ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-019

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité "Signal et Image"

présentée et soutenue publiquement par

Simon BOZONNET

le 2 Mai 2012

New Insights into Hierarchical Clustering and Linguistic Normalization for Speaker Diarization

Nouveaux points de vue sur la classification hiérarchique et normalisation pour la segmentation et le regroupement en locuteurs

THÈSE

Directeur de thèse: **Dr. Nicholas EVANS**
Co-encadrement de la thèse: **Prof. Bernard MERIALDO**

M. Jean-François BONASTRE, Professeur, LIA, Université d'Avignon, France
M. Laurent BESACIER, Professeur, LIG, Grenoble, France
M. John S. D. MASON, Professeur, Swansea University, UK
M. Xavier ANGUERA, Docteur, Telefonica, R&D Universitat Pompeu Fabra, Spain

Abstract

The ever-expanding volume of available audio and multimedia data has elevated technologies related to content indexing and structuring to the forefront of research. Speaker diarization, commonly referred to as the 'who spoke when?' task, is one such an example and has emerged as a prominent, core enabling technology in the wider speech processing research community. Speaker diarization involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering). Much progress has been made in the field over recent years partly spearheaded by the NIST Rich Transcription (RT) evaluations focus on meeting domain, in the proceedings of which are found two general approaches: top-down and bottom-up. The bottom-up approach is by far the most common, while very few systems are based on top-down approaches.

Even though the best performing systems over recent years have all been bottom-up approaches we show in this thesis that the top-down approach is not without significant merit. Indeed we first introduce a new purification component, improving the robustness of the top-down system and bringing an average relative Diarization Error Rate (DER) improvement of 15% on independent datasets, leading to competitive performance to the bottom-up approach. Moreover, while investigating the two diarization approaches more thoroughly we show that they behave differently in discriminating between individual speakers and in normalizing unwanted acoustic variation, i.e. that which does not pertain to different speakers. This difference of behaviours leads to a new top-down/bottom-up system combination outperforming the respective baseline systems. Finally, we introduce a new technology able to limit the influence of linguistic effects, responsible for biasing the convergence of the diarization system. Our novel approach is referred to as Phone Adaptive Training (PAT) by comparison to Speaker Adaptive Training (SAT) and shows an improvement of 11% relative improvement in diarization performance.

Résumé

Face au volume croissant de données audio et multimédia, les technologies liées à l'indexation de données et à l'analyse de contenu ont suscité beaucoup d'intérêt dans la communauté scientifique. Parmi celles-ci, la segmentation et le regroupement en locuteurs, répondant ainsi à la question 'Qui parle quand ?' a émergé comme une technique de pointe dans la communauté de traitement de la parole. D'importants progrès ont été réalisés dans le domaine ces dernières années principalement menés par les évaluations internationales du NIST (*National Institute of Standards and Technology*). Tout au long de ces évaluations, deux approches se sont démarquées : l'une est *bottom-up* et l'autre *top-down*. L'approche *bottom-up* est de loin la plus courante alors que seulement quelques systèmes sont basés sur l'approche dite *top-down*.

L'ensemble des systèmes les plus performants ces dernières années furent essentiellement des systèmes types *bottom-up*, cependant nous expliquons dans cette thèse que l'approche *top-down* comporte elle aussi certains avantages. En effet, dans un premier temps, nous montrons qu'après avoir introduit une nouvelle composante de purification des clusters dans l'approche *top-down*, nous obtenons une amélioration des performances de 15% relatifs sur différents jeux de données indépendants, menant à des performances comparables à celles de l'approche *bottom-up*.

De plus, en étudiant en détails les deux types d'approches nous montrons que celles-ci se comportent différemment face à la discrimination des locuteurs et la robustesse face à la composante lexicale. Ces différences sont alors exploitées au travers d'un nouveau système combinant les deux approches. Enfin, nous présentons une nouvelle technologie capable de limiter l'influence de la composante lexicale, source potentielle d'artefacts dans le regroupement et la segmentation en locuteurs. Notre nouvelle approche se nomme *Phone Adaptive Training* par analogie au *Speaker Adaptive Training* utilisé pour la reconnaissance de la parole et montre une amélioration de 11% relatifs par rapport au performances de référence.

“ The most exciting phrase to hear in science,
the one that heralds new discoveries, is not ’Eureka!’ (I found it!)
but ’That’s funny ...’ ”
(Isaac Asimov 1920 - 1992)

Acknowledgements

Research is like a game where no one really wins the only difference being that we never reach a state where we you could say, “okay, now the game is over”. However research is like a small victory every time we succeed in getting something new. Unexpected results can be good or bad, they will certainly lead to promising interpretations because they are unexpected! Since finally research is in my opinion not “How many?”, “How much?”, but “Why?” or “How?”.

Due to its attractiveness research is however time consuming and so I would like to thank a lot people who show understanding and support during my PhD. First I would like to thank my supervisor Dr. Nick Evans who always found a way to be available for discussions during my PhD and so I owe him much. I have to thank my co-supervisor Prof. Bernard Merialdo and the jury committee, some of whom travelling more than thousand kilometres to come to my defense, namely: Prof. Jean-François Bonastre, Prof. Laurent Besacier, Prof. John S. D. Mason and Dr. Xavier Anguera. Additionally I have to be very grateful to Dr. Corinne Fredouille who always replied my numerous emails and phone calls.

I have to say thank as well to my co-authors and colleague with who I worked on some different projects and/or papers, namely Oriol Vinyals, Mary Knox from the other side of the ocean, Jürgen Geiger from the cold Germany, Marijn Huijbregts and Félicien Vallet.

If I succeeded in my work it is thanks to my EURECOM’s colleagues also who were always pretty kind with me. I have first to formulate some huge thanks to Dr. Ravi Vipperla and Dr. Dong Wang who helped me a lot and with who I had often some interesting discussions! Many thanks to my officemates, Hajer, Rui, Rachid, Angela and Rémi who supported me during several

years without forgetting my other colleagues from the speech group: Moctar, Christelle, Federico, those with who I will/play(ed) music: Adrien, Xuran, Claudiu and more generally from the multimedia department: Antitza, Claudia, Safa, Giovanna, Houda, Nesli, Neslihan, Miriam, Mathilde, Jessica, Lionel, Xueliang, Yingbo, Usman (N.), Usman (S.), Ghislain, Giuseppe, Jose, Carmelo, Marco (P.). And since we are not that sectarian in this department I have to thank a lot as well Daniel & Carina, Sabir, Thomas, Tomek, Hendrik, Quentin, Gabriel, Tania, Ayse, Lei, Lorenzo, Faouzi, Marco (B.), Chen JB, Jelena, Wael, Daniel (C.). Two fervent supporters were my two German flatmates: Adrian & Fabian! Thank you both!

I wont forget my friends "from Lyon": Cécile, Camille, Chloé, Fannie, Mathilde, Marie, Perrine, Elisabeth, Emile, Stéphane, Olivier, Swann, Martin, Mathieu, Michael, Anthony, Delphine, Nicolas, Matthieu, neither those from Oyonnax: Nicos, Zoom, Sylvie, Arnaud, Yoann, Maurice and the musicians from Artfull: Luca, Anaïs, Fred, JP, Harry, Adrien and Sophie! And I am sure I unfortunately forgot some other friends... ...sorry for that!

Finally since Music was an important parameter for me to keep my hopes up I have to thank Nicole Blanchi and the "Choeur Régional PACA" with who I had the chance to be involved in an important number of prestigious concerts. On the same way, I have to thank Jean-François Jacomino (or Jeff!) and all the friends from the Big Band JMSU, where we had the chance to be enrolled in a collection of pleasant concerts!

And at last, I would like obviously to thank my family for their support during all this challenge, despite the distance they were always present!

Contents

List of Figures

List of Tables

Glossary

ADM All Distant Microphones

AHC Agglomerative Hierarchical Clustering

AIB Agglomerative Information Bottleneck

ASPG Adaptive Seconds Per Gaussian

ASR Automatic Speech Retranscription

CLR Cross Likelihood Ratio

DBN Dynamic Bayesian Network

DER Diarization Error Rate

DHC Divisive Hierarchical Clustering

DP Dirichlet Process

E-HMM Evolutive Hidden Markov Model

EM Expectation Maximization

GE Grand Échiquier (Name of the French TV-show corpus)

GLR Generalized Likelihood Ratio

GMM Gaussian Mixture Model

GSC Generalized Side-lobe Canceller

HDP Hierarchical Dirichlet Process

HMM Hidden Markov Model

ICD Inter-Channel Delay

ICR Information Change Rate

IHM Individual Headphone Microphones

IQR Inter-Quartile Range

KL Kullback-Leibler Divergence

KL2 Symmetric alternative of the Kullback-Leibler Divergence

LDA Linear Discriminant Analysis

LPC Linear Predictive Coefficient

MAP Maximum A Posteriori

MCMC Monte Carlo Markov Chains

MDM Multiple Distant Microphones

MFCC Mel-Frequency Cepstral Coefficients

MLLR Maximum Likelihood Linear Regression

MM3A Multiple Mark III Arrays

NIST National Institute of Standards and Technology

Over-clustering Producing less clusters than required

PAT Phone Adaptive Training

PDF Probability Density Function

PLP Perceptual Linear Prediction

RT Rich Transcription

SAD Speech Activity Detector

SAT Speaker Adaptive Training

SDM Single Distant Microphone

SIB Sequential Information Bottleneck

SNR Signal-to-Noise Ratio

SVM Support Vector Machines

TDOA Time-Delay-Of-Arrival

UBM Universal Background Model

Under-clustering Producing more clusters than required

VTLN Vocal Track Length Normalization

List Of Publications

Journal

- N. Evans, S. Bozonnet, D. Wang, C. Fredouille and R. Troncy. *A Comparative Study of Bottom-Up and Top-Down Approaches to Speaker Diarization.* IEEE Transactions On Audio Speech, and Language Processing (TASLP) special issue on New Frontiers in Rich Transcription, February 2012, Volume 20 no. 2.
- X. Anguera, S. Bozonnet, N. W. D. Evans, C Fredouille, G. Friedland and O. Vinyals. *Speaker diarization : A review of recent research.* IEEE Transactions On Audio Speech, and Language Processing (TASLP) special issue on New Frontiers in Rich Transcription, February 2012, Volume 20 no. 2.

Conference/Workshop

- S. Bozonnet R. Vipperla and N. Evans. *Phone Adaptive Training for Speaker Diarization.* Submitted to Interspeech 2012
- S. Bozonnet, D. Wang, N. Evans and R. Troncy. *Linguistic influences on bottom-up and top-down clustering for speaker diarization.* In ICASSP 2011, 36th International Conference on Acoustics, Speech and Signal Processing, May 22-27, 2011, Prague, Czech Republic, Prague, Czech Republic, 05 2011.
- S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland and C. Fredouille. *System output combination for improved speaker diarization.* In Proc. Interspeech 2010, 11th Annual Conference of the International Speech Communication Association, September 26-30, Makuhari, Japan

- S. Bozonnet, N. Evans, C. Fredouille, D. Wang and R. Troncy. *An Integrated Top-Down/Bottom-Up Approach To Speaker Diarization. In Proc.* Interspeech 2010, 11th Annual Conference of the International Speech Communication Association, September 26-30, Makuhari, Japan
- 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 Proc. ICASSP, Dallas, Texas, USA, March 14-19 2010.
- S. Bozonnet, F. Vallet, Evans, N. W. D. Evans, S. Essid, G. Richard, J. Carrive. *A Multimodal approach to initialisation for top-down speaker diarization of television shows*, EUPSICO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark
- C. Fredouille, S. Bozonnet and N. W. D. Evans. *The LIA-EURECOM RT'09 Speaker Diarization System.* In RT'09, NIST Rich Transcription Workshop, 2009, Melbourne, Florida.
- J. Geiger, V. Ravichander, S. Bozonnet, N. Evans, B. Schuller, G. Rigoll. *Convolutive Non-Negative Sparse Coding and Advanced Features for Speech Overlap Handling in Speaker Diarization* Submitted to Interspeech 2012
- R. Vipperla, J. Geiger, S. Bozonnet, W. Dong, N. W. D. Evans, B. Schuller; G. Rigoll. *Speech overlap detection and attribution using convolutive non-negative sparse coding.* In ICASSP 2012, 37th International Conference on Acoustics, Speech and Signal Processing, March 25-30, 2012, Kyoto, Japan
- R. Vipperla, S. Bozonnet, W. Dong, Evans, N. W. D. Evans, *Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization.* CHiME 2011, 1st International Workshop on Machine Listening in Multisource Environments, Interspeech, September 1st, 2011, Florence, Italy

Chapter 1

Introduction

1.1 Motivations

Since the late 20^{th} century, the mass of multimedia information has increased exponentially. In 2011-2012, statistics[1] show that an average of 60 hours of video is uploaded to *YouTube* every minute or the equivalent of 1 hour every second. 4 billion videos are watched every day. According to the evolution shown in Figure 1.1, this is twice more than in 2010 and we can still expect these numbers to grow year-after-year as the profiles of the curves infer.

To face the problem of processing huge amounts of multimedia information, automatic data indexing and content structuring are the only strategy. Different approaches exist already, mainly based on the video content analysis [Truong & Venkatesh, 2007]. However video uploaded on video-sharing websites come from devices of different natures including webcams, mobile phones, HD cameras, or homemade video clips involving the merging of audio and video streams which may not be originally recorded together, e.g. the video content can be a slideshow and cannot be considered as a real video.

A way to analyze the structure and annotate the different types of video for their indexation is to extract information from the audio stream, in order to, eventually, feed a fully video system in a second step. A collection of techniques aim to achieve the extraction of the audio information, they include emotion recognition, acoustic event detection, speaker recognition, language detection, speech recognition or speaker diarization. Whereas speaker and speech recognition correspond to, respectively, the recognition of

[1] source: `http://www.youtube.com/t/press_timeline`

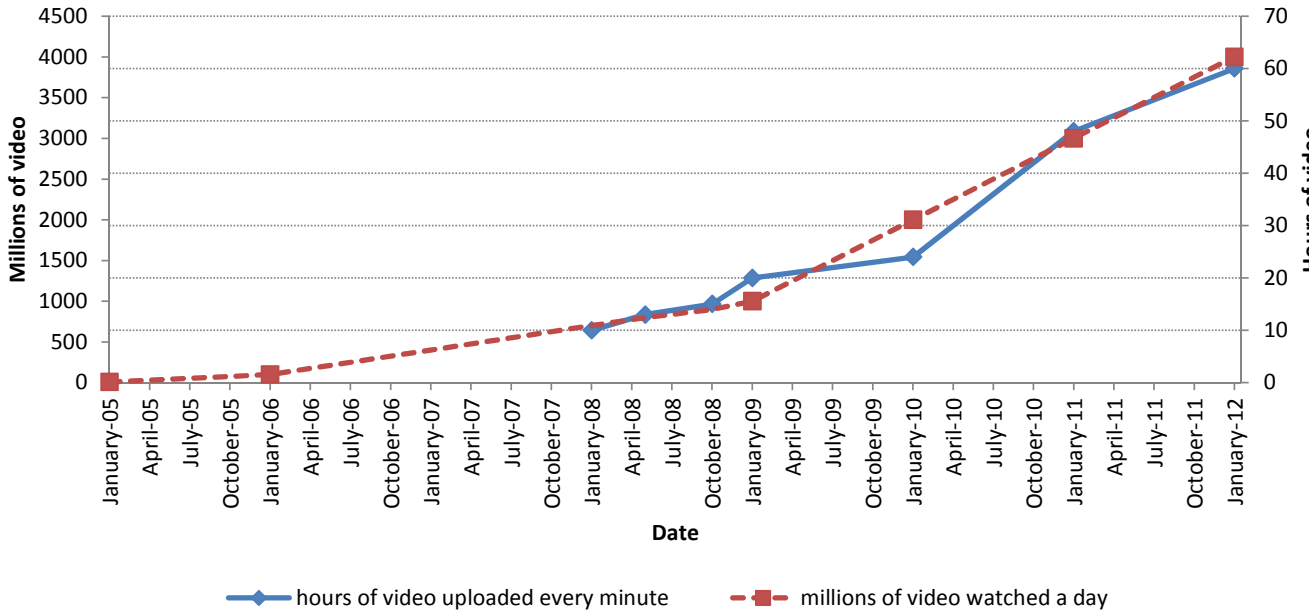


Figure 1.1: Evolution of the number of hours of video uploaded on *YouTube* from 2005 to 2012 (plain curve), and the millions of video watched per day (dashed line). Statistics issued from: `http://www.youtube.com/t/press_timeline`. Note that no data is available from 2005 to 2007 concerning the quantity of video uploaded every minute.

a person's identity or the transcription of their speech, speaker diarization relates to the problem of determining 'who spoke when'. More formally this requires the unsupervised identification of each speaker within an audio stream and the intervals during which each speaker is active.

Compared to music or other acoustic events, speech, due to its semantic content, is one of if not he most informative components in the audio stream. Indeed, speech transcription brings key information about the topic, while speaker recognition and/or speaker diarization reveal the speaker identities[1] through voice features. Due to it unsupervised nature, speaker diarization has utility in any application where multiple speakers may be expected and has emerged as an increasingly important and dedicated domain of speech research.

Indeed, speaker diarization first permits to index and extract the speakers in an audio stream in order to retrieve relevant information. Moreover, when some speaker a priori information is known, speaker diarization can be used as a preprocessing for

[1] or relative identities in the case of the unsupervised task of speaker diarization

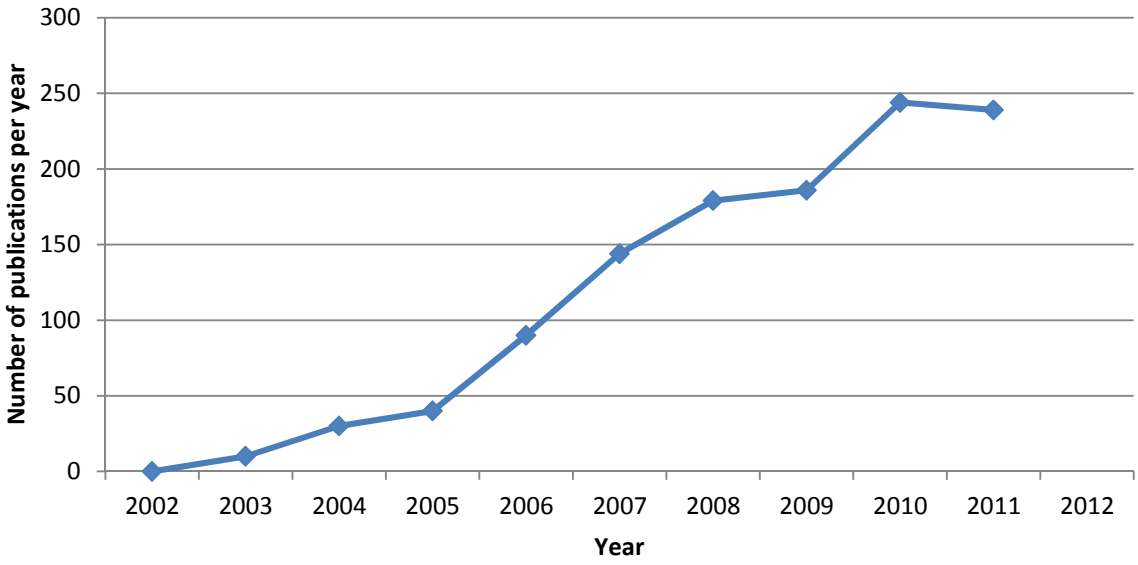


Figure 1.2: Number of citations per year in the field of Speaker Diarization. Source: *Google Scholar*

the task of speaker recognition to then determine the absolute identity of the speaker. Additionally, speaker diarization is considered as an important preprocessing step for Automatic Speech Retranscription (ASR) insofar as information about the speaker facilitates speaker adaptation e.g. Vocal Track Length Normalization (VTLN), Speaker Adaptative Training (SAT). Then, speaker specific speech models help to provide more accurate retranscription outputs.

