Interspeech'2005-Eurospeech)*, pages 2957–2960, September 4-8 2004.

[Pardo *et al.*, 2007] J.M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, 56 No. 9, 2007.

[Pardo *et al.*, 2012] J.M. Pardo, R. Barra-Chicote, R. San-Segundo, R. de Cordoba, and B. Martinez-Gonzalez. Speaker diarization features: The upm contribution to the rt09 evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):426 –435, feb. 2012.

[Pelecanos and Sridharan, 2001] Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *The Speaker and Language Recognition Workshop, Odyssey*, 2001.

[Pellegrini and Trancoso, 2009] T. Pellegrini and I. Trancoso. Error detection in automatic transcriptions using hidden markov models. In *Language and Technology Conference*, 2009.

[Peng *et al.*, 2005] H.C. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 27 No. 8:1226–1238, 2005.

[Perez-Freire and C-Garcia-Mateo, 2004] L. Perez-Freire and C-Garcia-Mateo. A multimedia approach for audio segmentation in tv broadcast news. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2004.

[Peterson and Barney, 1951] G.E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1951.

[Pfau *et al.*, 2001] T. Pfau, D. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI meeting recorder. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, pages 107–110, Madonna di Campiglio, Italy, 2001.

[Picone, 1993] Joseph W. Picone. Signal modeling techniques in speech recognition. In *PROCEEDINGS OF THE IEEE*, pages 1215–1247, 1993.

[Pietquin *et al.*, 2002] O. Pietquin, L. Couvreur, and P. Couvreur. Applied clustering for automatic speaker-based segmentation of audio material. *Belgian Journal of Operations Research, Statistics and Computer Science*, 41 No. 1-2:69–81, 2002.

[Platt *et al.*, 2000] John C. Platt, Nello Cristianini, and John Shawe-taylor. Large margin dags for multiclass classification. In *Advances in Neural Information Processing Systems*, pages 547–553. MIT Press, 2000.

[Poh and Bengio, 2004] Norman Poh and Samy Bengio. Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks? In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-04)*, 0 2004.

[Potamitis *et al.*, 2003] I. Potamitis, G. Tremoulis, and N. Fakotakis. Multi-speaker DOA tracking using interactive multiple models and probabilistic data association. In *Proceedings of European Conference on Speech Communication and Technology*, 2003.

[Przybocki *et al.*, 2004] Przybocki, M. A., Martin, and A. F. Nist speaker recognition evaluation chronicles. In *The Speaker and Language Recognition Workshop, 2004. IEEE Odyssey*, 2004.

[Przybocki *et al.*, 2006] M.A. Przybocki, A.F. Martin, and A.N. Le. Nist speaker recognition evaluation chronicles - part 2. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1 –6, june 2006.

[Przybocki *et al.*, 2007] Mark A. Przybocki, Alvin F. Martin, and A. N. Le. Nist speaker recognition evaluations utilizing the mixer corpora - 2004, 2005, 2006. *IEEE Transactions on Audio, Speech & Language Processing*, 15(7):1951–1959, 2007.

[R. Yantorno, 2001] R. Yantorno. The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A usable speech measure emplyed as a co-channel detection system. In *IEEE Workshop on Intelligent Signal Processing*, 2001.

[Rabiner and Juang, 1986] L. A. Rabiner and B. H. Juang. An introduction to Hidden Markov Models. *IEEE Acoustics, Speech and Signal Processing Magazine*, 3:4–16, 1986.

[Rabiner and Juang, 1993] L.A. Rabiner and B. H. Juang. Fundamentals of speech recognition, signal processing. *Signal Processing*, 1993.

[Rabiner and Schafer, 1978] L.R. Rabiner and W. Schafer. *Digital Processing of Speech Signal*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[Rabiner, 1989] L. A. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[Rabinkin, 1995] D.V. Rabinkin. *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University, 1995.

[R.Brunelli and Falavigna, 1995] R.Brunelli and D. Falavigna. Person Identification Using Multiple Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 17, No. 10:955–966, 1995.

[Rentzeperis *et al.*, 2006] Elias Rentzeperis, Andreas Stergiou, Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros Polymenakos. The 2006 athens information technology speech activity detection and speaker diarization systems. In Steve Renals, Samy Bengio, and Jonathan Fiscus, editors, *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 385–395. Springer Berlin / Heidelberg, 2006. 10.1007/11965152_34.