The task of speaker diarization is thus a prerequisite, enabling technology relevant to audio indexation, content structuring, automatic annotation or more generally, Rich Transcription (RT), either providing direct information about the structure and speaker content indexing or helping in a pre-processing step for speech retranscription or speaker recognition.

1.2 Objective of This Thesis

Speaker diarization is not a new topic and research in the field started mainly around 2002. As we observe in Figure 1.2, the number of publications in speaker diarization has increased year-after-year, showing the raising interest of the community and importance of the field. Among the different challenges tackled by the community, four main domains

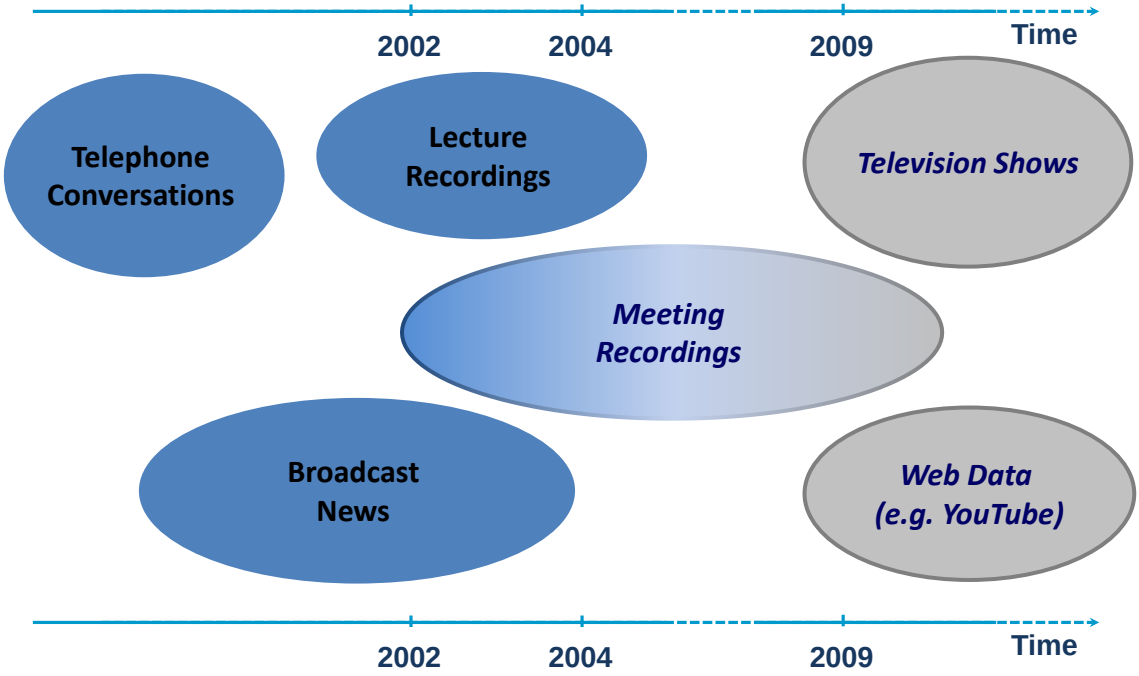


Figure 1.3: Different domains of application for the task of Speaker Diarization.

were addressed. In early 2000, the community first focused on telephone discussions (see Figure 1.3), which corresponds to a specific diarization challenge insofar as the number of speakers is known. Then the community turned to Broadcast News, including one dominant speaker and a few minor speakers. Around 2002 and 2004, the focus moved to lecture recordings and then meeting recordings. Meeting recordings, due to higher number of speakers and spontaneous speech (in comparison to the Broadcast News domain where the dialog is often scripted) becomes the most challenging diarization task and became the main focus of the community since 2004. Some other domains still deserve to be addressed, namely TV-shows, or more generally data issued from websites like *YouTube*.

This thesis relates to speaker diarization for meeting recordings since research in this domain is still very active, and meeting recordings are the focus of the recent international evaluations, this enables the comparison of performance with other state-of-the-art systems. Moreover, we have to highlight that meeting recordings, due to their specific characteristics, can be considered as general enough in terms of number of speakers and spontaneity of speech, and can be representative enough of an extensive part of the data

available on the Web.

Much progress has been made in the field over recent years partly spearheaded by the international NIST evaluations where two general approaches stand out: they are top-down and bottom-up. The bottom-up approach is by far the most common, while very few systems are based on top-down approaches.

Even though the best performing systems over recent years have all been bottom-up approaches, we want to show in this thesis that the top-down approach is not without significant merit and that each approach have its own benefits. The objective of this thesis can be formulated as follows:

- Is the bottom-up or top-down approach superior to the other?
- How do their behaviors differ?
- What are their specific weaknesses?
- How can we take the benefit of their behavioral differences?

1.3 Contributions

The main contributions of this thesis are four-fold. They are:

(i) a new post-purification process which, applied to the top-down approach, brings significant improvements in speaker diarization performance and makes the top-down approach comparable to the bottom-up scenario in terms of DER performance;

(ii) a comparative study which aims to show the differences in behaviors between the top-down and the bottom-up systems in a common framework and a set of Oracle experiments;

(iii) an integrated and a fused top-down/bottom-up system which confirm that, due to their different natures, the combination of the top-down/bottom-up systems brings improved performance which outperforms the original baselines;

(iv) a new phoneme normalization method which brings significant improvements on speaker diarization system.

The four contributions are described in more detail in the following.

(i) Novel Approach to cluster purification for Top-Down speaker diarization

Cluster purification is not a new topic in the field of speaker diarization, however previous works focus on the cluster purification of bottom-up systems. The first contribution of this thesis proposes a new purification component which is embedded in the top-down system baseline. It delivers improved stability across different datasets composed of conference meeting from five standard NIST evaluations and brings an average relative DER improvement of 15% on independent meeting datasets.

This work was presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2010 [Bozonnet et al., 2011].

(ii) Comparative study of Bottom-Up and Top-Down systems

The second contribution of this thesis is an analysis of the two different bottom-up and top-down clustering approaches otherwise known as agglomerative and divisive hierarchical clustering. Indeed, experimental results show that the purification work presented in the first contribution brings inconsistent improvements when applied to the bottom-up approach leading us to believe that each system has a specific behavior due to its particular nature. In order to set a complete and consistent analysis, two types of study are reported: an Oracle survey which aims to highlight the weaknesses of each system and a second survey which focuses more on the differences in convergence due to the different clustering scenarios. This study helps to understand the negative effect caused by the purification algorithm while applied on the bottom-up system.

- **Oracle Experiments**

  With the help of a set of Oracle experiments, sensitivity and robustness of the different components of the top-down baseline are analyzed in order to identify their possible weaknesses. The same framework is used for the bottom-up system. Experimental results show that, despite some common weaknesses mainly related to SAD performance and overlapping speech, both clustering algorithms present some specific shortcomings. Indeed, while the bottom-up scenario is almost independent to initialization, it is mainly sensitive to the merging and stopping criteria, particularly in case of cluster impurity. In

contrast, the top-down scenario is mainly sensitive to initialization and to the quality of the initial model which influences its discriminative capacity.

- **Behavior analysis and differences in terms of convergence**
  The second part of this analysis aims to focus on the effects in terms of convergence due to the bottom-up or top-down clustering direction. A theoretical framework including a formal definition of the task of speaker diarization and an analysis of the challenges that must be addressed by practical speaker diarization system are first derived leading us to believe that, theoretically, the final output should not depend on the clustering direction.

  However, we showed that, while ideally the models of a diarization system should be mainly speaker discriminative and independent of unwanted acoustic variations e.g. phonemes, the merging and splitting operations in the clustering process are likely to impact upon the discriminative power and phone-normalization of the intermediate and final speaker models, leading in practice to different behaviors and relative strengths and shortcomings. Indeed, our study shows that top-down systems are often better normalized toward phonemes and then more stable but suffer from lower speaker discrimination. This explains why they are likely to benefit from purification. In contrast, bottom-up clusterings are more speaker discriminative but, as a consequence of progressive merging, they can be sensitive to phoneme variations possibly leading to a non-optimal local maxima of the objective function.

This work was presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2011 [Bozonnet et al., 2011]. An extended version of the work including a more complete analysis is published in the IEEE Transactions on Audio Speech and Language Processing (TALSP), special issue on New Frontiers in Rich Transcription in 2012 [Evans et al., 2012].

(iii) Top-Down / Bottom-up combination system

The previous contribution highlights the distinct properties in terms of model reliability and discrimination of the bottom-up and top-down approaches. These specific behaviors suggest that there is some potential for system combination.

The third contribution of this thesis presents some novel ways to combine the top-down and bottom-up approaches harnessing the strengths of each system and thus to improve performance and stability. Two system combinations have been investigated:

- **Fused system**

  The fused system aims to run simultaneously and independently the top-down and bottom-up systems in order to then combine their outputs. We proposed a new approach which first maps the different clusters extracted from each of the system outputs based on some constraints on their confusion matrix and on their acoustic contents. Thanks to this mapping, a first selection of clusters is made. Then, some iterative unmatched clusters are introduced according to their acoustic distances to the mapped clusters where only the most confident frames are kept. A final realignment is made to associate the unclassified frames. Thanks to this scenario we achieved up to 13% relative improvement in diarization performance.

  This work was presented at the Annual Conference of the International Speech Communication Association (Interspeech) in 2010 [Bozonnet et al., 2010], and a deeper analysis of the effect of the system fusion was published in the IEEE Transactions on Audio Speech and Language Processing (TALSP) , special issue on New Frontiers in Rich Transcription in 2012 [Evans et al., 2012].

- **Integrated system**

  An alternative approach to combine the top-down and bottom-up systems is an integrated approach which aims to fuse the two systems at their heart. The systems are run simultaneously, the top-down system calling the bottom-up system as a subroutine during its execution, in order to improve the quality of newly introduced speaker models. Experimental results show a relative improvement on three different datasets including meetings and TV-shows and gives up to 32% relative improvement in diarization performance.

  This work was presented at the Annual Conference of the International Speech Communication Association (Interspeech) in 2010 [Bozonnet et al., 2010].

(iv) Phoneme normalization for speaker diarization

The last contribution of this thesis relates to a new technology able to limit the influence of linguistic effects, analyzed in our comparative study as a drawback which may bias the convergence of the diarization system. By comparison to Speaker Adaptive Training (SAT), we propose an analogous way to reduce the linguistic components in the acoustic features. Our approach is referred to as Phone Adaptive Training (PAT). This technique is based on Constraint Maximum Likelihood Linear Regression (CMLLR) which aims to suppress the unwanted components through a linear feature transformation. Experimental results show an improvement of 11% relative improvement in diarization performance.

1.4 Organization

This thesis is organized in 8 chapters as follows:

In Chapter 2 a full survey is given to assess the state-of-the-art and progress in the field including the main approaches, their specificities and the ongoing problems.

Chapter 3 introduces the official metric, datasets and protocols as defined by NIST in order to then describe two state-of-the-art baseline systems: a bottom-up and a top-down approach and their respective performance.

Chapter 4 presents an Oracle study, which, thanks to 'blame game' experiments, aims to evaluate the sensitivity and the robustness of the different components of the top-down and bottom-up baseline systems and compare their weaknesses.

In Chapter 5 a new purification component is proposed for the baseline systems. After a description of the algorithm, purification is integrated into the top-down system and then the bottom-up system and an analysis of the performance is reported.

A comparative study of the top-down and bottom approach is detailed in Chapter 6, including first a formal definition of the task and the challenge of speaker diarization. Then a qualitative and experimental comparison is carried out, showing the differences of behavior of the two systems toward unwanted variation like the lexical content.

Chapter 7 introduces a system combination which takes the benefit of the difference of behaviors highlighted in Chapter 6 in order to design a more efficient system. Two scenarios are considered and their respective performances are examined.

Finally Chapter 8 introduces a new way to normalize the feature space, called Phone Adaptive Training (PAT), in order to attenuate the lexical effect considered as the main unwanted phone variation in Chapter 6. A description of the technique is first given, followed by some experimental results.

Conclusions are given in Chapter 9 summarizing the major contributions and results obtained in this thesis and points to some potential avenues for improvement and future work.

Chapter 2

State of The Art

Speaker diarization, commonly referred to as the 'who spoke when?' task, involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering) via unsupervised identification as illustrated in Figure 2.1.

Speaker diarization has been mainly applied on four domains namely telephone conversation, broadcast news and recorded lectures or meetings. In this chapter we review the main techniques used for the task of speaker diarization focusing on research over the recent years that relates predominantly to speaker diarization for conference meeting. Section 2.1 presents the main approaches used by the community. Section 2.2 details the possible different components used by these approaches and Section 2.3 introduces the hot topics and the current research directions in the field. Note that main part of this work was published in our article [Anguera et al., 2011].

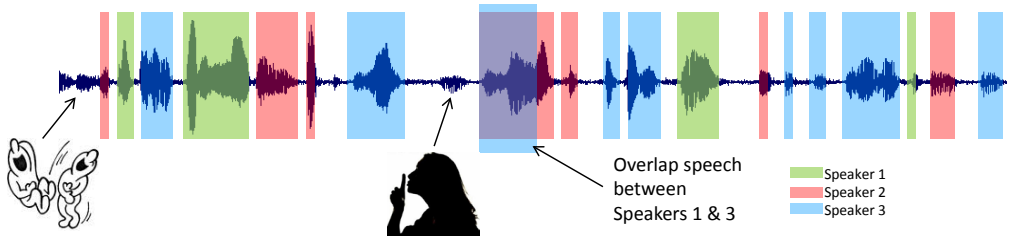


Figure 2.1: Example of audio diarization on recorded meeting including laughs, silence and 3 speakers.

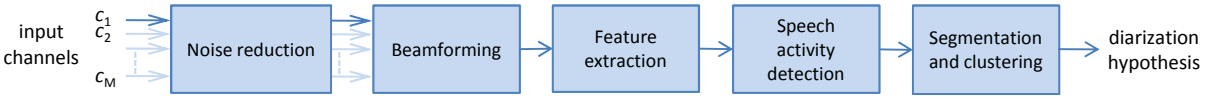
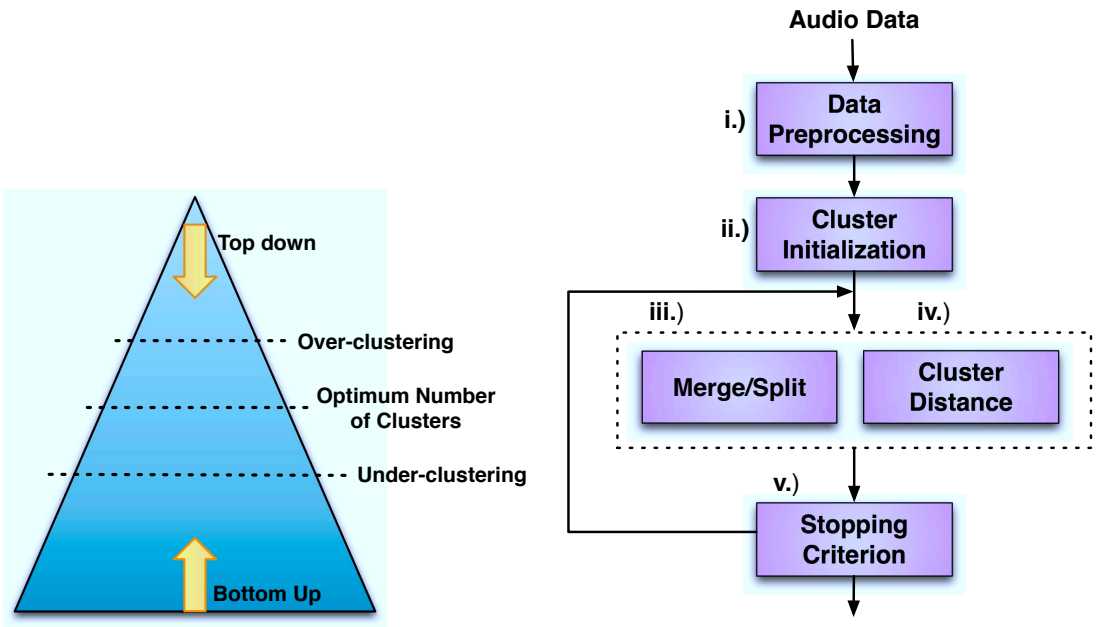


Figure 2.2: An overview of a typical speaker diarization system with one or multiple input channels.

2.1 Main Approaches

Current state-of-the-art systems to speaker diarization can be mainly categorized into two classes: they are bottom-up and top-down approaches. As illustrated in Figure 2.3(a), the top-down approach is first initialized with one (or very few) cluster and aims to iteratively split the clusters in order to reach an optimal number of clusters, ideally equal to the number of speakers. In contrast, the bottom-up approach is initialized with many clusters, in excess of the expected number of speakers, and then the clusters are merged iteratively until reaching the optimal amount of clusters. If the system provides more clusters than the real number of speakers, it is said to under-cluster, on the contrary, if the number of clusters is lower than the number of speakers, the system is said to over-cluster. Generally bottom-up and top-down systems are based on Hidden Markov Models (HMMs) where each state is associated with a Gaussian Mixture Model (GMM) and aims to characterize a single speaker. State transitions represent the speaker turns.

In this section, the standard bottom-up and top-down approaches are briefly outlined as well as two recent alternatives: one based on information theory and a second one based on a non-parametric Bayesian approach. Although these new approaches have not been reported previously in the context of official evaluations i.e. NIST RT evaluations, they have shown strong potential on official datasets and are thus included here. Some other works propose sequential single-pass segmentation and clustering approaches as well [Jothilakshmi et al., 2009; Kotti et al., 2008; Zhu et al., 2008], however their performance tends to fall short of the state-of-the-art, so they are not reported here.



(a) (b)

Figure 2.3: General diarization system: (a) Alternative clustering schemas, (b) General speaker diarization architecture. Pictures published with the kind permission of Xavier Anguera (Telefonica - Spain)

2.1.1 Bottom-Up Approach - Agglomerative Hierarchical Clustering

Bottom-up approach, so called agglomerative hierarchical clustering (AHC or AGHC) is the most popular in the literature. Its strategy aims to initialize the system in under-clustering the speech data in a number of clusters which exceeds the number of speakers. Then, successively, clusters are merged until only one cluster remains for each speaker. Different initializations have been proposed, including for example k-means clustering, however many systems finally kept a uniform initialization, where the speech stream is split into equal length abutted segments. Nonetheless this simpler approach leads to comparable performance[Anguera et al., 2006c]. In a second step, the bottom-up approach iteratively selects the two closest clusters and merges them. Generally a GMM model is trained on each cluster. Upon merging, a new GMM model is trained on the new merged cluster. To identify the closest clusters, standard distance metrics, as those described in Section 2.2.3 are used. After each cluster merging, the frames are reassigned to the clusters thanks to a Viterbi decoding for example. The whole scenario is repeated iteratively until some stopping criterion is reached, upon which it should ideally remain one cluster per speaker. Common stopping criterion include thresholded approaches such as the Bayesian Information Criterion (BIC) [Wooters & Huijbregts, 2008], Kullback-Leibler (KL)-based metrics [Rougui et al., 2006], the Generalized Likelihood Ratio (GLR) [Tsai et al., 2004] or the recently proposed T_s metric [Nguyen et al., 2008]. Bottom-up systems involved in the NIST RT evaluations [Nguyen et al., 2009; Wooters & Huijbregts, 2008] have performed consistently well.

2.1.2 Top-Down Approach - Divisive Hierarchical Clustering

In contrast with the previous approach, the top-down approach first models the entire audio stream with a single speaker model and successively adds new models to it until the full number of speakers are deemed to be accounted for. A single GMM model is trained on all the speech segments available, all of which are marked as unlabeled. Using some selection procedure to identify suitable training data from the non-labeled segments, new speaker models are iteratively added to the model 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. Top-down approaches are far less popular than their bottom-up counterparts. Some examples include [Fredouille et al., 2009; Fredouille & Evans, 2008; Meignier et al., 2001]. Whilst they are generally out-performed by the best bottom-up systems, top-down approaches have performed consistently and respectably well against the broader field of other bottom-up entries. Top-down approaches are also extremely computationally efficient and can be improved through cluster purification [Bozonnet et al., 2010].