[Reynolds and Torres-Carrasquillo, 2005] D.A. Reynolds and P. Torres-Carrasquillo. Approaches and Applications of Audio Diarization. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 5, 2005.

[Reynolds *et al.*, 2000] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, pages 19–41, 2000.

[Reynolds *et al.*, 2003] D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Qin Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, , and Bing Xiang. The supersid project: exploiting high-level information for high-accuracy speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2003.

[Reynolds, 1994] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, 1994.

[Reynolds, 1995] D. A. Reynolds. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions ASSP Magazine*, 3(1):72–83, 1995.

[Reynolds, 1996] D. A. Reynolds. The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 113–116, 1996.

[Reynolds, 1997] D.A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *European Conference on Speech Communication and Technology, Eurospeech*, 1997.

[Reynolds, 1998] D. A. Reynolds. Blind clustering of speech utterances based in speaker and language characteristics. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 1998.

[Reynolds, 2002] D.A. Reynolds. An overview of automatic speaker recognition technology. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2002.

[Reynolds, 2003] D.A. Reynolds. Channel robust speaker verification via feature mapping. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2003.

[Rodrguez-Liares *et al.*, 2003] Leandro Rodrguez-Liares, Carmen Garca-Mateo, and Jos Luis Alba-Castro. On combining classifiers for speaker authentication. *Pattern Recognition*, 36:347–359, 2003.

[Rosenberg *et al.*, 1992] A. Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong. The use of cohort normalized scores for speaker recognition. In *International Conference on Spoken Language Processing, ICSLP*, pages 599–602, 1992.

[Ross *et al.*, 2008] David Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1):125–141, May 2008.

[Rougi and et al., 2006] J. Rougi and et al. Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2006.

[Salah *et al.*, 2008a] A. A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten, and E.J. Pauwels. Multimodal identification and localization of users in a smart environment. *Journal on Multimodal User Interfaces*, 2(2):77–91, 2008.

[Salah *et al.*, 2008b] Albert Ali Salah, Ramon Morros, Jordi Luque, Carlos Segura, Javier Hernando, Onkar Ambekar, Ben Schouten, and Eric Pauwel. Multimodal identification and localization of users in a smart environment. *Journal on Multimodal User Interfaces*, 2(2):75–91, 2008.

[Sankar and et al., 1995] A. Sankar and et al. Training data clustering for improved speech recognition. In *European Conference on Speech Communication and Technology, Eurospeech*, 1995.

[Scheirer and Slaney, 1997] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331 –1334 vol.2, apr 1997.

[Schwart *et al.*, 1982] R. Schwart, L. S. Roucos., and M. Berouti. The application of probability denhity estimation to text-independt-nt speaker identification. *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 1649–1652, May 1982.

[Schwarz, 1973] G. Schwarz. A sequential student test. *The Annals of Statistics*, pages 1003–1009, 1973.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

[Schlkopf and Smola, 2002] B. Schlkopf and A. Smola. Learning with Kernels. In *MIT Press, Cambridge, MA*, 2002.

[Segundo, 2006] Rubn San Segundo. Propuesta de evaluacin de sistemas albayzin-2006: Segmentacin e identificacin de hablantes. Website, http://jth2006.unizar.es/evaluacion/albayzin06.html, 2006.

[Segura *et al.*, 2007] C. Segura, A. Abad, J. Hernando, and C. Nadeu. Multispeaker Localization and Tracking in Intelligent Environments. In *Lecture Notes on Computer Science, LNCS: second International CLEAR Evaluation Workshop 2007*, volume 4625. Springer-Verlag, 2007.

[Shaobing and Gopalakrishnan, 1998] C. Shaobing and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[Shriberg *et al.*, 2001] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *European Conference on Speech Communication and Technology, Eurospeech*, volume 2, pages 1359–1362, Aalborg, Denmark, 2001.

[Shriberg *et al.*, 2005] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3-4):455 – 472, 2005. ¡ce:title¿Quantitative Prosody

Modelling for Natural Speech Description and Generation¡/ce:title¿ ¡xocs:full-name¿International Conference on Speech Prosody¡/xocs:full-name¿.