2.1.3 Other Approaches

A recent alternative approach, though also bottom-up in nature, is inspired from rate-distortion theory and is based on an information-theoretic framework [Vijayasenan et al., 2007]. It 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 mutual information, which measures the mutual dependence of two variables [Vijayasenan et al., 2009]. Only a single global GMM is tuned for the full audio stream, and mutual information is computed in a new space of relevance variables defined by the GMM components. The approach aims at minimizing the loss of mutual information between successive clusterings while preserving as much information as possible from the original dataset. Two suitable methods have been reported: the agglomerative information bottleneck (aIB) [Vijayasenan et al., 2007] and the sequential information bottleneck (sIB) [Vijayasenan et al., 2009]. Even if this new system does not lead to better performance than parametric approaches, results comparable to state-of-the-art GMM systems are reported and are achieved with great savings in computation.

Alternatively, Bayesian machine learning became popular by the end of the 1990s and has recently been used for speaker diarization. The key component of Bayesian inference is that it does not aim at estimating the parameters of a system (*i.e.* to perform point estimates), but rather the parameters of their related distribution (hyperparameters). This allows for avoiding any premature hard decision in the diarization problem and for automatically regulating the system with the observations (*e.g* the complexity of the model is data dependent). However, the computation of posterior distributions often requires intractable integrals and, as a result, the statistics community has developed approximate inference methods. Monte Carlo Markov Chains (MCMC) were

first used [McEachern, 1994] to provide a systematic approach to the computation of distributions via sampling, enabling the deployment of Bayesian methods. However, sampling methods are generally slow and prohibitive when the amount of data is large, and they require to be run several times as the chains may get stuck and not converge in a practical number of iterations.

Another alternative approach, known as Variational Bayes, has been popular since 1993 [Hinton & van Camp, 1993; Wainwright & Jordan, 2003] and aims at providing a deterministic approximation of the distributions. It enables an inference problem to be converted to an optimization problem by approximating the intractable distribution with a tractable approximation obtained by minimizing the Kullback-Leibler divergence between them. In [Valente, 2005] a Variational Bayes-EM algorithm is used to learn a GMM speaker model and optimize a change detection process and the merging criterion. In [Reynolds et al., 2009] Variational Bayes 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 and differ from the nonparametric Bayesian systems introduced in Section 2.3.6.

Finally, the recently proposed speaker binary keys [Anguera & Bonastre, 2010] have been successfully applied to speaker diarization in meetings [Anguera & Bonastre, 2011] with similar performance to state-of-the-art systems but also with considerable computational savings (running in around 0.1 times real-time). Speaker binary keys are small binary vectors computed from the acoustic data using a UBM-like model. Once they are computed all processing tasks take place in the binary domain. Other works in speaker diarization concerned with speed include [Friedland et al., 2010; Huang et al., 2007] which achieve faster than real-time processing through the use of several processing tricks applied to a standard bottom-up approach ([Huang et al., 2007]) or by parallelizing most of the processing in a GPU unit ([Friedland et al., 2010]). The need for efficient diarization systems is emphasized when processing very large databases or when using diarization as a preprocessing step to other speech algorithms.

2.2 Main Algorithms

Figure 2.3(b) shows a block diagram of the generic modules which make up most speaker diarization systems. The data preprocessing step (Figure 2.3(b)-i) tends to be somewhat

domain specific. For meeting data, preprocessing usually involves noise reduction (such as Wiener filtering for example), multi-channel acoustic beamforming (see Section 2.2.1), the parameterization of speech data into acoustic features (such as MFCC, PLP, etc.) and the detection of speech segments with a speech activity detection algorithm (see Section 2.2.2). Cluster initialization (Figure 2.3(b)-ii) depends on the approach to diarization, *i.e.* the choice of an initial set of clusters in bottom-up clustering [Anguera et al., 2006a,c; Nguyen et al., 2009] (see Section 2.2.3) or a single segment in top-down clustering [Fredouille et al., 2009; Fredouille & Evans, 2008]. Next, in Figure 2.3(b)-iii/iv, a distance between clusters and a split/merging mechanism (see Section 2.2.4) is used to iteratively merge clusters [Ajmera, 2003; Nguyen et al., 2009] or to introduce new ones [Fredouille et al., 2009]. Optionally, data purification algorithms can be used to make clusters more discriminant [Anguera et al., 2006b; Bozonnet et al., 2010; Nguyen et al., 2009]. Finally, as illustrated in Figure 2.3(b)-v, stopping criteria are used to determine when the optimum number of clusters has been reached [Chen & Gopalakrishnan, 1998; Gish & Schmidt, 1994].

2.2.1 Acoustic beamforming

A specific characteristic of meeting recordings is the way they are recorded. Indeed meetings take place mainly in a room where often multiple microphones are located at different positions[Janin et al., 2004; McCowan et al., 2005; Mostefa et al., 2007]. Different types of microphone can be used including lapel microphones, desktop microphones positioned on the meeting room table, microphone arrays or wall-mounted microphones (intended for speaker localization). The availability of multiple channels captured by microphones of different natures and located at different location gives some potential for new speaker diarization approaches.

NIST introduced in the RT‘04 (Spring) evaluation the multiple distant microphone (MDM) condition. Since 2004, different systems handling multiple channels have been proposed. We can cite[Fredouille et al., 2004] who propose to perform speaker diarization on each channel independently and then to merge the individual outputs. To achieve the fusion of the outputs, the longest speaker intervention in each channel is selected to train a new speaker in the final segmentation output.

In the same year, [Jin et al., 2004] introduced a late-stage fusion approach where speaker segmentation is performed separately in all channels and diarization is applied

only taking into account the channel whose speech segments have the best signal-to-noise ratio (SNR).

Another approach aims to combine the acoustic signals from the different channels in order to make a single pseudo channel and perform a regular mono-channel diarization system. In [Istrate et al., 2005] for example, multiple channels are combined with a simple weighted sum according to their signal-to-noise (SNR) ratio. Though straightforward to implement, it does not take into account the time difference of arrival between each microphone channel and might easily lead to a decrease in performance.

Since the NIST RT'05 evaluation, the most common approach to multi-channel speaker diarization involves acoustic beamforming as initially proposed in [Anguera et al., 2005] and detailed in [Anguera et al., 2007]. Main of the RT participants use the free and open-source acoustic beamforming toolkit known as *BeamformIt* [Anguera, 2006] which consists of an enhanced delay-and-sum algorithm to correct misalignments due to the time-delay-of-arrival (TDOA) of speech to each microphone. Speech data can be optionally preprocessed using Wiener filtering [Wiener, 1949] to attenuate noise, for example, using [Adami et al., 2002a]. To perform the beamforming process, a reference channel is first 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. Various additional algorithms are available in the *BeamformIt* toolkit to select the optimum reference channel and to stabilize the TDOA values between channels before the signals are summed. Finally, the TDOA estimates themselves are made available as outputs and have been used successfully to improve diarization, as explained in Section 2.3.1.

Note that other algorithms can provide better beamforming results for some cases, however, delay-and-sum beamforming is the most reliable one when no a priori information on the location or nature of each microphone is known. Alternative beamforming algorithms include maximum likelihood (ML) [Seltzer et al., 2004] or generalized sidelobe canceler (GSC) [Griffiths & Jim, 1982] which adaptively find the optimum parameters, and minimum variance distortionless response (MVDR) [Woelfel & McDonough, 2009] when prior information on ambient noise is available. All of these have higher computational requirements and, in the case of the adaptive algorithms, there is the risk to

converge to inaccurate parameters, especially when processing microphones of different nature.

2.2.2 Speech Activity Detection

Speech Activity Detection (SAD) involves the labeling of speech and non-speech segments. SAD can have a significant impact on speaker diarization performance for two reasons. The first stems directly from the standard speaker diarization performance metric, namely the diarization error rate (DER), which takes into account both the false alarm and missed speaker error rates (see Section 3.2 for more details on evaluation metrics); poor SAD performance will therefore lead to an increased DER. The second follows from the fact that non-speech segments can disturb the speaker diarization process, and more specifically the acoustic models involved in the process [Wooters et al., 2004]. Indeed, the inclusion of non-speech segments in speaker modeling leads to less discriminant models and thus increased difficulties in segmentation. Consequently, a good compromise between missed and false alarm speech error rates has to be found to enhance the quality of the following speaker diarization process.

SAD is a fundamental task in almost all fields of speech processing (coding, enhancement, and recognition) and many different approaches and studies have been reported in the literature [Ramirez et al., 2007]. Initial approaches for diarization tried to solve speech activity detection on the fly, *i.e.* by having a non-speech cluster be a by-product of the diarization. However, it became evident that better results are obtained using a dedicated speech/non-speech detector as pre-processing step. In the context of meetings non-speech segments may include silence, but also ambient noise such as paper shuffling, door knocks or non-lexical noise such as breathing, coughing and laughing, among other background noises. Therefore, highly variable energy levels can be observed in the non-speech parts of the signal. Moreover, differences in microphones or room configurations may result in variable signal-to-noise ratios (SNRs) from one meeting to another. Thus SAD is far from being trivial in this context and typical techniques based on feature extraction (energy, spectrum divergence between speech and background noise, and pitch estimation) combined with a threshold-based decision have proved to be relatively ineffective.

Model-based approaches tend to have better performances and rely on a two-class detector, with models pre-trained with external speech and non-speech

data [Anguera et al., 2005; Fredouille & Senay, 2006; Van Leeuwen & Konečný, 2008; Wooters et al., 2004; Zhu et al., 2008]. Speech and non-speech models may optionally be adapted to specific meeting conditions [Fredouille & Evans, 2008]. Discriminant classifiers such as Linear Discriminant Analysis (LDA) coupled with Mel Frequency Cepstrum Coefficients (MFCC) [Rentzeperis et al., 2006] or Support Vector Machines (SVM) [Temko et al., 2007] have also been proposed in the literature. The main drawback of model-based approaches is their reliance on external data for the training of speech and non-speech models which makes them less robust to changes in acoustic conditions. Hybrid approaches have been proposed as a potential solution. In most cases, an energy-based detection is first applied in order to label a limited amount of speech and non-speech data for which there is high confidence in the classification. In a second step, the labeled data are used to train meeting-specific speech and non-speech models, which are subsequently used in a model-based detector to obtain the final speech/non-speech segmentation [Anguera et al., 2006; Nwe et al., 2009; Sun et al., 2009; Wooters & Huijbregts, 2008]. Finally, [El-Khoury et al., 2009] combines a model-based with a 4Hz modulation energy-based detector. Interestingly, instead of being applied as a preprocessing stage, in this system SAD is incorporated into the speaker diarization process.

2.2.3 Segmentation

In the literature, the term 'speaker segmentation' is sometimes used to refer to both segmentation and clustering. Whilst some systems treat each task separately many of present state-of-the-art systems tackle them simultaneously, as described in Section 2.2.5. In these cases the notion of strictly independent segmentation and clustering modules is less relevant. However, both modules are fundamental to the task of speaker diarization and some systems, such as that reported in [Zhu et al., 2008], apply distinctly independent segmentation and clustering stages. Thus the segmentation and clustering models are described separately here.

Speaker segmentation is core to the diarization process and aims at splitting the audio stream into speaker homogeneous segments or, alternatively, to detect changes in speakers, also known as speaker turns. The classical approach to segmentation performs a hypothesis testing using the acoustic segments in two sliding and possibly overlapping,

consecutive windows. For each considered change point there are two possible hypotheses: first that both segments come from the same speaker (H_0), and thus that they can be well represented by a single model; and second that there are two different speakers (H_1), and thus that two different models are more appropriate. In practice, models are estimated from each of the speech windows and some criteria are used to determine whether they are best accounted for by two separate models (and hence two separate speakers), or by a single model (and hence the same speaker) by using an empirically determined or dynamically adapted threshold [Lu et al., 2002; Rougui et al., 2006]. This is performed across the whole audio stream and a sequence of speaker turns is extracted.

Many different distance metrics have appeared in the literature. Next we review the dominant approaches which have been used for the NIST RT speaker diarization evaluations during the last 4 years. The most common approach is that of the Bayesian Information Criterion (BIC) and its associated ΔBIC metric [Chen & Gopalakrishnan, 1998] which has proved to be extremely popular *e.g.* [Ben et al., 2004; Li & Schultz, 2009; van Leeuwen & Huijbregts, 2007]. The approach requires the setting of an explicit penalty term which controls the trade-off between missed turns and those falsely detected. It is generally difficult to estimate the penalty term such that it gives stable performance across different meetings and thus new, more robust approaches have been devised. They either adapt the penalty term automatically, *i.e.* the modified BIC criterion [Chen & Gopalakrishnan, 1998; Mori & Nakagawa, 2001; Vandecatseye et al., 2004], or avoid the use of a penalty term altogether by controlling model complexity [Ajmera et al., 2004]. BIC-based approaches are computationally demanding and some systems have been developed in order to use the BIC only in a second pass, while a statistical-based distance is used in a first pass [Lu & Zhang, 2002]. Another BIC-variant metric, referred to as cross-BIC and introduced in [Anguera & Hernando, 2004; Anguera et al., 2005], involves the computation of cross-likelihood: the likelihood of a first segment according to a model tuned from the second segment and vice versa. In [Malegaonkar et al., 2006], different techniques for likelihood normalization are presented and are referred to as bilateral scoring.

A popular and alternative approach to BIC-based measures is the Generalized Likelihood Ratio (GLR), *e.g.* [Delacourt & Wellekens, 2000; Siu et al., 1991]. In contrast to the BIC, the GLR is a likelihood-based metric and corresponds to the ratio between the two aforementioned hypotheses, as described in [Gangadharaiah et al., 2004; Jin et al.,

2004; Shrikanth & Narayanan, 2008]. To adapt the criterion in order to take into account the amount of training data available in the two segments, a penalized GLR was proposed in [Liu & Kubala, 1999].

The last of the dominant approaches is the Kullback-Leibler (KL) divergence which estimates the distance between two distributions [Siegler et al., 1997]. However, the KL divergence is asymmetric, and thus the KL2 metric, a symmetric alternative, has proved to be more popular in speaker diarization when used to characterize the similarity of two audio segments [Siegler et al., 1997; Zhu et al., 2006; Zochová & Radová, 2005].

Finally, in this section we include a newly introduced distance metric that has shown promise in a speaker diarization task. The Information Change Rate (ICR), or entropy can be used to characterize the similarity of two neighbouring speech segments. The ICR determines the change in information that would be obtained by merging any two speech segments under consideration and can thus be used for speaker segmentation. Unlike the measures outlined above, the ICR similarity is not based on a model of each segment but, instead, on the distance between segments in a space of relevance variables, with maximum mutual information or minimum entropy. One suitable space comes from GMM component parameters [Vijayasenan et al., 2007]. The ICR approach is computationally efficient and, in [Han & Narayanan, 2008], ICR is shown to be more robust to data source variation than a BIC-based distance.

2.2.4 Clustering

Whereas the segmentation step operates on adjacent windows in order to determine whether or not they correspond to the same speaker, clustering aims at identifying and grouping together same-speaker segments which can be localized anywhere in the audio stream. Ideally, there will be one cluster for each speaker. The problem of measuring segment similarity remains the same and all the distance metrics described in Section 2.2.3 may also be used for clustering, *i.e.* the KL distance as in [Rougui et al., 2006], a modified KL2 metric as in [Ben et al., 2004], a BIC measure as in [Moraru et al., 2005] or the cross likelihood ratio (CLR) as in [Aronowitz, 2007; Barras et al., 2004].

However, with such an approach to diarization, there is no provision for splitting segments which contain more than a single speaker, and thus diarization algorithms can only work well if the initial segmentation is of sufficiently high quality. Since this is rarely the case, alternative approaches combine clustering with iterative resegmentation,

hence facilitating the introduction of missing speaker turns. Most present diarization systems thus perform segmentation and clustering simultaneously or clustering on a frame-to-cluster basis, as described in Section 2.2.5. The general approach involves Viterbi realignment where 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. In order to make the Viterbi decoding more stable, it is common to use a Viterbi buffer to smooth the state, cluster or speaker sequence to remove erroneously detected, brief speaker turns, as in [Fredouille et al., 2009]. Most state-of-the-art systems employ some variations on this particular issue.

An alternative approach to clustering involves majority voting [Friedland & Vinyals, 2008; Hung & Friedland, 2008] whereby short windows of frames are entirely assigned to the closest cluster, *i.e.* that which attracts the most frames during decoding. This technique leads to savings in computation but is more suited to online or live speaker diarization systems.

2.2.5 One-Step Segmentation and Clustering

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. Such a method was initially proposed in [Ajmera, 2003] for a bottom-up system and has subsequently been adopted by many others [Anguera et al., 2005; Friedland et al., 2009; Luque et al., 2008; Pardo et al., 2006a; Van Leeuwen & Konečný, 2008; Wooters & Huijbregts, 2008]. For top-down algorithms it was initially proposed in [Meignier et al., 2001] as used in their latest system [Fredouille et al., 2009].

In all cases the different acoustic classes are represented using HMM/GMM models. EM training or MAP adaptation is used to obtain the closest possible models given the current frame-to-model assignments, 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. This step is useful when a class is created/eliminated so that the resulting class distribution is allowed to adapt to the data.

The one-step segmentation and clustering approach, although much slower, constitutes a clear advantage versus sequential single-pass segmentation and clustering approaches [Jothilakshmi et al., 2009; Kotti et al., 2008; Zhu et al., 2008]. On the one

hand, early errors (mostly missed speaker turns from the segmentation step) can be later corrected by the re-segmentation steps. On the other hand, most speaker segmentation algorithms use only local information to decide on a speaker change while when using speaker models and Viterbi realignment all data is taken into consideration.

When performing frame assignment using Viterbi algorithm a minimum assignment duration is usually enforced to avoid an unrealistic assignment of very small consecutive segments to different speaker models. Such minimum duration is usually made according to the estimated minimum length of any given speaker turn.

2.2.6 Purification of Output Clusters

The segmentation and clustering steps follow a greedy strategy i.e. they take decisions on the basis of information at hand without worrying about the effect these decisions may have in the future. Final outputs may result in a speaker segmentation that is not optimal and correspond to a local minimum. It is then possible to apply a post processing step in order to refine the clustering outputs. Cluster purification aims to first select the best frames for each cluster and retake a decision for all the other speech data considered as less confident.

In [Anguera et al., 2006b] a purification component for a bottom-up diarization system is proposed. It involves in selecting first the best speech segment in each cluster according to its likelihood. Then a ΔBIC score is computed between the best segment and all other segments in the same cluster. According to a threshold, either the cluster is declared to be pure else it is split into two clusters, then all models are retrained and the data are realigned.

In[Ning et al., 2006] proposed a post processing for a agglomerative Hierarchical clustering called 'cross EM refinement'. This algorithm based on the idea of cross validation and EM algorithm aims to avoid some possible over-fitting and split randomly and equally each cluster into two parts. Then the first part is used to retrain the cluster model and labels are update on the second part. Then the role of each part is reversed.

2.3 Current Research Directions

In this section we review those areas of work which are still not mature and which have the potential to improve diarization performance. We first discuss the trend in recent

NIST RT evaluations to use spatial information obtained from multiple microphones, which are used by many in combination with MFCCs to improve performance. Then, we discuss the use of prosodic information which has led to promising speaker diarization results. Also addressed in this section is the 'Achilles heel' of speaker diarization for meetings, which involves overlapping speech; many researchers have started to tackle the detection of overlapping speech and its correct labeling for improved diarization outputs. We then consider a recent trend towards multimodal speaker diarization including studies of multimodal, audiovisual techniques which have been successfully used for speaker diarization, at least for laboratory conditions. Finally we consider general combination strategies that can be used to combine the output of different diarization systems. The following summarizes recent work in all of these areas.