[Shriberg, 2005] E. Shriberg. Spontaneous speech: How people really talk and why engineers should care. In *European Conference on Speech Communication and Technology, Eurospeech*, pages 1781–1784, Lisbon, Portugal, 2005.

[Siegler and et al., 1997] M.A. Siegler and et al. Automatic segmentation, classification and clustering of broadcast news audio. In *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 *European Conference on Speech Communication Technology*, pages 2437–2440, September 2005.

[Sjlander, 1997] Kre Sjlander. Snack toolkit v2.2.10. WebSite, http://www.speech.kth.se/snack/, 1997.

[Skosan and Mashao, 2006] Marshalleno Skosan and Daniel Mashao. Modified segmental histogram equalization for robust speaker verification. *Pattern Recognition Letters*, 27(5):479 – 486, 2006.

[Slomka *et al.*, 1998] S. Slomka, Sridharan, S., and V. Chandran. A comparison of fusion techniques in mel-cepstral based speaker idencication. In *In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, pages 225–228, November 1998.

[smartflow, 2002] Smartflow system, 2002.

[Snir *et al.*, 1998] Marc Snir, Steve Otto, Steven Huss-Lederman, David Walker, and Jack Dongarra. *MPI-The Complete Reference, Volume 1: The MPI Core*. MIT Press, Cambridge, MA, USA, 2nd. (revised) edition, 1998.

[Solomonoff *et al.*, 2007] Alex Solomonoff, William M. Campbell, and Carl Quillen. Nuisance attribute projection, speech communication. *Elsevier Science BV, Amsterdam, The Netherlands*, May 2007.

[Soong *et al.*, 1985] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang. A vector quantization approach to speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 387–390, March 1985.

[Stafylakis *et al.*, 2010] T. Stafylakis, V. Katsouros, and G. Carayannis. The segmental bayesian information criterion and its applications to speaker diarization. *Selected Topics in Signal Processing, IEEE Journal of*, 4(5):857 –866, oct. 2010.

[Stafylakis1 and Katsouros, 2011] Themos Stafylakis1 and Vassilis Katsouros. *Speech and Language Techonologies: Chapter 12, A Review of Recent Advances in Speaker Diarization with Bayesian Methods*. Ivo Ipši, 2011.

[Stanford *et al.*, 2003] V. Stanford, J. Garofolo, O. Galibert, M. Michel, and C. Laprun. The NIST smart space and meeting room projects: Signals, acquisition, annotation, and metrics. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 4, pages 736–739, 2003.

[Stevens, 1937] S.S. Stevens. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[Stiefelhagen *et al.*, 2007] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R. Travis Rose, Martial Michel, and John Garofolo. The clear 2007 evaluation. In *Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625, pages 3–34. Springer Lecture Notes in Computer Science, 2007.

[Stolcke *et al.*, 2005] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. Mllr transforms as features in speaker recognition. In *European Conference on Speech Communication and Technology, Eurospeech*, pages 2425–2428, 2005.

[Sundaram *et al.*, 2003] N. Sundaram, R.E. Yantorno, B.Y. Smolenski, and A.N. Iyer. Usable speech detection using linear predictive analysis - a model based approach. In *International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS*, pages 231–235, Awaji Island, Japan, 2003.

[Svaizer and others, 1997] P. Svaizer et al. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 231–234, Munich, Germany, 1997.

[Swartling and et al., 2006] M. Swartling and et al. Direction of Arrival Estimation for Multiple Speakers Using Time-Frequency Orthogonal Signal Separation. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 4, pages 833–836, 2006.

[T. Gustafsson and B. Rao and M. Trivedi, 2003] T. Gustafsson and B. Rao and M. Trivedi. Source Localization in Reverberant Environments: Modeling and Statistical Analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6):791–803, 2003.

[Tangelder and Schouten, 2006] JWH Tangelder and BAM Schouten. Sparse face representations for face recognition in smart environments. In *International Conference on Pattern Recognition, ICPR*, 2006.

[Tannenbaum *et al.*, 2001] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. Condor – a distributed job scheduler. In Thomas Sterling, editor, *Beowulf Cluster Computing with Linux*. MIT Press, October 2001.