2.3.1 Time-Delay Features

Estimates of inter-channel delay may be used not only for delay-and-sum beamforming of multiple microphone channels, as described in Section 2.2.1, but also for speaker localization. If we assume that speakers do not move, or that appropriate tracking algorithms are used, then estimates of speaker location may thus be used as additional features, which have nowadays become extremely popular. Much of the early work, *e.g.* [Lathoud & Cowan, 2003], requires explicit knowledge of microphone placement. However, as is the case with NIST evaluations, such a priori information is not always available. The first work [Ellis & Liu, 2004] that does not rely on microphone locations led to promising results, even if error rates were considerably higher than that achieved with acoustic features. Early efforts to combine acoustic features and estimates of inter-channel delay clearly demonstrated their potential, *e.g.* [Ajmera et al., 2004], though this work again relied upon known microphone locations.

More recent work, and specifically in the context of NIST evaluations, reports the successful combination of acoustic and inter-channel delay features [Pardo et al., 2006a, 2007, 2006b] when they are combined at the weighted log-likelihood level, though optimum weights were found to vary across meetings. Better results are reported in [Anguera et al., 2007] where automatic weighting based on an entropy-based metric is used for cluster comparison in a bottom-up speaker diarization system. A complete front-end for speaker diarization with multiple microphones was proposed in [Anguera et al., 2007]. Here a two-step TDOA Viterbi post-processing algorithm together with a dynamic

output signal weighting algorithm were shown to greatly improve speaker diarization accuracy and the robustness of inter-channel delay estimates to noise and reverberation, which commonly afflict source localization algorithms. More recently an approach to the unsupervised discriminant analysis of inter-channel delay features was proposed in [Evans et al., 2009] and results of approximately 20% DER were reported using delay features alone.

In the most recent NIST RT evaluation, in 2009, all but one entry used estimates of inter-channel delay both for beamforming and as features. Since comparative experiments are rarely reported it is not possible to assess the contribution of delay features to diarization performance. However, those who do use delay features report significant improvements in diarization performance and the success of these systems in NIST RT evaluations would seem to support their use.

2.3.2 Use of Prosodic Features in Diarization

The use of prosodic features for both speaker detection and diarization is emerging as a reaction to the theoretical inconsistency derived from using MFCC features both for speaker recognition (which requires invariance against words) and speech recognition (which requires invariance against speakers) [Wölfel et al., 2009]. In [Friedland et al., 2009] the authors present a systematic investigation of the speaker discriminability of 70 long-term features, most of them prosodic features. They provide evidence that despite the dominance of short-term cepstral features in speaker recognition, a number of long-term features can provide significant information for speaker discrimination. As already suggested in [Shriberg, 2007], the consideration of patterns derived from larger segments of speech can reveal individual characteristics of the speakers' voices as well as their speaking behavior, information which cannot be captured using a short-term, frame-based cepstral analysis. The authors use Fisher LDA as a ranking methodology and sort the 70 prosodic and long-term features by speaker discriminability. The combination of the top-ten ranked prosodic and long-term features combined with regular MFCCs leads to a 30% relative improvement in terms of DER compared to the top-performing system of the NIST RT evaluation in 2007. An extension of the work is provided in [Imseng & Friedland, 2010]. The article presents a novel, adaptive initialization scheme that can be applied to standard bottom-up diarization algorithms. The initialization method is a combination of the recently proposed 'adaptive seconds per

Gaussian' (ASPG) method [Imseng & Friedland, 2009] and a new pre-clustering method in addition to a new strategy which automatically estimates an appropriate number of initial clusters based on prosodic features. It outperforms previous cluster initialization algorithms by up to 67% (relative).

2.3.3 Overlap Detection

The process of overlapping speech in speaker diarization is a problem which remains largely unsolved. Indeed, the main part of the current speaker diarization systems permit only to assign one speaker to each segment, while overlapping speech is very common in domains like multi-party meetings. Consequences on the overall DER are high missed speech errors when overlapped speech is omitted and can be a substantial fraction of the DER. Moreover without some means of detection, segments of overlapping speech lead to impurities in speaker specific models and hence reduce segmentation performance. Approaches to overlap detection were thoroughly assessed in [Çetin & Shriberg, 2006; Shriberg et al., 2001] and, even whilst applied to ASR as opposed to speaker diarization, only a small number of systems actually detects overlapping speech well enough to improve error rates [Boakye, 2008; Boakye et al., 2008; Trueba-Hornero, 2008].

In [Otterson & Ostendorf, 2007] the authors demonstrated a theoretical improvement in diarization performance by adding a second speaker during overlap regions using a simple strategy of assigning speaker labels according to the labels of the neighboring segments, as well as by excluding overlap regions from the input to the diarization system. However, this initial study assumed ground-truth overlap detection. In [Trueba-Hornero, 2008] a real overlap detection system was developed, as well as a better heuristic that computed posterior probabilities from diarization to post process the output and include a second speaker on overlap regions. The main bottleneck of the achieved performance gain is mainly due to errors in overlap detection, and more work on enhancing its precision and recall is reported in [Boakye, 2008; Boakye et al., 2008]. The main approach consists of a three state HMM-GMM system (non-speech, non-overlapped speech, and overlapped speech), and the best feature combination is MFCC and modulation spectrogram features [Kingsbury et al., 1998], although comparable results were achieved with other features such as root mean squared energy, spectral flatness, or harmonic energy ratio. The reported performance of the overlap detection is 82% precision and 21% recall, and yielded a relative improvement of 11% DER. However, assuming reference

overlap detection, the relative DER improvement goes up to 37%. This way, this area has potential for future research efforts.

2.3.4 Audiovisual Diarization

An empirical study to review definitions of audiovisual synchrony and examine their empirical behavior is presented in [Nock et al., 2003]. The results provide justifications for the application of audiovisual synchrony techniques to the problem of active speaker localization in broadcast video. Zhang et al. [2006] present a multi-modal speaker localization method using a specialized satellite microphone and an omni-directional camera. Though the results seem comparable to the state-of-the-art, the solution requires specialized hardware. The work presented in [Noulas & Krose, 2007] integrates audiovisual features for on-line audiovisual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to discussions with two to three people on two short test scenarios. Another use of DBN, also called factorial HMMs [Ghahramani & Jordan, 1997], is proposed in [Noulas et al., 2009] 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 [Tamura et al., 2004] the authors demonstrate that the different shapes the mouth can take when speaking facilitate word recognition under tightly constrained test conditions (*e.g.* frontal position of the subject with respect to the camera while reading digits).

Common approaches to audiovisual speaker identification involve identifying lip motion from frontal faces, *e.g.* [Chen & Rao, 1996; Fisher & Darrell, 2004; Fisher et al., 2000; Rao & Chen, 1996; Siracusa & Fisher, 2007]. Therefore, the underlying assumption is that motion from a person comes predominantly from the motion of the lower half of their face. In addition, gestural or other non-verbal behaviors associated with natural body motion during conversations are artificially suppressed, *e.g.* for the CUAVE database [Patterson et al., 2002]. Most of the techniques involve the identification of one or two people in a single video camera only where short term synchrony of lip motion and speech are the basis for audiovisual localization. In a real scenario the subject behavior is not controlled and, consequently, the correct detection of the mouth is not always feasible. Therefore, other forms of body behavior, *e.g.* head gestures, which are also visible manifestations of speech [McNeill, 2000] are used. While there has been relatively

little work on using global body movements for inferring speaking status, some studies have been carried out [Campbell & Suzuki, 2006; Hung & Friedland, 2008; Hung et al., 2008; Vajaria et al., 2006] that show promising initial results.

However, until the work presented in [Friedland et al., 2009], approaches have never considered audiovisual diarization as a single, unsupervised joint optimization problem. The work in [Friedland et al., 2009], though, relies on multiple cameras. The first article that discusses joint audiovisual diarization using only a single, low-resolution overview camera and also tests on meeting scenarios where the participants are able to move around freely in the room is [Friedland et al., 2009]. 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". This solution to the localization problem has properties that may not be observed either by audio-only diarization nor by video-only localization, such as increased robustness against various issues present in the channel. In addition, in contrast to audio-only speaker diarization, this solution provides a means for identifying speakers beyond clustering numbers by associating video regions with the clusters.

2.3.5 System Combination

System or component combination is often reported in the literature as an effective means for improving performance in many speech processing applications. However, very few studies related to speaker diarization have been reported in recent years. This could be due to the inherent difficulty of merging multiple output segmentations. Combination strategies, due to the unsupervised nature of the diarization task, have to accommodate differences in temporal synchronization, outputs with different number of speakers, and the matching of speaker labels. Moreover, systems involved in the combination have to exhibit segmentation outputs that are sufficiently orthogonal in order to ensure significant gains in performance when combined. 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] for instance, the authors combine two different algorithms based on the Information Bottleneck framework. In [El-Khoury et al., 2008], 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 (parametrizations based on Gaussianized against non-Gaussianized MFCCs) are combined for the speaker diarization of phone calls conversations. The combination approach relies on both systems identifying some common clusters which are then considered as the most relevant. All the segments not belonging to these common clusters are labeled as misclassified and are involved in a new re-classification step based on a GMM modeling of the common clusters and a maximum likelihood-based decision.

2.3.6 Alternative Models

Among the clustering structures recently developed some differ from the standard HMM insofar as they are fully nonparametric (that is, the number of parameters of the system depends on the observations). The Dirichlet process (DP) [Ferguson, 1973] allows for converting the systems into Bayesian and nonparametric systems. The DP mixture model produces infinite Gaussian mixtures and defines the number of components by a measure over distributions. The authors of [Valente, 2006] illustrate the use of the Dirichlet process mixtures, showing an improvement compared to other classical methods. [Teh et al., 2006] propose another nonparametric Bayesian approach, in which a stochastic hierarchical Dirichlet process (HDP) defines a prior distribution on transition matrices over countably infinite state spaces, that is, no fixed number of speakers is assumed, nor found through either split or merging approaches using classical model selection approaches (such as the BIC criterion). Instead, this prior measure is placed over distributions (called a random measure), which is integrated out using likelihood-prior conjugacy. The resulting HDP-HMM leads to a data-driven learning algorithm which infers posterior distributions over the number of states. This posterior uncertainty can be integrated out when making predictions effectively averaging over models of varying complexity. The HDP-HMM has shown promise in diarization [Fox et al., 2008], yielding similar performance to the standard agglomerative HMM with GMM emissions, while

requiring very little hyper-parameter tuning and providing a statistically sound model. Globally, these non parametric Bayesian approaches did not bring a major improvement compared to classical systems as presented in Section 2.2. However, they may be promising insofar as they do not necessarily need to be optimized for certain data compared to methods cited in Section 2.1. Furthermore, they provide a probabilistic interpretation on posterior distributions (*e.g.* number of speakers).

Chapter 3

Protocols & Baseline Systems

Much progress has been made in speaker diarization over recent years partly spearheaded by the National Institute of Standards and Technology (NIST) Rich Transcription (RT) evaluations [NIST, 2002, 2003, 2004, 2006, 2007, 2009] in the proceedings of which are found two general approaches: top-down or divise hierarchical clustering (DHC) and bottom-up or agglomerative hierarchical clustering (AHC). Even though the best performing systems over recent years have all been bottom-up approaches we believe that the top-down approach is not without significant merit. Results on the NIST RT'09 dataset show that the top-down approach gives extremely competitive results[1] and is significantly less computationally demanding than bottom-up approaches.

In this chapter we first describe the official protocols and metric proposed by NIST and then introduce the different datasets used in the Rich Transcription evaluations. A TV talk-shows dataset used later to assess the robustness of the baselines is also introduced. Then details of the bottom-up and top-down hierarchical clustering considered as our baselines are presented. Finally experimental results for the different baseline systems are given.

3.1 Protocols

Since 2004, NIST has organized a series of benchmark evaluations within the Rich Transcription (RT) campaigns[2]. These evaluations which include the task of speaker di-

[1] on the multiple distant microphone (MDM) condition (even though we did not use estimates of inter-channel delay as features) and on the single distant microphone (SDM) condition

[2] See http://nist.gov/speech/tests/rt.

arization, aim to facilitate transcription and annotation technology for human-to-human speech. Due to its international scope, the RT evaluations have had an instrumental role in assessing the state-of-the-art and in providing standard evaluation protocols, performance metrics and common datasets. An important characteristic of these evaluations is that there is no a priori information available to the participants (e.g. number of speakers, speaker identities, etc.) with the exception of the nature of the recording (e.g. conference meetings, broadcast news, etc.) and the language (English). Standard formats for data input and output are defined and evaluation participants may use external data for building world models and/or for normalization purposes.

Having considered broadcast news, lectures or coffee breaks domain, the most recent RT evaluation focused on conference meetings, a particularly challenging domain for speaker diarization due to its spontaneous speaking style. For this reason the work presented in this thesis also targets the meeting domain. The meetings provided in the RT evaluations were recorded using multiple microphones of different types and qualities which are positioned on the participants (e.g. lapel microphone) or in different locations around the meeting room. By grouping these microphones into different classes, NIST proposed several contrastive evaluation conditions. These include: individual headphone microphones (IHM), single distant microphones (SDM), multiple distant microphones (MDM), multiple mark III arrays (MM3A[1]) and all distant microphones (ADM).

The MDM condition is defined as the core, required condition, where participants have the possibility to use data recorded simultaneously from a number of distributed table-top microphones. Standard practice in this case involves acoustic beamforming [Anguera, 2006] in order to obtain a single pseudo channel and may utilize localization or inter-channel delay (ICD) features [Anguera et al., 2005; Ellis & Liu, 2004; Evans et al., 2009] which, if integrated with traditional acoustic features, can lead to better diarization performance [Anguera et al., 2005].

In contrast, the SDM condition allows only the use of data recorded from one microphone (usually the most centrally located) and cannot therefore exploit speech enhancement with beamforming of multiple channels or the use of ICD. In this thesis we mainly show results for SDM condition since we consider them to be the most representative of standard meeting room recording equipment.

[1] MM3A microphones are those exclusively found within the arrays built and provided by NIST. These are usually not included within the MDM condition, they are included within the ADM condition.

3.2 Metrics

NIST defines a standard diarization output which contains a hypothesized speaker activity including starting and stopping times of speech segments. Speaker labels are used solely to identify the multiple interventions of a given speaker, but do not reflect their real identity. In order to estimate the quality of the hypothesis, the outputs are compared to the ground-truth reference in order to obtain the overall Diarization Error Rate (DER) also defined by NIST. The DER metric can be defined as the time-weighted sum of three sources of error:

- **Missed Speech** (MS): percentage of speech in the ground-truth which is not in the hypothesis;
- **False Alarm speech** (FA): percentage of speech in the hypothesis which is not in the ground-truth;
- **Speaker Error** (*SpkErr*): percentage of speech assigned to the wrong speaker (while ignoring the overlapped speech)

The DER can be determined with and without the inclusion of overlapping speech segments. When scoring the segments of overlapping speech, the DER reflects errors in the estimated number of simultaneous speakers (in the NIST RT evaluations up to 4 overlapping speakers are considered in the scoring) and errors in the speaker label. Errors on the estimated number of speakers lead to an increase of the MS when fewer speakers than the real number are hypothesized or the FA when too many speakers are hypothesized. In case of errors on the speaker label, the respective speaker error of each of the overlap speaker is included in the *SpkErr*.

The DER is determined according to Equation 3.1

$$DER = SAD_{error} + SpkErr = \underbrace{MS + FA}_{\text{SAD Error}} + SpkErr \tag{3.1}$$

More precisely, the DER is computed as the fraction of speaker time that is not correctly attributed, based on an optimal mapping. The mapping is performed according to a standard dynamic programming algorithm defined by NIST, between speakers in

the ground-truth and those in the speaker diarization hypothesis. The DER can be formally defined as:

$$DER = \frac{\sum_{\forall i}\{D_i^R \cdot (\max(N_i^R, N_i^S) - N_i^C)\}}{\sum_{\forall i}\{D_i^R \cdot N_i^R\}} \tag{3.2}$$

where D_i^R denotes the duration of the i-th reference segment, and where N_i^R and N_i^S are respectively the number of speakers according to the reference and the number of speakers in the diarization hypothesis. N_i^C is the number of speakers that are correctly matched by the diarization system. Note that with overlapping speech, N_i^R,N_i^S and N_i^C can be larger than one.

As can be seen from Equation 3.2 the DER is time-weighted, i.e. it attributes less importance to speakers whose overall speaking time is small. Additionally, a non-scoring collar of 250ms is generally applied either side of the ground-truth segment boundaries to account for inevitable inconsistencies in precise start and end point labeling. For the TV shows with one dominant speaker and multiple relatively inactive speakers (typical examples can be found in the 'Grand Échiquier' corpus, see 3.3.2), the DER is not always a relevant metric, since it can be very small even if only a single speaker is detected.

Note that, since 2006, the primary metric of the RT evaluations includes the overlapping speech error. However since the systems reported in this thesis assume only a single speaker at a time and do not detect or handle overlapped speech, we refer often to the metric without scoring overlapped speech. In this case N_i^R,N_i^S and N_i^C are either zero or one. Where possible we nonetheless report both scores: with and without the scoring of overlap.

3.3 Datasets

In the work outlined in this manuscript, the majority of the experiments are performed on meeting domain, i.e. involving the NIST RT meeting corpus. However, in order to assess the robustness of the systems to different data, some additional work involving a corpus of TV-talk shows, known as the Grand Échiquier dataset, is also described in Section 3.3.2.

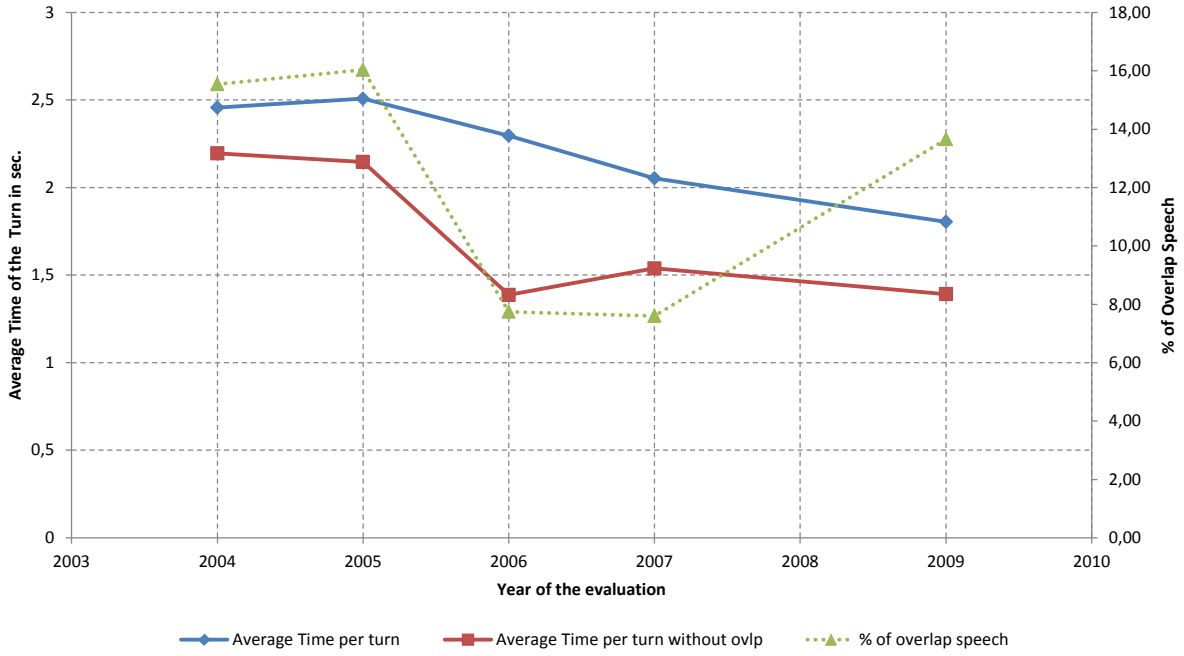


Figure 3.1: Analysis of the percentage of overlap speech and the average duration of the turns for each of the 5 NIST RT evaluation datasets. Percentages of overlap speech are given over the total speech time

3.3.1 RT Meeting Corpus

For each NIST RT evaluation since 2004 a new database of annotated audio meetings was collected[1]. A total of five conference meeting evaluation datasets is available.

Figure 3.1 shows the difference between RT evaluation datasets in terms of percentage of overlap speech and turn duration. For RT'04, RT'05 and RT'09 we see a percentage of overlap speech in the order of 15%, while the datasets from 2006 and 2007 involve around 8% of overlap speech. While looking at the average turn duration, which can be defined as the average time during which there is no change in speaker activity (same speaker, same condition: overlap/no overlap), we observe that the last three evaluations: RT'06, '07 and '09 have shorter average turn durations, although we do not consider overlap speech. This brings strikingly to the fore the fact that the speech present in the three last evaluations may be considered as more spontaneous and more interactive, leading to smaller turn durations. According to these first observations we therefore expect the

[1] The ground-truth keys are released later so that they may be used by the community for their own research and development independently of official NIST evaluations

RT‘06,‘07 and ‘09 datasets to be more challenging.

For the work reported in this thesis, and for consistency with previous work [Fredouille & Evans, 2008; Fredouille et al., 2004], all the experimental systems were optimized on a development dataset of 23 meetings from the NIST RT‘04, ‘05 and ‘06 evaluations. Performance was then assessed on the independent RT‘07 and RT‘09 datasets. Note that there is no overlap between development and evaluation datasets although they may contain shows recorded from the same site and possibly identical speakers.

3.3.2 GE TV-Talk Shows Corpus

Through some other work [Bozonnet et al., 2010] we also conducted speaker diarization assessments on a database of TV talk-shows known as the Grand Échiquier’ (GE) database. Since these results allow us to evaluate the robustness of speaker diarization system (i.e. to variations in dominant speaker floor time), it is described here. Baseline results for the GE database are reported in Section 3.5.

This corpus is comprised of over 50 French-language TV talk-show programs from the 1970-80s and was made popular among both national and European multimedia research projects, e.g. the European K-Space network of excellence [K-Space, K-Space]. Each show focuses on a main guest and other supporting guests, who are both interviewed by a host presenter. The interviews are punctuated with film excerpts, live music, audience applause and laughter. Aside from this, silences during speaker turns can be very short or almost negligible; compared to meetings, where speakers often pause to collect their thoughts or to reflect before responding to a question, TV show speech tends to be more fluent and sometimes almost scripted. This is perhaps due to the fact that the main themes and discussions are prepared in advance and known by the speakers.

Table 3.1 highlights more quantitative differences between NIST RT conference meetings from the RT‘09 dataset and 7 TV shows from the GE database, which have thus far been annotated manually according to standard NIST RT protocols [NIST, 2009]. Upon comparison of the first 3 lines of Table 3.1 we observe that TV-talk shows are on average much longer than conference meeting (147 minutes vs. 25 minutes) and, with noise (e.g. applause) and music removed, the quantity of speech is twice that for RT data (50 minutes vs. 21 minutes). Note, however, that the average segment duration is slightly smaller for RT‘09 than for GE (2 sec. vs 3 sec.). These preliminary findings

Attribute	GE	NIST RT'09
No. of shows	7	7
Avg. Evaluation time	147 min.	25 min.
Total speech	50 min.	21 min.
Avg. No. of segments	1033	882
Avg. segment length	3 sec.	2 sec.
Avg. Overlap	5 min.	3 min.
Avg. % Overlap / Total speech	10 %	14 %
Avg. No. speakers	13	5
most active	1476 sec.	535 sec.
least active	7 sec.	146 sec.

Table 3.1: A comparison of Grand Échiquier (GE) and NIST RT'09 database characteristics.

may suggest that TV-shows will present more of a challenge due to the greater levels of intra-speaker variability within a same show.

Moreover, differences in terms of speaker statistics have to be considered as well. Indeed the average number of speakers, and the average floor time for the most and least active speakers in each show are not comparable for both domains. On average there are 13 speakers per TV show but only 5 speakers per conference meeting. This might be expected given the longer average length of TV shows. Given a larger number of speakers we can expect a smaller average inter-speaker difference than for meetings and hence increased difficulties in speaker diarization.

Furthermore, we see that the spread in floor time is much greater for the GE dataset than it is for the RT'09 dataset. The average speaking time for the most active speaker is 1476 seconds for the GE dataset (cf. 535 sec. for RT'09) and corresponds to the host presenter in each case. The average speaking time for the least active speaker is only 7 seconds (cf. 146 sec. for RT'09) and corresponds to one of the minor supporting guests. Speakers with such little data are extremely difficult to detect and thus this aspect of the TV show dataset is likely to pose significant difficulties for speaker diarization. Note however that the overall DER is not very sensitive to such speakers insofar as each speaker's contribution to the diarization performance metric is time weighted. Additionally, the presence of one or two dominant speakers means that lesser active speakers will be comparatively harder to detect, even if they too have a significant floor time.

Finally, the amount of overlapping speech (averages of 5 minutes cf. 3 minutes per

show), or 10% (GE) vs. 14% (RT‘09) while considering the fraction of the total amount of speech, shows that there is proportionally slightly less overlap speech in the GE dataset than there is in the RT‘09 dataset, but compared to other RT datasets, the overlap speech rate can still be considered as quite high.

Even if there is a shade less overlap speech, the nature of TV shows thus presents unique challenges not seen in meeting data, mainly: the presence of music and other background non-speech sounds, a greater spread in speaker floor time, a greater number of speakers and shorter pauses.

3.4 Baseline System Description

The top-down system is based on the work of LIA [Fredouille & Evans, 2008], while the bottom-up system is based on the work of ICSI [Wooters & Huijbregts, 2008] and more recently I2R [Nguyen et al., 2009] .

3.4.1 Top-Down System

The top-down system described hereafter corresponds to the official system used for LIA-EURECOM’s joint submission to the most recent RT‘09 evaluation [Fredouille et al., 2009] and was developed using the freely available open source ALIZE toolkit [Bonastre et al., 2005]. The system can be decomposed into 5 steps including Pre-Processing, Speech Activity Detection (SAD), Speaker Segmentation and Clustering, Resegmentation and Normalization. Among a number of modifications made to the system used for the RT‘07 evaluation [Fredouille & Evans, 2008] are the use of delay and sum beamforming for the multiple distant microphone (MDM) condition and significant changes to the speaker segmentation algorithm, notably in terms of initialization and speaker modeling which will be highlighted in the following.

1. **Pre-Processing**

   All audio files are treated with Wiener filter noise reduction [Adami et al., 2002b]. Then, if multiple microphones are available (MDM condition) a single virtual channel for each show is created using the BeamformIt v2 toolkit [Anguera, 2006; Anguera et al., 2007] with a 500ms analysis window and a 250ms frame rate. This latter stage is not necessary for the SDM condition. Note that this is the only

difference between the diarization systems used for the MDM and SDM conditions and no delay features are used in any other steps.

2. **Speech Activity Detection (SAD)**

   After preprocessing, speech activity detection (SAD) system is performed in order to isolate useful speech data. SAD is composed of a two-state hidden Markov model (HMM), where each state is associated with 32-component Gaussian mixture model (GMM) trained with an EM/ML algorithm on a large amount of external speech and non-speech data from the RT'04 and RT'05 evaluations[1]. The system utilizes 12 LFCCs and energy augmented by their first and second order derivatives, extracted every 10ms using a 20ms window. First, a single iteration of speech/non-speech Viterbi alignment is performed using equiprobable state transition probabilities in the 2-state HMM and a Viterbi buffer[2] equal to 30 frames. Then the models are adapted by Maximum A Posteriori (MAP) adaptation to ensure that the models adjust to the prevailing ambient conditions, before Viterbi realignment is applied. These two steps are repeated a maximum of 10 times until no more changes occur between two consecutive segmentations. Finally some heuristic duration rules are applied to remove rapid transitions between speech and non-speech states and thus to smooth the output.

3. **Speaker Segmentation and Clustering**

   Working directly on the SAD output, (the previous pre-segmentation stage used in the RT'07 system [Fredouille & Evans, 2008] was removed), the second-stage speaker segmentation and clustering can be considered as the core of the system. It relies on an Evolutive Hidden Markov Model (E-HMM) [Meignier et al., 2000, 2006] where each E-HMM state aims to characterize a single speaker and the transitions represent the speaker turns. All possible changes between speakers are authorized and a Viterbi buffer[2] of 30 frames is used. Here the signal is characterized by 20 unnormalized LFCCs plus energy coefficients computed every 10ms using a 20ms window.

[1] Note that this training set is totally independent of any development set or evaluation set used for later work

[2] The Viterbi buffer allows a fixed state persistence and makes the system more stable

The segmentation and clustering process for each audio show can be defined as follows:

(a) **Initialization:** The E-HMM has only one state, S0 as shown in the Stage 1 of Figure 3.2. A world model of 16 Gaussian components is trained by EM on all of the speech data (cf. 128 Gaussian components for the system described in [Fredouille & Evans, 2008]). An iterative process is then started where a new speaker is added at each iteration.

(b) **Speaker Addition:** At the n^{th} iteration a new speaker model Sn is added to the E-HMM: the longest segment with a minimum duration of 6 seconds (cf. maximum likelihood criterion with 3 sec. minimum in [Fredouille & Evans, 2008]) is selected among all of the segments currently assigned to S0. The selected segment is attributed to Sn and is used to estimate a new GMM with EM training (cf. MAP adaption for the LIA RT'07 system.)

(c) **Adaptation/Decoding loop:** The objective is to detect all segments belonging to the new speaker Sn. All speaker models are re-estimated through a Viterbi realignment and EM learning, according to the current segmentation (EM Algorithm) and a new segmentation is obtained via Viterbi decoding. This realignment/learning loop is repeated while a significant number of changes are observed in the speaker segmentation between two successive iterations.

(d) **Speaker model validation and stop criterion:** The current segmentation is analyzed in order to decide if the newly added speaker model Sn is relevant, according to some heuristic rules on the total duration assigned to speaker Sn. The minimum speaker time allowed is 10 seconds. The stop criterion is reached if there are no more segments greater than 6 seconds in duration available in S0 with which to add a new speaker, otherwise the process goes back to step (b).

Figure 3.2 illustrates the 4 steps described above, during the addition of speaker models S1 and S2 (Stages 2 and 3).

4. **Resegmentation**

The segmentation and clustering stage followed by a resegmentation step which

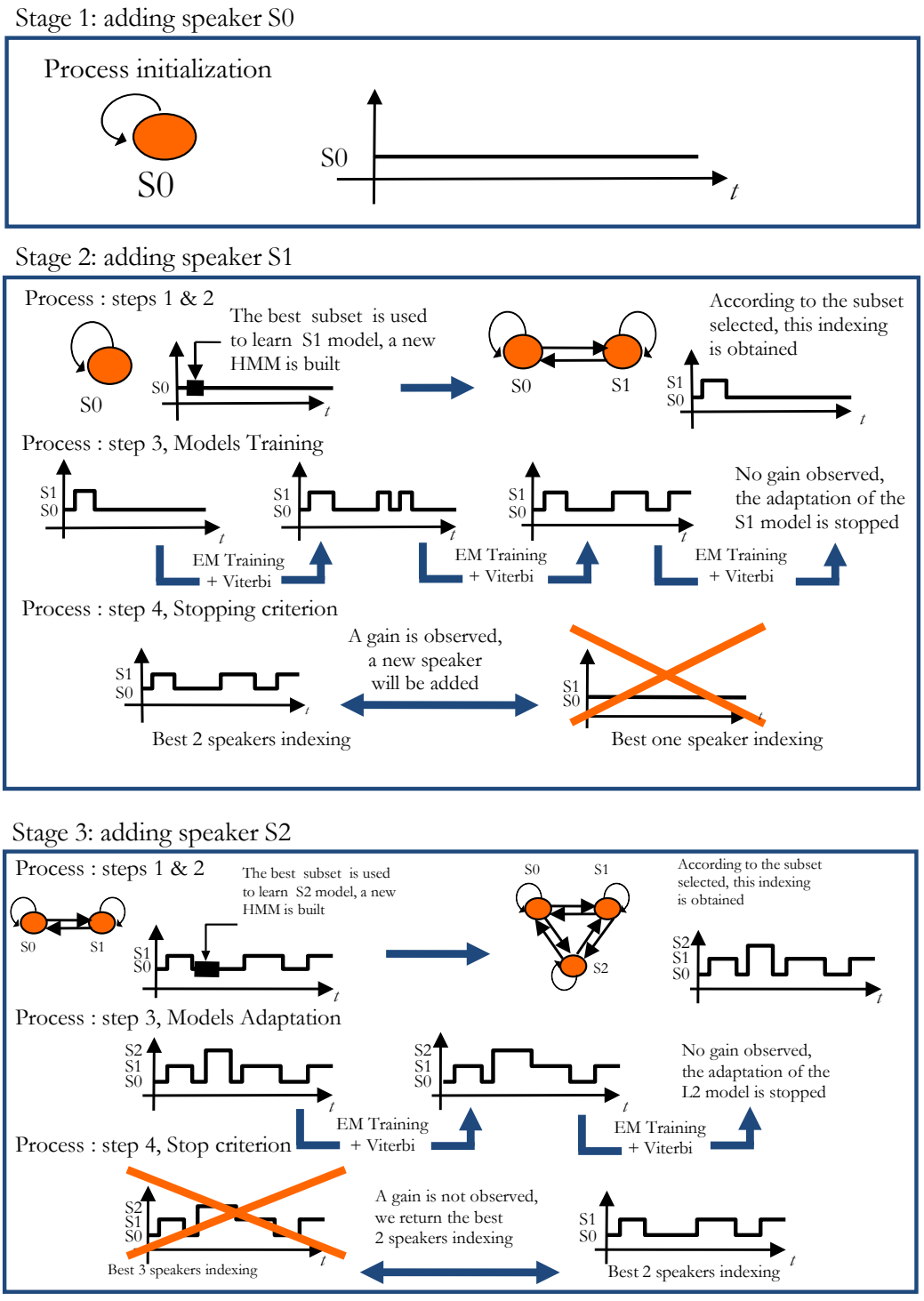


Figure 3.2: Top-down Speaker Segmentation and Clustering: case of 2 Speakers, picture published with the kind permission of Sylvain Meignier (LIUM) and Corinne Fredouille (LIA)

aims to refine the segmentation outputs and to remove irrelevant speakers (e.g. speakers with too few segments). A new HMM is generated from the segmentation output and an iterative speaker model training/Viterbi decoding loop is launched. In contrast to the segmentation stage, here speaker models are adapted by MAP adaptation from an universal background model (UBM) trained on a Speaker Recognition corpus[1]. Note that during the resegmentation process, all the boundaries (except speech/non-speech boundaries) and segment labels are re-examined.

5. **Normalization and Resegmentation**
   Finally a normalization and resegmentation stage is applied using feature vectors composed of 16 LFCCs, energy, and their first derivatives are extracted every 10 ms using a 20ms window. Vectors are normalized, speech segment by speech segment, to fit a zero-mean and unity-variance distribution and a last resegmentation is then applied as described above.

3.4.2 Bottom-Up System

Compared to the top-down strategy, bottom-up systems are much more popular and have consistently obtained the best performance in NIST RT evaluations [NIST, 2007, 2009]. For this reason we chose to put the focus on two systems well representative of the bottom-up clustering state-of-the-art according to. The first bottom-up system is that proposed by ICSI in [Wooters & Huijbregts, 2008]. The second system is our implementation of that proposed by I2R as published in [Nguyen et al., 2009]. On account of a collaboration with ICSI, we were able to work with ICSI's official outputs, thus all results related to this system shown in the following correspond to the official outputs unless otherwise stated. The I2R system was implemented using the open source ALIZE toolkit [Bonastre et al., 2005] and so all related experimental results correspond to our own experimental outputs and cannot be considered as I2R's official outputs. Some details of our implementation are given below.

Moreover it is important to note that the original ICSI and I2R systems are both capable of using time-delay features for MDM conditions in order to help discriminate the

[1]Compared to a speaker diarization corpus this database contains data from many more speakers (in the order of 400)

speakers. In our work however, we are principally interested in the SDM conditions and thus, all details related to time-delay features for speaker discrimination are deliberately omitted. Their only possible use reported here aims to improve the audio quality through a beamforming.

3.4.2.1 ICSI Bottom-up System

ICSI's bottom-up system is an example of Agglomerative Hierarchical Clustering (AHC). Mainly the SAD process and the AHC algorithm are described in the following. Note that a similar front-end acoustic processing, as presented in Subsection 3.4.1, is performed and includes noise reduction and beamforming.

1. **Speech Activity Detection (SAD)**
   As for SAD used in the top-down system, a first model-based speech/non-speech segmentation is performed with a 2-state HMM that contains two GMM models trained previously on speech and non-speech data respectively issued from broadcast news. Only the labels with a high confidence score are kept. Then, among the data classified as non-speech, two sub-clusters are made: regions with low energy (labeled as 'silence') and regions with high energy and high zero-crossing rate labeled as 'non-speech sounds'. Three models corresponding to each of these classes: silence/non-speech sounds/speech are trained and all the data are then reassigned. A final check is made to decide whether the non-speech sounds and the speech are similar enough (BIC similarity) in which case they are merged.

2. **Agglomerative Hierarchical Clustering**
   AHC is applied on the concatenated speech data (with non-speech removed). The system initially over-segments the data into K clusters (where K exceeds the anticipated number of speakers). Then an ergodic hidden Markov model (HMM) is built where the initial number of states is equal to the number of clusters (K). Each of the states is associated with a single probability density function (PDF), and then a probabilistic model is trained for each of the K states. A minimum duration for each state is set to 2.5 seconds[1]. Several iterations of model training and Viterbi alignments are then performed in order to refine the initial models.

[1]Note that this parameter can be compared to the Viterbi buffer in the top-down system introduced in Section 3.4.1

Finally the most closely matching clusters are iteratively merged according to the following procedure:

(a) Run a Viterbi decoding to realign the data;

(b) Retrain the models with an EM algorithm using the new segmentation obtained in step (a);

(c) Select the pair of the closest clusters according to the largest ΔBIC score that is higher than 0.0;

(d) If no pair is detected then the algorithm stops, else the pair detected in step (c) is merged and a new model for the fused cluster is trained;

(e) Go back to step (a)

The stopping criterion as the merging criterion are based on an inter-cluster distance measure which corresponds to a variation of the commonly used Bayesian Information Criterion (BIC) [Chen & Gopalakrishnan, 1998]. It is explained in the following.

Assume we have 2 clusters (C_x, C_y), then ΔBIC aims to compare two hypotheses:

- (H_1) a situation where (C_x, C_y) correspond to two different speakers:
  $\Longleftrightarrow C_x \in Speaker_x; C_y \in Speaker_y; Speaker_x \neq Speaker_y$
- (H_2) a situation where (C_x, C_y) correspond to one same speaker:
  $\Longleftrightarrow C_x \cup C_y = C_z; C_z \in Speaker_x; Speaker_x = Speaker_y$

According to [Chen & Gopalakrishnan, 1998], ΔBIC can be expressed as follows:

$$\begin{aligned} \Delta BIC(C_x, C_y) &= BIC(H_1) - BIC(H_2) \\ &= n_z \log|\Sigma_z| - n_x \log|\Sigma_x| - n_y \log|\Sigma_y| \qquad (3.3) \\ &\quad -\lambda \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log n_z \end{aligned}$$

(3.4)

Where: $n_z = n_x + n_y$

n_x,n_j are the number of frames assigned to each cluster

Σ_x,Σ_y are the covariance matrices for each cluster

Σ_z is the covariance matrix shared by both clusters

λ is a tunable parameter

The ICSI system uses a variation of ΔBIC, as reported in [Ajmera et al., 2004], and does not require the tunable parameter λ present in the original algorithm [Chen & Gopalakrishnan, 1998]. This is achieved by ensuring that, for any given ΔBIC comparison, the difference between the number of free parameters in the two hypotheses is zero.