[Teh *et al.*, 2004] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2004.

[Temko and Nadeu, 2006] A. Temko and C. Nadeu. Classification of acoustic events using svm-based clustering schemes. *Pattern Recognition*, 39:682–694, 2006.

[Temko *et al.*, 2007] A. Temko, D. Macho, and C. Nadeu. Enhanced SVM training for robust speech activity detection. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2007.

[Temko *et al.*, 2008] Andrey Temko, Dusan Macho, and Climent Nadeu. Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. *Pattern Recognition*, 41(5):1814–1823, 2008.

[Thomas P. Wilson and Zimmerman, 1984] John M. Wiemann Thomas P. Wilson and Don H. Zimmerman. Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3 No. 3:159–183, 1984.

[Tranter and Reynolds, 2004] S. Tranter and D.A. Reynolds. Speaker diarization for broadcast news. In *The Speaker and Language Recognition Workshop, Odyssey*, 2004.

[Tranter and Reynolds, 2006] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557 –1565, sept. 2006.

[Tsai and Wang, 2006] W.-H. Tsai and H.-M. Wang. On maximizing the within-cluster homogeneity of speaker voice characteristics for speech utterance clustering. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2006.

[Valente and Wellekens, 2004] F. Valente and C. Wellekens. Variational bayesian speaker clustering. In *The Speaker and Language Recognition Workshop, Odyssey*, 2004.

[Valente *et al.*, 2010] F. Valente, P. Motlicek, and D. Vijayasenan. Variational bayesian speaker diarization of meeting recordings. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4954 –4957, march 2010.

[Valente *et al.*, 2011] Fabio Valente, Deepu Vijayasenan, and Petr Motlcek. Speaker diarization of meetings based on speaker role n-gram models. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 4416–4419. IEEE, 2011.

[van Leeuwen and Huijbregts, 2006] D. van Leeuwen and M. Huijbregts. The AMI speaker diarization system for NIST RT06s meeting data. In *Machine Learning for Multimodal Interaction*, volume 4299/2006, pages 371–384. Springer Berlin/Heidelberg, 2006.

[van Leeuwen and Konečný, 2008] David van Leeuwen and Matej Konečný. Progress in the amida speaker diarization system for meeting data. In Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, editors, *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 475–483. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-68585-2_44.

[Vandecatseye and et al., 2004] Vandecatseye and et al. The cost278 pan-european broadcast news database. In *LREC*, 2004.

[Vapnik, 1998] V. N. Vapnik. *Statistical learning theory*. John Wiley & Sons, New York, 1998.

[Vijayasenan *et al.*, 2007] D. Vijayasenan, F. Valente, and H. Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. In *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on*, pages 250 –255, dec. 2007.

[Vijayasenan *et al.*, 2008] D. Vijayasenan, F. Valente, and H. Bourlard. Combination of agglomerative and sequential clustering for speaker diarization. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4361 –4364, 31 2008-april 4 2008.

[Vijayasenan *et al.*, 2009] D. Vijayasenan, F. Valente, and H. Bourlard. An information theoretic approach to speaker diarization of meeting data. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(7):1382 –1393, sept. 2009.

[Vijayasenan *et al.*, 2011] Deepu Vijayasenan, Fabio Valente, and Herv Bourlard. Multistream speaker diarization of meetings recordings beyond mfcc and tdoa features. *Speech Communication*, 54(1):55 – 67, 2011.

[Vilaplana *et al.*, 2006] V. Vilaplana, C. Martínez, J. Cruz, and F. Marques. Face recognition using groups of images in smart room scenarios. In *International Conference on Image Processing, ICIP*, 2006.

[Vinyals and Friedland, 2008] O. Vinyals and G. Friedland. A hardware-independent fast logarithm approximation with adjustable accuracy. In *Proceedings of IEEE International Symposium on Multimedia (ISM 08)*, 2008.

[Viola and Jones, 2001] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, volume 1, pages 511–518, 2001.

[Wactlar *et al.*, 1996] Howard D. Wactlar, Er G. Hauptmann, and Michael J. Witbrock. Informedia tm: News-on-demand experiments in speech recognition. In *ARPA SLT workshop*, 1996.