3.4.2.2 I2R Bottom-up System

I2R's system [Nguyen et al., 2009] differs from ICSI's system mainly in its initialization, and its merging and stopping criteria. We detail hereafter these two particular steps and the configuration we chose for our implementation.

1. **Pre-processing & SAD**
   In exactly the same fashion as the top-down system in 3.4.1, Wiener filtering noise reduction and beamforming are first performed on each of the MDM channels to obtain a single pseudo channel for subsequent processing. For practical reasons, the SAD process from the top-down approach is then applied, instead of the I2R's SAD published in [Nguyen et al., 2009]. Note that the top-down SAD performances are comparable to I2R's SAD outputs.

2. **Initialization: Sequential EM**
   The diarization system is initialized with 30 homogeneous clusters of uniform length and a 4-component GMM is trained by EM/ML on the data in each cluster. Each cluster is then split into segments of 500ms in length and the top 25% of segments which best fit the GMM are identified and marked as classified. The remaining 75% of worst-fitting segments are then gradually reassigned to their closest GMMs, K segments at a time (the value of K is not published in [Nguyen et al., 2009], however our implementation shows that the system is not overly sensitive to this parameter), with iterative Viterbi realignment and adaptation until all segments are classified.

3. **Agglomerative Hierarchical Clustering**

   After the Segmental EM initialization, conventional AHC is performed. Models are retrained with 16 Gaussian components. Cluster merging is controlled with the Information Change Rate (ICR) criterion [Han et al., 2008]. ICR is a BIC-like criterion and is defined for two clusters C_x, C_y as a normalized version of the Generalized Likelihood Ratio (GLR):

$$ICR(C_x, C_y) \triangleq \frac{1}{n_x + n_y} \log GLR(C_x, C_y) \tag{3.5}$$

   where

$$GLR(C_x, C_y) = \frac{P(x \cup y|H_1)}{P(x \cup y|H_2)} \tag{3.6}$$

   and where H_1 and H_2 are the same hypotheses that the ones set in 3.4.2.1. Parameters x and y are the feature vectors related to each of the clusters C_x, C_y, and n_x, n_y are the respective size of each cluster (number of assigned features).

   If each cluster C_x, C_y and $C_z = C_x \cup C_y$ is modeled by a probability density function (PDF) f_X, f_Y and f_Z with the following parameters $\theta_{f_X}, \theta_{f_Y}$ and θ_{f_Z} then the GLR can be rewritten as:

$$GLR(C_x, C_y) = \frac{p(x|f_X; \theta_{f_X}) \cdot p(y|f_Y; \theta_{f_Y})}{p(z|f_Z; \theta_{f_Z})} \tag{3.7}$$

   In this way, clusters are sequentially merged with embedded Viterbi realignment until only a single cluster remains. Each intermediate segmentation hypothesis is retained for subsequent processing.

4. **Choice of the Best Segmentation**

   After the set of hypothesized segmentations is determined, the best is selected according to metric which estimates the segmentation quality. The original work [Nguyen et al., 2009] used the *Rho* clustering quality metric [Nguyen et al., 2008], however we use the T_s metric [Nguyen et al., 2008] since we find that it leads to better performance. The T_s clustering quality metric is based on the inter and intra-feature vector distribution and works as follows:

Let $C^{(i)}$ be a segmentation of speech data X into K_i clusters $C^{(i)} = \{C_1^{(i)}, C_2^{(i)}, ..., C_{K_i}^{(i)}\}$. We denote by $d(x_m, x_n)$ the distance between two feature vectors x_m, x_n and define the population of intra-cluster distances by D_{intra} and the population of inter-cluster distances by D_{inter} as defined below:

$$D_{intra} = \bigcup_{i=1}^{K} D(C_i, C_i) \tag{3.8}$$

$$D_{inter} = \bigcup_{1 \le i < j \le K} D(C_i, C_j) \tag{3.9}$$

$$\text{where } D(C_i, C_j) = \{d(x_m, x_n) | x_m \in C_i, x_n \in C_j, \forall m \forall n\} \tag{3.10}$$

If we assume that **the distributions of the two populations D_{intra} and D_{inter} to be Gaussian**, we can measure their separation with the T_s metric according to:

$$T_s = \frac{m_{inter} - m_{intra}}{\sqrt{\frac{\sigma_{inter}^2}{n_{inter}} + \frac{\sigma_{intra}^2}{n_{intra}}}} \tag{3.11}$$

where $m_{inter}, \sigma_{inter}, n_{inter} (m_{intra}, \sigma_{intra}, n_{intra})$ are respectively the mean, standard deviation and size of D_{inter} (D_{intra}).

5. **Post-Processing**

This final post-processing step described in the following is not included in I2R's system, but was found to bring some improvements. Similar to the resegmentation and normalization steps described for the top-down system, speaker models are retrained by MAP adaptation with 128 components and several repetitions of Viterbi realignment and adaptation are performed to improve the segmentation. Speakers with less than 8 seconds of data are removed and the process is repeated until a stable diarization hypothesis is reached. Then a final resegmentation is performed, but this time using features which are normalized segment-by-segment to fit a zero-mean and unity-variance distribution. This step also uses the MAP adaptation of a background model with 128 components.

System	Dev. Set	RT07	RT09	GE
Top-down	22.7/20.0	18.3/15.0	26.0/21.5	40.4/36.0
Bottom-Up (I2R)	21.7/18.9	23.8/20.8	19.1/13.5	33.7/29.0
Bottom-up (ICSI)	-/-*	21.3/17.9	31.2/26.5	-/-*

Table 3.2: % Speaker diarization performance for Single Distant Microphone (SDM) conditions in terms of DER with/without scoring the overlapped speech, for the Dev. Set and the RT'07, RT'09 and GE datasets. *Note that results for ICSI's system corresponds to the original outputs and have not been forthcoming for the Dev. Set and GE.

3.5 Experimental Results

Performance of the different baseline systems presented in the Section 3.4 are illustrated in Table 3.2 for the development dataset, for two RT datasets and the GE TV-show dataset. More details for RT'07, RT'09 evaluation datasets are given in Tables 3.3 and 3.4.

All results in Table 3.2 are reported with/without scoring the overlap speech. For all of the 3 systems we can observe a large difference in performance with and without the scoring of overlap speech on the RT'09 and GE datasets. The degree of overlapping speech is known for being particularly high on the RT'09 and GE datasets (14% and 10% cf. 8% for RT'07) and thus this is only to be expected.

When comparing top-down performance to the best bottom-up baseline system we can observe that the top-down baseline delivers the best results for RT'07 dataset, it shows some competitive scores for the development set, but it is outperformed by I2R bottom-up system for RT'09 and GE datasets. Among the two bottom-up systems, results on RT'07 and RT'09 show that none is definitely better and while ICSI's system performs better on RT'07 dataset, I2R's system provides the best baseline on RT'09.

Tables 3.3 and 3.4 give the SAD error, the speaker error and the overall DER for each of the meetings of RT'07 and RT'09 datasets. As we described in Section 3.4, the top-down system and I2R's systems have the same SAD process which outperforms SAD performance for ICSI's system (3.4% vs 6.1% for RT'07, and 3.2% vs 9.9% for RT'09). While looking at the speaker error, it is interesting to highlight that the tendency in terms of variation of the speaker error is not always the same according to the system: e.g. while I2Rs system performs very well for the meeting *NIST_20080307-0955*, the two

other systems perform more than 3 times worse; conversely when the top-down system outputs a speaker error of 0.5% for the meeting *VT_20050408-1500*, I2R's bottom-up system performs with 22.9% speaker error. We can hypothesize from these results, a difference of behavior between these two types of clustering which may then suffer from different weaknesses and leading to different performance.

In the results related to RT'09 dataset we can notice a meeting (*NIST_20080201-1405*) for which all of the three systems perform poorly. The difficulty of this meeting was already reported by the community [Anguera et al., 2011], and can be attribute to the high degree of overlap speech and the very small speaker turns caused by the spontaneity of the speech.

3.6 Discussion

This chapter introduces the official protocols used for the diarization challenge in the NIST RT evaluations and the Diarization Error Rate, the official metric to estimate the quality of the hypothesized diarization output. The different datasets used throughout the remainder of this thesis as described with an emphasis on their main characteristics. We present 3 official baseline systems representative of the state-of-the-art, and experimental results for each on independent development and evaluation datasets.

Experimental results show that top-down strategy leads to competitive results and outperforms the bottom-up strategy on one dataset. Each of the systems seem to have their own strengths and weaknesses while none is consistently better than the others. In this context we detail in the next chapter a comparative study for these 2 clustering strategies in order to understand their difference in behavior.

Table 3.3: Results for RT'07 dataset with SDM conditions without scoring the overlap speech. Given in the following order: the Speech Activity Detector error (SAD), the Speaker Error (S_{Error}), and the DER

Meetings ID RT'07	Top-Down			Bottom-up (I2R)			Bottom-Up (ICSI)		
	SAD	S_{Error}	DER	SAD	S_{Error}	DER	SAD	S_{Error}	DER
CMU_20061115-1030	5.0	10.3	15.3	5.0	31.1	36.1	11.5	20.1	31.6
CMU_20061115-1530	5.5	12.0	17.5	5.5	12.5	18.0	5.9	11.0	16.9
EDI_20061113-1500	3.0	30.0	33.0	3.0	19.3	22.3	6.2	20.0	26.2
EDI_20061114-1500	3.1	25.2	28.3	3.1	29.5	32.6	6.0	14.3	20.3
NIST_20051104-1515	1.8	6.7	8.5	1.8	6.2	8.0	2.7	1.8	4.5
NIST_20060216-1347	3.1	5.1	8.2	3.1	6.0	9.1	4.6	3.0	7.6
VT_20050408-1500	3.7	0.5	4.2	3.7	22.9	26.6	8.6	7.8	16.4
VT_20050425-1000	1.6	7.3	8.9	1.6	12.6	14.2	3.5	20.0	23.5
Overall Error	3.4	11.6	15.0	3.4	17.4	20.8	6.1	11.8	17.9

Table 3.4: Same as in 3.3 but for RT'09 dataset

Meetings ID RT'09	Top-Down			Bottom-up (I2R)			Bottom-Up (ICSI)		
	SAD	S_{Error}	DER	SAD	S_{Error}	DER	SAD	S_{Error}	DER
EDI_20071128-1000	5.9	2.2	8.1	5.9	5.9	11.8	16.2	2.0	18.2
EDI_20071128-1500	5.1	35.2	40.3	5.1	19.8	24.9	5.9	5.2	11.1
IDI_20090128-1600	0.9	11.3	12.2	0.9	3.5	4.4	11.2	4.0	15.2
IDI_20090129-1000	3.8	10.1	13.9	3.8	8.3	12.1	6.5	13.9	20.4
NIST_20080201-1405	3.6	55.6	59.2	3.6	40.0	43.6	17.9	43.5	61.4
NIST_20080227-1501	1.4	11.2	12.6	1.4	6.2	7.7	6.6	33.6	40.2
NIST_20080307-0955	1.9	27.0	28.9	1.9	6.1	8.0	5.9	38.9	44.8
Overall Error	3.2	18.3	21.5	3.2	10.2	13.5	9.9	16.7	26.5

Chapter 4

Oracle Analysis

In Chapter B.2 we introduced two main techniques for the task of speaker diarization involving bottom-up and top-down hierarchical clustering. Although these technologies represent the state-of-the-art in the field, one could still wonder what their real strength and weakness are and how they can be improved.

In this chapter we analyze the performance of each step of the two approaches. To achieve this goal, a global 'blame game' as defined in [Huijbregts & Wooters, 2007] is carried out in order to detect the sensitive steps of each system through a series of oracle experiments. Section 4.1 first introduces the protocol and dataset used for this oracle study, then the oracle setup used for the top-down system is described in Section 4.2 and experimental results are given. The same approach is followed in Section 4.3 for the bottom-up scenario. Finally a comparison of the oracle observations is presented in Section 4.4.

4.1 Oracle Protocol

The term *Oracle* comes from Latin and means 'to speak'. It refers in the classical antiquity to a person considered to be a source of prophetic predictions of the future inspired by the gods. With the same analogy, an oracle experiment is a setup where the system can make use of all available knowledge, even the ground-truth transcripts. In that sense the system is an *Oracle* which knows everything.

Oracle experiments were already used in the field of speaker diarization. In [Huijbregts & Wooters, 2007] oracle experiments were performed in order to high-

light the impact of overlapped speech in a bottom-up system. In [Han et al., 2008] oracle experiments were used to analyze the performance of different stopping criteria. Finally in [Huijbregts et al., 2012] a complete analysis, a so-called 'blame game' of the bottom-up system introduced by ICSI and reported in Section 3.4.2.1 was performed. Thanks to a full set of oracle experiments the impact in terms of DER of each of the system component was quantified and some improvements in the system were proposed.

In this chapter we follow the same oracle framework as in [Huijbregts et al., 2012; Huijbregts & Wooters, 2007] but for our top-down baseline system. We hypothesize that components perform independently and the overall error corresponds to the sum of the error of each component. Assuming this, we can then replace all experimental components by their corresponding oracle setup and then iteratively place back in the system the experimental setup to measure the contribution of each component. In order to make a fair comparison and run some consistent experiments, we keep exactly the same dataset and acoustic conditions than in [Huijbregts et al., 2012]. The dataset used for all the oracle experiments is composed of 27 meetings and shown in Table 4.1. The reference transcripts were obtained by forced alignment of the reference speech transcriptions in order to avoid inconsistencies in the placement of segment boundaries[1]. The same recording conditions are considered i.e. a single pseudo channel is extracted from the MDM conditions where noise reduction is first applied followed by beamforming. No delay features are exploited.

4.2 Oracle Experiments on Top-Down Baseline

The 'blame game', as defined in [Huijbregts et al., 2012], aims to compute the contribution in terms of DER of each system component thanks to the use of all the available knowledge, including the official ground-truth. During this analysis we assumed that the performance of each component is mostly **independent** of the performance of the others. We accept that this hypothesis is approximate and that changing one component may impact on subsequent steps. However oracle experiments permit to give a first diagnosis of the weaknesses of a system with a limited amount of experiments. We first

[1]The realignment was made by Marijn Huijbregts and kindly shared with us, allowing a strict comparison between our top-down oracle experiments and those of the bottom-up system published in [Huijbregts et al., 2012]

Meetings ID		
AMI_20041210-1052	EDI_20050218-0900	NIST_20051104-1515
AMI_20050204-1206	EDI_20061113-1500	NIST_20060216-1347
CMU_20050228-1615	EDI_20061114-1500	TNO_20041103-1130
CMU_20050301-1415	ICSI_20000807-1000	VT_20050304-1300
CMU_20050912-0900	ICSI_20010208-1430	VT_20050318-1430
CMU_20050914-0900	NIST_20030623-1409	VT_20050408-1500
CMU_20061115-1030	NIST_20030925-1517	VT_20050425-1000
CMU_20061115-1530	NIST_20051024-0930	VT_20050623-1400
EDI_20050216-1051	NIST_20051102-1323	VT_20051027-1400

Table 4.1: List of meetings used for these oracle experiments. All of these 27 meetings are extracted from our development set issued from RT'04 '05 '06 '07 datasets and are the same data used for the Blame Game in [Huijbregts et al., 2012].

describe five different oracle experiments with our top-down baseline system described in 3.4.1. Note that some of these experiments are specific to the system and are different from the oracle analysis of the bottom-up system presented in 4.3.

4.2.1 Experiments

In order to assess the performance of separate system components we first replace all components by an oracle setup and measure the DER. Then, in a top-down fashion, the actual components are successively placed back into the system such that subsequent steps are still oracle. We have to emphasize that, due to its iterative nature, i.e. the loop between each speaker addition and realignment, it is not possible to perform the experiments perfectly top-down, but the list of experiments we propose aims to minimize this effect. Note moreover that the pre-processing step is not evaluated.

Experiment 1: Perfect Topology:

In this first experiment, all steps are substituted by an oracle setup. The perfect SAD ground-truth is used. However, since our top-down system is not able to score overlapping speech, some missed speech will be included in the SAD error. Each of the speaker models is iteratively introduced into the E-HMM and trained on the totality of the data of each speaker. The generic model $S0$ is optimally trained at each iteration with the rest of the speakers not yet included in the E-HMM.

Despite these optimal conditions, we cannot expect to get perfect performances for different reasons. First the system is not able to handle overlapping speech, second the speaker modeling cannot be perfect due to the limited complexity of the GMMs.

Experiment 2: Speech Activity Detection:

In the second experiment the actual SAD component is put back into the system in order to evaluate its contribution in DER. All other steps are still oracle. The speaker models are trained on the ground-truth as previously, according to the SAD reference, but the Viterbi realignment are performed on the experimental SAD outputs. Note that while changing the SAD we may expect a difference of speaker error since first, the Viterbi decoding is applied speech segment by speech segment and second, the state alignment to a non-speech frame (case of false alarm) may deteriorate the Viterbi decoding in the neighborhood of this frame. The difference of error between experiments 1 & 2 can be attributed to the SAD component.

Experiment 3: Speaker Initialization:

The third experiment differs from Experiment 2 since the new added speakers are now trained on data chosen automatically by the speaker diarization algorithm. At each speaker addition, the system uses the longest speech segment left in the cluster $S0$ and trains a new speaker model. Note, however, that the model related to $S0$ is still trained artificially on the data belonging to the speakers out of the current speaker inventory. The stopping criterion is still controlled by an oracle setup, i.e. the hypothesis which minimizes the DER is kept.

Experiment 4: S0 training:

This experiment aims to show the importance of $S0$ being independent from the other models i.e. $S0$ must theoretically be composed of only non-introduced speakers. The setup is the same as for Experiment 3, except that the model related to $S0$ is now trained according to the segmentation hypothesis. Here again the stopping criterion is optimized artificially.

Experiment 5: Stop Criterion:

In this last experiment all components are placed back in the system except the parameter deciding the minimum speaker time which is still artificially computed (Orcale). This last experiment aims to estimate the sensitivity and strength of the system toward the stop criterion. Note that the difference in performance between this experiment and the

experimental baseline enables an estimation of the contribution of the minimum speaker time for speaker validation.

4.2.2 Experimental Results

Results are illustrated in Table 4.2 and show both SAD and DER scores for each of the five experiments both with and without the scoring of the overlapping speech. Since at each following experiment, one step of the original approach is placed back into the system, and assuming that the components perform independently of each other, the increase in DER can be considered as the contribution to the total error of the component in the system. For the following analysis we will focus on the results whit scoring the overlap speech for consistency with the work in [Huijbregts et al., 2012].

In the first experiment, referred to as *Perfect Topology* all steps are oracle. Even if the SAD reference was used we still get a SAD error of 3.50% while scoring the overlap speech since our system is not able to handle the overlap speech. This error rate is reported in Table 4.3 as the contribution in DER due to the overlap speech. The global DER for this experiment shows a speaker error of 3.36% despite the perfect oracle setup. This error can be explained since the speaker modeling and the Viterbi alignment, due to their probabilistic nature and their limited complexity cannot perform perfectly.

While adding the actual SAD step into the system, we note an increase in DER of 4.83%. The new DER includes the increase of SAD error (+3.70%) and of speaker error (+1.13% compared to the *Perfect Topology*). This is explained by the segmental Viterbi decoding and the speaker modeling which cannot be as accurate as before while introducing non-speech frames as highlighted in [Fredouille & Evans, 2007].

In experiments 3, 4 and 5, the speaker addition is made experimentally as proposed in the original system. In experiment 3, we first constrain artificially the general model attributed to $S0$ in order that it is independent from speaker models already added. Despite this constraint, we observe an increase of DER of 0.76% due to the new model initialization. While removing the constraint for the training of $S0$ in experiment 4, the overall DER deteriorates by 4.20%. Note however that the effect of the speaker model initialization and the quality of the general model $S0$ are closely tied together and can hardly be dissociated. Indeed, in the case of a perfect training of $S0$ totally independent of the already introduced speakers, the choice and the initialization of a new speaker

oracle Experiment	**With Scoring Ovlp**		**Without Scoring Ovlp**	
	SAD(%)	DER(%)	SAD(%)	DER(%)
1. Perfect Topology	3.50	6.86	0.00	3.43
2. Speech Activity Detection	7.20	11.69	4.00	8.52
3. Speaker Initialization	7.20	12.45	4.00	9.36
4. S0 training	7.20	16.65	4.00	13.59
5. Stop criterion	7.20	17.83	4.00	14.77
Top-Down Baseline System	7.20	18.74	4.00	15.74

Table 4.2: The SAD and DER error rates for six oracle experiments on the top-down system with and without scoring the overlap speech. Details of each of the experiments are given in Section 4.2.2

Error Name	**With Scoring Ovlp**		**Without Scoring Ovlp**	
	DER(%)	Relative	DER(%)	Relative
Overlapping speech	3.50	18.68%	0.00	0.00%
Speech Activity Detection	4.83	25.77%	5.09	32.34%
Modeling/Alignment	3.36	17.93%	3.43	21.79%
Models initialization	0.76	4.06%	0.84	5.34%
Robustness of S0 model	4.2	22.41%	4.23	26.87%
Stop clustering too early/late	1.18	6.30%	1.18	7.50%
Minimum Time Speaker accepted	0.91	4.86%	0.97	6.16%
System (Sum of the DERs)	18.74	100.00%	15.74	100.00%

Table 4.3: Contribution of each of the top-down system component to the overall DER

model among the data associated to the cluster $S0$ will obviously be less noisy and less likely to lead to a redundant speaker.

Finally we compare results for experiments 4 and 5 which aim to evaluate the sensitivity of the system to the stopping criterion. We note that the use of the experimental stopping criterion leads to an increase in DER of 1.18%. Examining the final baseline and experiment 6 permits us to attribute an increase in DER of 0.91% to the minimum speaker time allowed.

Table 4.3 summarizes all the DER contributions with and without the scoring of overlap speech. For both situations the same trend can be observed: the SAD error and the quality of the general model $S0$ are the main weaknesses of the system and can be held

accountable for almost 50% of the DER. The effect of $S0$ not being totally independent from the already added speakers leads to a system not discriminative enough. As a result, after Viterbi decoding, a lot of speech is assigned to $S0$ instead of the correct corresponding speaker, leading to some possible artifacts for new speaker initialization.

Another weakness highlighted by this set of experiments, except that of overlapped speech which is not processed by our system, is the inaccuracy in terms of modeling and alignment. A comparison of these contributions with those obtained with a bottom-up system are discussed in Section 4.4.

4.3 Oracle Experiments on Bottom-up Baseline

Huijbregts et al. report comparable experiments in [Huijbregts et al., 2012] for a bottom-up approach comparable to ICSI's system. Since we used exactly the same corpus and the same acoustic conditions we report in this section the results published in [Huijbregts et al., 2012] to facilitate a comparison of the two approaches[1].

4.3.1 Experiments

Huijbregts et al. proposed a set of six oracle experiments in order to highlight the contribution of each component to the DER, assuming each component to be **mostly independent** of the performance of others. All components are first replace by their corresponding oracle setup, then the actual components are successively placed back into the system in a top-down fashion. Their results are reproduced in Table 4.4. A short description of the oracle experiments is reported here, but more details can be found in [Huijbregts et al., 2012].

Experiments to test the quality of the merging algorithm, the cluster initialization, the model combination and the stop criterion are specific to the bottom-up nature of the clustering and are described hereafter, while other experiments have comparable protocols to those presented in Section 4.2.1. In all experiments, downstream components are always replaced by their oracle setup.

Merging Algorithm:

The experiment aims to test the influence of the actual merging algorithm on the final result. The system first creates 16 initial clusters with the help of the ground-truth to

[1]Results reproduced with the kind permission of Marijn Huijbregts

insure that each model is trained with the speech of one speaker. The decision about which models to merge and when to stop is performed according to the Oracle.

Cluster Initialization:

The initial 16 clusters are created by splitting the speech data randomly

Merge Candidate Selection:

The clusters to merge are selected according to the original selection based on the BIC criterion.

Stop Criterion:

The component deciding when to stop the merging process is replaced by its original implementation.

4.3.2 Experimental Results

Error Name	**With Scoring Ovlp**	
	DER(%)	Relative
Overlapping speech	3.50	21.21%
Speech Activity Detection	3.20	19.39%
Modeling/Alignment	2.20	13.33%
Merging algorithm	1.19	7.21%
Non-perfect initial clusters	0.80	4.85%
Combining wrong models	3.35	20.30%
Stop Speaker Addition too early/lat	2.26	13.70%
System (Sum of the DERs)	16.50	100.00%

Table 4.4: Contribution of each of the bottom-up system component to the overall DER as published in [Huijbregts & Wooters, 2007] for the dataset shown in Table 4.1. Results reproduced with the kind permission of Marijn Huijbregts.

By comparing the consecutive oracle experiments, a part of the overall diarization error rate is assigned to each of the components of the bottom-up system. Table 4.4 lists the contribution of each component to the total DER. Results show that overlapping speech, SAD and the merging criterion are together responsible for more than 60% of the overall error.

4.4 Discussion

Tables 4.3 and 4.4 present a fair performance comparison over the same dataset of each component in the top-down and bottom-up clustering algorithms. The overall DER shows that the bottom-up approach slightly outperforms the top-down system with an overall DER of 16.50% vs. 18.74%. However, it must be emphasized that the overall SAD error is a bit lower for the bottom-up system i.e. an estimate of the SAD error can be found if we consider the speaker error to be independent of the SAD quality. In fact, we observe an increase of DER of 4.83% for the top-down system while using the experimental SAD, vs 3.20% for the bottom-up, which leads approximately to a higher SAD error of 1.60% absolute for the top-down system.

The contribution of the modeling / alignment seems to be higher in terms of absolute DER for the top-down approach (3.36% vs. 2.20% for the bottom-up approach). This is due to the iterative nature of the top-down approach. Indeed, compared to a bottom-up system, modeling and realignment have to be performed for each speaker addition, accumulating thereby consecutive errors due to modeling/realignment imperfections.

The stopping criterion is a common component to both of the systems although precise approaches differ. It is important to notice that the stopping criterion for the bottom-up scenario has an important role and contributes to almost 14% of the DER, while it represents only 6% of the DER for the top-down approach. Moreover the contribution of the merging criterion represents 20.30% of the overall DER in the bottom-up system. The contribution of these two criteria together corresponds to more than one third of the overall DER and confirms as explained in [Han et al., 2008] the low robustness of BIC and ΔBIC criteria mainly in case of cluster impurity.

In contrast, while the bottom-up system is almost independent to its initialization (an increase of 0.80% DER is observed while doing a random initialization instead of a supervised initial splitting), the top-down system is very sensitive to the quality of the $S0$ model which should, in a perfect world, be trained on speakers out of the current speaker inventory[1] which affect directly the model initialization[2].

[1] The speaker inventory corresponds to the speakers already introduced in the E-HMM

[2] We pick up the longest segment in the cluster $S0$ to introduce a new speaker

As a conclusion it is worth noting that, except for the SAD error and the presence of overlap speech which are some common problems to both systems, bottom-up and top-down clustering have some specific weaknesses. Indeed, while the bottom-up system is almost independent of its initialization, it is mainly sensitive to performance of the components located at the bottom of the system: e.g. merging and stopping criteria can perform poorly, particularly in case of cluster impurity. In contrast, the top-down scenario is mainly sensitive to the steps situated at the top of the system, namely the initialization and the training of the general model $S0$ which influences its discriminative capacity.

Chapter 5

System Purification

Chapter 4 shows through a set of oracle experiments that top-down clustering compared to the bottom-up approach suffers from low speaker discrimination mainly due to the quality of the general model $S0$. In this chapter we investigate the possibility to correct some artifacts caused by the low speaker discrimination, with the help of a new purification component we published in [Bozonnet et al., 2010]. The new purification process is applied after the segmentation and clustering process as a post-processing. This approach to purification is first added to the top-down system, then, its effect on the bottom-up system is investigated also.

The remainder of this chapter is organized as follows. Section 5.1 describes the new purification algorithm. Experiments with the top-down approach are presented in Section 5.2, while Section 5.3 details experiments conducted with the bottom-up system.

5.1 Algorithm Description

Purification is not a new idea and several different purification approaches have been reported, e.g. [Anguera et al., 2006b]. In contrast to this previous work using bottom-up systems we here seek to demonstrate the potential for cluster purification specifically in *top-down* approaches. Our approach is based on sequential initialization which was first proposed in [Nguyen et al., 2009] by I2R-NTU researchers at the NIST RT'09 evaluations [NIST, 2009]. This system is described in 3.4.2.2.

Sequential initialization algorithm used in [Nguyen et al., 2009] initializes 30 homogeneous clusters split randomly. We have found it necessary to modify this approach in

Figure 5.1: Scenario of the diarization system including the new added cluster purification component.

order to bring its potential to the E-HMM system. Indeed, in our system and as shown in Figure 5.1, purification is applied after segmentation and clustering, which produces a number of clusters (generally only a few more than the true number of speakers) each of which, ideally, corresponds to a single speaker. Of course there remains the distinct potential for impurities and our experiments on development data have shown that speaker clusters are typically between 50% and 95% pure.

Thus, in contrast to the bottom-up approach, where the initial clustering is generally random and uniform, our cluster purification algorithm operates on clusters which should already contain a dominant speaker. The original algorithm was intended for clusters of relatively lower initial purity and we have found that, the same algorithm with little modifications, can, in some cases, reduce cluster impurity.

The modified algorithm first trains, by EM/ML, a 16-component GMM on the data of each cluster identified by the segmentation and clustering component (vs 4-component GMM in [Nguyen et al., 2009]). Each cluster is then split into segments of 500ms in length and the top 55% of segments which best fit the GMM are identified and marked as classified (vs. 25% of segments in [Nguyen et al., 2009]). The remaining 45% of worst-fitting segments are then gradually reassigned to their closest GMMs, with iterative Viterbi decoding and adaptation until all segments are classified. As for the segmentation and clustering component, the system utilizes 20 unnormalized LFCCs plus energy coefficients computed every 10ms using a 20ms window.

5.2 Experimental Work with the Top-Down System

Experiments presented in this section aim to demonstrate the improvements in diarization performance obtained on the top-down system while adding the new cluster purification algorithm described in Section 5.1.

We report experiments on a development dataset comprising meeting shows from the NIST RT'04, '05 and '06 datasets (23 shows in total). This set alone was used to optimize the purification algorithm and is the same used for baseline optimization reported in Section 3.4. In addition we present results on a separate evaluation set, namely the NIST RT'07 dataset (8 shows) and also validate improvements in performance on unseen data in the NIST RT'09 evaluation dataset (7 shows). Additionally to assess the stability of the system, performances are tested on the TV-show corpus Grand Échiquier (GE)(7 shows).

In order to give a more meaningful assessment of our core diarization system, independently of beamforming performance and fused delay features, we only report results on the SDM condition. Diarization performance is assessed according to the standard setup introduced in 3.2. All analyses in terms of DER are made without scoring the overlapping speech.

5.2.1 Diarization Performance

Table 5.1 illustrates a comparison of speaker diarization performance for the SDM condition using the two different top-down system variations (with and without purification) and the four different datasets (columns 2 to 9). All results are given with (OV) and without (NOV) the scoring of overlap speech.

The purification algorithm has a small effect on the Development Set and leads to a relative improvement of 9% (18.3% cf. 20.0%) over the top-down baseline. Results are almost identical on the RT'07 dataset (4% relative improvement) but are markedly improved on the RT'09 dataset. Here results of 21.5% without purification and 16.0% with purification correspond to a relative improvement of 26% (18% with scoring overlapping speech). Finally, results on the GE corpus show a small improvement (6% relative). Thus the purification algorithm gives as good or better results and helps to stabilize the results across the three datasets.

Table 5.2 details the SAD error, the speaker error (S_{Error}) and the DER for each show of the RT'07 and RT'09 datasets without scoring the overlapping speech. For the RT'07 dataset, the 8 first lines of Table 5.2 indicates that globally the speaker error decreases after purification. However while it is the case for main of the meetings, we observe that performance over one show is significantly deteriorated. Indeed, for the meeting *CMU_20061115-1530* we notice a deterioration of the speaker error of 16% absolute.

System	Dev. Set		RT'07		RT'09		GE	
	OV	**NOV**	**OV**	**NOV**	**OV**	**NOV**	**OV**	**NOV**
Top-down Baseline	22.7	20.0	18.3	15.0	26.0	21.5	40.4	36.0
Top-down Baseline+Pur.	21.1	18.3	17.8	14.4	21.1	16.0	38.5	33.9

Table 5.1: A comparison of diarization performance on the Single Distant Microphone (SDM) condition and four different datasets: a development set (23 meetings from RT'04, RT'05, RT'06), an evaluation (RT'07), a validation (RT'09) and a TV-show dataset: Grand Échiquier(GE). Results reported for two different systems: the top-down baseline as described in Section 3.4.1 and the same system using cluster purification (Top-down Baseline+Pur.). Results illustrated with(OV)/without(NOV) scoring overlapping speech.

In contrast, some shows are improved more or less significantly when purification is applied e.g. the speaker error of the meeting *EDI_20061113-1500* decreases by more than 18% absolute. The last 7 lines of Table 5.2, details the performance of the system for the RT'09 dataset. Compared to performance over the RT'07 dataset, we observe a consistent improvement for the speaker error of each show including improvement until 19% absolute speaker error (*EDI_20071128-1500*).

It is of interest to understand why the algorithm performs significantly better on the RT'09 dataset than on the development dataset on which it was optimized and in the following we analyze the effect of purification on the cluster quality thanks to a measure of the purity.

5.2.2 Cluster Purity

To help explain this behavior we measured the cluster purity statistics before and after purification. For this we introduce an additional metric (%Pur) which is specifically designed to assess the performance of the purification algorithm. Among all of the data assigned to any one cluster we simply determine the percentage of data that corresponds to the most dominant speaker, as determined according to reference transcriptions. The %Pur metric is the average purity for all speaker models after segmentation and clustering and performance is gauged by comparing %Pur before and after purification. Note that the DER is not appropriate for assessing purity as it penalizes the case where there are more models than speakers - this is generally the case with our algorithm (the later resegmentation stage aims to reduce their number). Thereafter the final DER metric is

			Top-down Baseline		Top-down Baseline + Purification	
	Meeting ID	SAD	S_{Error}	DER	S_{Error}	DER
RT07 SDM	CMU_20061115-1030	5.0	10.3	15.3	9.9	14.9
	CMU_20061115-1530	5.5	12.0	17.5	27.5	33.0
	EDI_20061113-1500	3.0	30.0	33.0	11.8	14.8
	EDI_20061114-1500	3.1	25.2	28.3	24.5	27.6
	NIST_20051104-1515	1.8	6.7	8.5	6.3	8.1
	NIST_20060216-1347	3.1	5.1	8.2	5.2	8.4
	VT_20050408-1500	3.7	0.5	4.2	0.5	4.2
	VT_20050425-1000	1.6	7.3	8.9	4.8	6.4
RT09 SDM	EDI_20071128-1000	5.9	2.2	8.1	1.2	7.1
	EDI_20071128-1500	5.1	35.2	40.3	15.9	21.0
	IDI_20090128-1600	0.9	11.3	12.2	7.4	8.3
	IDI_20090129-1000	3.8	10.1	13.9	7.2	11.0
	NIST_20080201-1405	3.6	55.6	59.2	41.2	44.9
	NIST_20080227-1501	1.4	11.2	12.6	7.8	9.2
	NIST_20080307-0955	1.9	27.0	28.9	27.0	28.9

Table 5.2: Details of the DER with and without adding the purification step presented in Section 5.1 for the Evaluation Set: RT'07, and the Validation Set: RT'09 for the SDM conditions. All results are given without scoring the overlapping speech

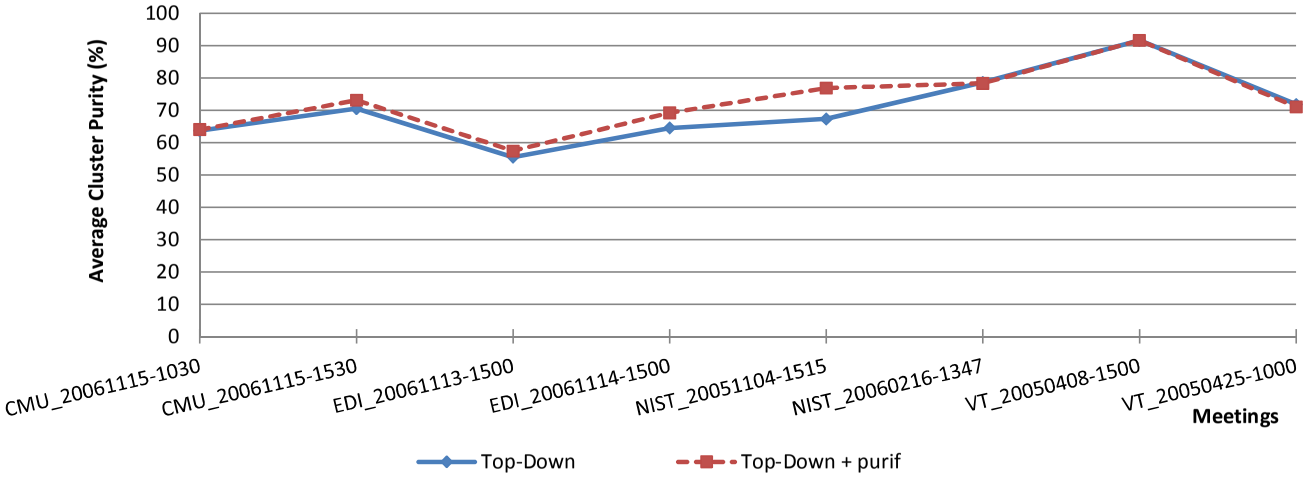
System	**Dev. Set**	**RT'07**	**RT'09**
Top-down Baseline	70.4/42.6/91.2	74.6/60.4/91.5	68.2/47.2/83.9
Top-down Baseline + Pur.	70.5/43.7/91.4	75.6/65.6/91.5	69.7/54.2/84.7

Table 5.3: Cluster purities (%Pur) without (Top-down Baseline) and with (Top-down Baseline + Pur.) purification for the Development Set, the Evaluation Set: RT'07, and the Validation Set: RT'09. Results for SDM condition. Note that compared to the similar Table published in [Bozonnet et al., 2010], results here are given for SDM conditions (vs. Multiple Distant Microphones (MDM) in [Bozonnet et al., 2010])

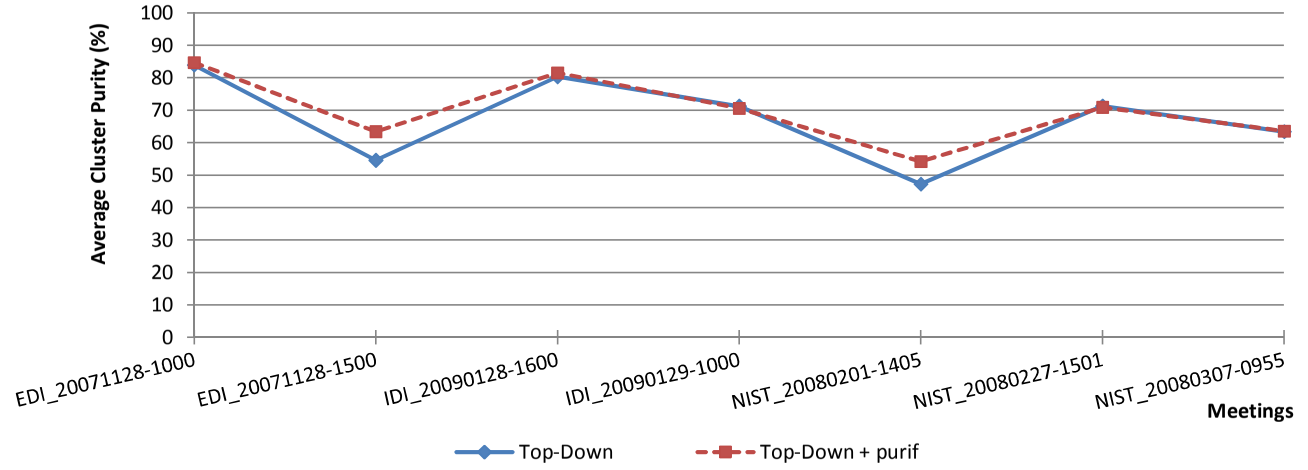
the most suitable and is that used everywhere else in this thesis.