[Wan and Renals, 2005] V. Wan and S. Renals. Speaker verification using sequence discriminant support vector machines. *Speech and Audio Processing, IEEE Transactions on*, 13(2):203 – 210, march 2005.

[Wegmann *et al.*, 1999] Steven Wegmann, Francesco Scattone, Ira Carp, Larry Gillick, Robert Roth, and Jon Yamron. Dragon systems' 1997 broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 60–65, 1999.

[Whu and et al., 2005] Z. Whu and et al. Multi-level fusion of audio and visual features for speaker identification. In *ICB 2006*, volume 3832, pages 493–499, 2005.

[Wiener, 1949] Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley, 1949.

[Wilcox *et al.*, 1994] L. Wilcox, F. Chen, D. Kumber, and V. Balasubramanian. Segmentation of speech using speaker identification. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, volume 41 No. 1-2, pages 161–164, 1994.

[Willsky and Jones, 1976] A.S. Willsky and H.L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21:161–164, 1976.

[Wolf, 1972] J.J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51:2044–2056, 1972.

[Woodland *et al.*, 1998] Philip Woodland, Thomas Hain, Sue Johnson, S. E. Johnson, Thomas Niesler, Steve Young, Andreas Tuerk, and S. J. Young. Experiments in broadcast news transcription. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 909–912, 1998.

[Woodland, 2002] P.C. Woodland. The development of the HTK broadcast news transcription system: An overview. In *Speech Communcations*, volume 37, pages 47–67, 2002.

[Wooters and et al., 2004] C. Wooters and et al. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In *Fall 2004 Rich Transcription Workshop (RT04)*, 2004.

[Wooters and Huygbregts, 2008] C. Wooters and M. Huygbregts. The ICSI RT07s Speaker Diarization System. In *Lecture Notes on Computer Science, LNCS: CLEAR 2007 and RT 2007*, volume 4625. Springer-Verlag, 2008.

[Wrigley *et al.*, 2005] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals. Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, 13(1):84–91, 2005.

[Yamashita and Matsunaga, 2005] M. Yamashita and S. Matsunaga. Spectral cross-correlation features for audio indexing of broadcast news and meetings. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2005.

[Yilmaz and Rickard, 2004] O. Yilmaz and S. Rickard. Blind source separation based on space-time-frequency masking. *IEEE Transactions on Signal Processing*, 52 No. 7:1830–1847, 2004.

[Young *et al.*, 1993] S.J. Young, Woodland P.C., and Byrne W.J. Htk version 3.4.1: User, reference and programmer manual., September 1993.

[Yuan *et al.*, 2004] Wang Yuan, Wangm Yunhong, and T. Tan. Combining Fingerprint and Voiceprint Biometrics for Identity Verification: an Experimental Comparison. In *ICBA*, pages 663–670, 2004.

[Zelenák *et al.*, 2010] M. Zelenák, C. Segura, and J. Hernando. Overlap detection for speaker diarization by fusing spectral and spatial features. In *Annual Conference of the International Speech Communication Association, Interspeech*, pages 2302–2305, Makuhari, Japan, 2010.

[Zelenák *et al.*, 2011] M. Zelenák, C. Segura, J. Luque, and J. Hernando. Simultaneous speech detection with spatial features for speaker diarization. *IEEE Transactions on Audio, Speech and Language Processing*, 2011.

[Zhang and et al., 2004] D. Zhang and et al. An adaptative model of person identification combining speech and image information. In *ICARCV*, 2004.

[Zheng *et al.*, 2005] R. Zheng, S. Zhang, and B. Xu. A comparative study of feature and score normalization for speaker verification. In *Lecture Notes in Computer Science, LNCS*, volume 2832/2005, pages 531–538. Springer Berlin/Heidelberg, 2005.

[Zhu and et al., 2005] X. Zhu and et al. Combining speaker identification and bic for speaker diarization. In *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP*, 2005.

[Zhu *et al.*, 2008] X. Zhu, C. Barras, L. Lamel, and J-L. Gauvain. Multi-stage speaker diarization for conference and lecture meetings. In Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, editors, *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 533–542. Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-68585-2_49.

[Zochova and Radova, 2005] P. Zochova and V. Radova. Modified distbic algorithm for speaker change detection. In *International Conference on Spoken Language Processing, ICSLP*, 2005.