Table 5.3 illustrates the purity for all three datasets both with and without purification. Average/minimum/maximum cluster purities are shown in each case for the three different datasets. Results show that, in all cases, the average cluster purity increases after purification. Of particular note, is the general increase in the minimum cluster purity (with the exception of the Development set), whereas the maximum purity only changes for the RT'09 dataset. Note that the lowest purities before purification (average, minimum and maximum) all correspond to the RT'09 dataset and also that the biggest improvement in minimum purity (54% cf. 47%) is also achieved on the RT'09 dataset. This goes someway to explain the behavior noted above but it is nonetheless of interest to see the improvement in purity across the individual shows.

Figures 5.4a and 5.4b illustrate the %Pur metrics before and after purification (solid and dashed profiles respectively) for each of the 8 files of RT'07 and 7 files in the RT'09 dataset (horizontal axis). For both datasets, we observe that purity is improved or unchanged after the purification component, but never deteriorates. Moreover results show that, where initial models are already of high purity (e.g. the first and third shows in Figure 5.4b), then purification has little effect. However, when initial clusters are of relatively poor purity (e.g. the second or fifth shows in Figure 5.4b) then purification leads to a marked improvement. For these particular shows the cluster purity increases from 55% to 63% with purification (second show) and from 47% to 54% (fifth show). With few exceptions this behavior is typical of that across the other datasets. Since initial cluster purities are particularly bad for the RT'09 dataset (illustrated in Table 5.3), it is thus of no surprise that the effect of purification is greatest here. Even so, we note that other researchers have found that this dataset was more 'difficult' compared to previous datasets and the performance of our new system is also slightly inferior to that on the Development Set and RT'07 set even if the purification system reduces the difference.



(a) NIST RT'07 dataset (SDM condition)



(b) NIST RT'09 dataset (SDM condition)

Table 5.4: (a): %Pur metrics for the NIST RT'07 dataset (SDM condition) before and after purification (solid and dashed profiles respectively); (b): same for NIST RT'09 dataset

The addition of the purification component in the top-down system leads to DER improvements, but are at the expense of a small increase in computational cost. Compared to the top-down system, as described in Section 3.4.1 which achieved a speed factor[1] of 1.5, the purification algorithm introduces a negligible overhead in processing time which increases the speed factor of our new system to 1.6. Compared to the speed factors of other systems published in the proceedings of the NIST RT evaluations our new system is still among the most efficient[2].

5.3 Experimental Work with the Bottom-Up System

Purification of output clusters with the algorithm described in Section 5.1 shows a consistent improvement on the top-down system baseline. In this section we apply the same algorithm as a post processing to the bottom-up system described in Section 3.4.2.2.

5.3.1 Diarization Performance

Similar to Table 5.1, Table 5.5 illustrates a comparison of speaker diarization performance for the SDM condition using the bottom-up system with and without post purification. Results for the same four different datasets (columns 2 to 9) are given with(OV) and without(NOV) the scoring of the overlap speech.

The purification algorithm has almost no effect on the Development Set (0.1 absolute % difference) and leads to a relative improvement of 6% (19.6% cf. 20.8%) over the bottom-up baseline on the RT'07 dataset. However for RT'09 dataset a large deterioration of 61% relative is observed (41% relative deterioration without scoring the overlap speech). Moreover, results on the GE corpus also show a deterioration in performance. Thus, compared to results for the top-down system, the purification algorithm leads to inconsistent improvements on the bottom-up system and can even deteriorate average performance. In order to understand why the algorithm performs significantly worse on the RT'09 dataset than on the RT'07 dataset, we focus in the following on the evolution of the cluster purity.

[1]The submission criteria of the NIST RT evaluations [NIST, 2009] require the reporting of system efficiency in terms of a speed factor which gauges the efficiency of the system in relation to real time.

[2]For the NIST RT'09 evaluation the speed factor for bottom-up approach was at least 4.0

System	Dev. Set		RT'07		RT'09		GE	
	OV	**NOV**	**OV**	**NOV**	**OV**	**NOV**	**OV**	**NOV**
Bottom-up (I2R)	21.7	18.9	23.8	20.8	19.1	13.5	33.7	29.0
Bottom-up+Pur.(I2R)	21.6	18.8	22.7	19.6	27.0	21.8	33.9	29.1

Table 5.5: A comparison of diarization performance on the SDM condition and four different datasets: a development set (23 meetings from RT'04, RT'05, RT'06), an evaluation (RT'07), a validation (RT'09) and a TV show dataset: Grand Échiquier(GE). Results reported for two different systems: the bottom-up baseline (I2R) as described in Section 3.4.2.2 and the same system using cluster purification (Bottom-up+Pur.). Results illustrated with(OV)/without(NOV) scoring overlapping speech.

System	Dev. Set	RT'07	RT'09
Bottom-up(I2R)	72.0/37.5/91.2	70.3/57.5/91.0	68.1/52.8/78.1
Bottom-up(I2R) + Pur.	71.7/37.5/91.3	71.4/58.2/91.9	66.4/36.9/77.3

Table 5.6: cluster purities (%Pur) without (Bottom-up Baseline) and with (Bottom-up Baseline + Pur.) purification for the Development Set, the Evaluation Set: RT'07, and the Validation Set: RT'09. Results for SDM condition.

5.3.2 Cluster Purity

Cluster purity statistics before and after purification are shown in Table 5.6. Average/minimum/maximum cluster purities are given for the same four datasets as in Section 5.2.2. While for the top-down system a consistent purification improvement was observed on each dataset, on the bottom-up system, improvements in terms of cluster purity are only seen on the RT'07 dataset. Indeed, purification deteriorates on the Development set and the RT'09 dataset. When we look at the minimum and maximum cluster purity, we note a small improvement for the development and RT'07 set, but a large deterioration for the minimum cluster purity for the RT'09 set (a decrease from 52.8% to 36.9%). This is consistent with the poor performance in terms of DER observed for the RT'09 dataset in 5.3.1.

5.4 Conclusion

In this chapter we introduced a new purification component which brings some consistent improvements in the top-down system. Purification leads to a new top-down baseline which produces comparable results to the bottom-up approach and delivers improved stability across different datasets composed of conference meetings from five standard NIST evaluations and a TV-show corpus. An average relative DER improvement of 15% can be observed on independent meeting datasets.

However, in contrast to the top-down system, results show that performance can sometimes deteriorate when purification is applied to bottom-up clustering. From these observations we hypothesize that, in practice, the nature of the system outputs is significantly different depending on the type of clustering. This leads us to investigate the two diarization approaches more thoroughly and to study their relative merits. This it the subject of the next chapter.

Chapter 6

Comparative Study

Chapter 5 shows that purification brings some consistent improvements to the top-down system, leading to comparable results to the bottom-up approach with neither system being consistently superior to the other. Results show, however, that performance can sometimes deteriorate when purification is applied to bottom-up strategies. These observations lead us to investigate the two diarization approaches more thoroughly and to study their relative merits.

In this chapter we propose to first present in Section 6.1 an original theoretical framework which we published in [Evans et al., 2012] including a formal definition of the task of speaker diarization and an analysis of the challenges that must be addressed by practical speaker diarization systems. We then report in Section 6.2 a qualitative comparison highlighting the relative merits of top-down and bottom-up clustering approaches in terms of discrimination between individual speakers and normalization of unwanted acoustic variation, i.e. that which does not not pertain to different speakers. Finally Section 6.3 presents an experimental validation of the hypothesized behaviors.

6.1 Theoretical Framework

In this section we propose a theoretical framework for the speaker diarization task. Although it is not the only possible approach, the formulation presented is representative of state-of-the-art technologies based on probabilistic modeling. All the assumptions made in theory development are consistent with modern speaker diarization systems

that have been entered into the official NIST RT evaluations [NIST, 2009], including the two top-down and bottom-up baseline scenarios presented in Chapter B.2.

Based on the probabilistic framework, we analyze the main challenges that must be addressed in related practical systems. This analysis leads naturally to the two principal approaches to speaker diarization, namely the bottom-up and top-down clustering approaches that are studied and compared later in this chapter.

6.1.1 Task Definition

Speaker diarization can be defined as an optimization task on the space of speakers given the audio stream that is under evaluation. We first assume that non-speech segments have been removed from the acoustic stream and that features are extracted such that the remaining speech information is represented by a stream of acoustic features O. Letting S represent a speaker sequence and G a segmentation of the audio stream by S, then the task of speaker diarization can be formally defined as follows:

$$(\tilde{S}, \tilde{G}) = \underset{S,G}{\operatorname{argmax}} P(S, G|O) \tag{6.1}$$

where $\tilde{S}$ and $\tilde{G}$ represent respectively the optimized speaker sequence and segmentation, i.e. who (S) spoke when (G). We can factorize the posterior probability in (6.1) by applying the Bayesian rule:

$$\begin{aligned} (\tilde{S}, \tilde{G}) &= \underset{S,G}{\operatorname{argmax}} \frac{P(S, G)P(O|S, G)}{P(O)} \\ &= \underset{S,G}{\operatorname{argmax}} P(S, G)P(O|S, G) \end{aligned} \tag{6.2}$$

where $P(O)$ is suppressed since it is independent of S and G. Equation (6.2) shows that two models are required in order to solve the optimization task:

- an **acoustic model** which describes the acoustic attributes of each speaker, constituting the likelihood $P(O|S, G)$,
- a **speaker turn model** which describes the probability of a turn between speakers with a given segmentation, constituting the prior $P(S, G)$

Usually the acoustic models are implemented as Gaussian mixture models (GMMs). Letting S_i denote the i-th speaker in S, and O_i its corresponding speech segment according to G, we have the following likelihood:

$$P(O|S,G) = \prod_{\forall \text{ speaker } i} P(O_i|\lambda_{S_i}, G), \tag{6.3}$$

where λ_{S_i} denotes the GMM speaker model for speaker S_i.

By applying various different assumptions one can obtain different forms of the speaker turn model. For example, if we assume that the speaker labels either side of the turn are irrelevant and take only the utterance duration into account then we have the following duration model:

$$P(S,G) = P(G), \tag{6.4}$$

where $P(G)$ can be modeled with a normal or Poisson distribution for example. Alternatively, and as is common in practice, one may assume a uniform distribution and thus omit the turn model entirely. Substituting (6.3) and (6.4) into (6.2) we obtain:

$$(\tilde{S}, \tilde{G}) = \underset{S,G}{\operatorname{argmax}}\, P(G) \prod_i P(O_i|\lambda_{S_i}, G), \tag{6.5}$$

which provides a full solution to the speaker diarization problem.

6.1.2 Challenges

In practice, the implementation of a practical speaker diarization system is rather more complex than may first appear from the basic framework presented above. The first challenge involves the optimization of the speaker sequence S in (6.5). This is not straightforward since the inventory of S is unknown, i.e. we do not know how many speakers N are present within the acoustic stream. This means that it is not possible to optimize the speaker sequence S without a jointly-optimized speaker inventory.

Second, although we suppose that a set of acoustic models can reliably represent the acoustical characteristics of the speakers, the speech signal O is rather complex. Whilst the acoustic models depend fundamentally on the speaker, they also depend on a number of other nuisance factors such as the linguistic content, for example the words or phones pronounced, which are not related specifically to the speaker.

In the following we assume for simplicity that the major nuisance variation relates only to the phone class of uttered speech, which we denote as Q, though other acoustic classes are also valid. Due to its significant effect on the speech signal, Q should appear in the solutions and must be addressed appropriately.

To formulate a solution which addresses these two challenges, we first introduce the speaker inventory Δ, and let $\Gamma(\Delta)$ represent all possible speaker sequences. Returning to equations (6.1) and (6.2) we can derive the solution as follows:

$$\begin{aligned}(\tilde{S}, \tilde{G}, \tilde{\Delta}) &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S, G|O) \\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S, G)P(O|S, G) \end{aligned} \tag{6.6}$$

While marginalizing the likelihood $P(O|S, G)$ over all the possible phone classes Q, we can derive:

$$\begin{aligned}(\tilde{S}, \tilde{G}, \tilde{\Delta}) &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S, G) \sum_{\forall Q} P(O, Q|S, G) \\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S, G) \sum_{\forall Q} P(O|S, G, Q)P(Q|S, G) \\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S, G) \sum_{\forall Q} P(O|S, G, Q)P(Q) \end{aligned} \tag{6.7}$$

where Q is naturally independent of G and we have further assumed it to be independent of the speaker S.

The solution reveals two important issues that any practical speaker diarization system must address. First, the speaker inventory Δ must be optimized together, not only with the speaker sequence S, but also the segmentation G. There is no analytical solution for Δ and so a trial-and-error search is typically conducted. This search can be either from a smaller inventory to a larger inventory, or from a larger inventory to a smaller inventory. These strategies correspond respectively to the top-down and bottom-up approaches to speaker diarization.

Secondly, when comparing (6.6) and (6.7), we see that:

$$P(O|S, G) = \sum_{\forall Q} P(O|S, G, Q)P(Q). \tag{6.8}$$

This means that in the optimization task one should either use a phone-independent model $P(O|S,G)$ and apply (6.6), or a phone-dependent model $P(O|S,G,Q)$ with prior knowledge of $P(Q)$ and apply (6.7). Due to its simplicity and effectiveness, most speaker diarization systems nowadays adopt the former approach. For such a system $P(O|S,G)$ must be trained with speech material containing all possible phones, otherwise Q will be not marginalized. In other words, for a phone-independent system, acoustic speaker models must be *normalized* across phones Q to ensure that the resulting model is phone-independent, otherwise optimization according to (6.6) will be suboptimal.

In summary, a practical diarization system should incorporate an effective search strategy to optimize the speaker inventory Δ, and a set of well-trained speaker models to infer the speaker sequence S and segmentation G. Ideally, the models should be most discriminative for speakers and fully normalized across phones. From this perspective, the direction in which the optimal speaker inventory is searched for (bottom-up or top-down) is inconsequential. Searching from either direction will in any case arrive at the optimal inventory[1].

However, the merging (bottom-up) or splitting (top-down) operations in the search process are likely to impact upon the **discriminative power** and **phone-normalization** of the intermediate and final speaker models. Therefore, the two approaches will exhibit different behaviors and relative strengths and shortcomings in practice.

6.2 Qualitative Comparison

The bottom-up and top-down approaches to speaker diarization are fundamentally opposing strategies. The bottom-up approach is a specific-to-general strategy whereas the top-down approach is general-to-specific. The latter will produce more reliably trained models as relatively more data are available for training. However, the models are likely to be less discriminative until sufficient speakers and their data are liberated to form distinct speaker models. The bottom-up approach, in contrast, is initialized with a larger number of models and is there more likely to discover specific speakers earlier in the process, however the models may be weakly trained until sufficient clusters are merged.

[1]We assume that the number of speakers is known approximately so that the bottom-up approach is initialized with more clusters than true speakers in order to avoid the risk of over-clustering.

The two approaches thus have their own strengths and weaknesses and are therefore likely to exhibit different behavior and results. In the following we discuss some particular characteristics in further detail with the aim of better illuminating their .

6.2.1 Discrimination and Purification

A particular advantage of the bottom-up approach rests in the fact that it is likely to capture comparatively purer models. Whilst they may correspond to a single speaker, they may also correspond to some other acoustic unit, for example a particular phone class. This is particularly true when short-term cepstral-based features are used, though recent work with prosodic features has potential to encourage convergence specifically toward speakers [Friedland et al., 2009]. In contrast, since it initially trains only a small number of models using relatively larger quantities of data, the top-down approach effectively normalizes phone classes, but it also normalizes speakers at the same time. To achieve the best discriminative power *across speakers*, a purification step becomes essential for both approaches: for the bottom-up approach, it is necessary to purify the resulting models of interference from phone variation, whereas for the top-down approach it is necessary to purify the resulting models of data from other speakers. Purifying phones involves phone recognition which is usually rather costly; purifying speakers, however, is much easier with some straightforward assumptions. We have achieved significant improvements in diarization performance using purification in our top-down approach as presented in Section 5.2.

6.2.2 Normalization and Initialization

Theoretically, the EM algorithm ensures that both the bottom-up and top-down approaches will converge to a local maximum of the objective function for a fixed size Δ. If the differences between speakers is the dominant influence in the acoustic space then we can safely assume that the local maximum represents an optimal diarization on speakers, as opposed to any other acoustic class. In this case, initial models are not predominantly important, and thus both bottom-up and top-down approaches will tend to provide similar diarization results. However, in addition to the speaker the acoustic signal bears a significant influence from the linguistic contents, and more specifically the phones. Therefore, the local maximums of the objective function may correspond to

phones Q instead of speakers S if the speaker models are not well normalized, i.e. Q is not fully marginalized. This analysis highlights a major advantage of the top-down approach to speaker diarization: by drawing new speakers from a potentially well-normalized background model, newly introduced speaker models are potentially more reliable than those generated by linear initialization and model merging in the bottom-up approach.

An interesting point derived from the above analysis is that the bottom-up and top-down approaches, which possess distinct properties in terms of model reliability and discrimination, are likely to result in different local maximums of the objective function, suggesting that their combination may thus provide for more reliable diarization. Previous work would seem to support this observation [Meignier et al., 2006]. We report our work on system combination in Chapter 7.

6.3 System Output Analysis

In this Section we present some experimental works which aim to validate the behaviors highlighted in Section 6.2 in terms of speaker discrimination and phone normalization. In that regard, an analysis of the phone distribution and the cluster purity of the system outputs is carried out and accounts for the inconsistencies in system performance outlined above.

6.3.1 Phone Normalization

According to the arguments presented in Section 6.2 bottom-up approaches are relatively more likely than top-down approaches to convergence to sub-optimal local maxima of Equation (6.2). These are likely to correspond to nuisance variation and, whilst other acoustic classes are also relevant, we hypothesize here that the phones uttered are among the most significant competing influences in the acoustic space.

To help confirm this, or otherwise, we measured the difference in the phone distribution between each pair of clusters in the diarization hypothesis. The phone distribution is computed as the fraction of speech time attributed to each phone and thus requires a phone-level reference to determine the phone class of each frame. This was accomplished by a forced alignment of the phone transcription of each word in the reference annotation

Table 6.1: Inter-cluster phone distribution distances.

	Mean		Variance	
System	**RT'07**	**RT'09**	**RT'07**	**RT'09**
Top-down	0.11	0.10	0.006	0.004
Bottom-up (I2R)	0.17	0.14	0.014	0.013
Bottom-up (ICSI)	0.16	0.23	0.005	0.017

to the corresponding speech. The phone distribution of each cluster is used to calculate the average inter-cluster distance D as follows:

$$D = \binom{N}{2}^{-1} \sum_{n=1}^{N} \sum_{m=n+1}^{N} D_{\mathrm{KL2}}(C_n||C_m), \tag{6.9}$$

where N is the size of the speaker inventory Δ, i.e. the number of clusters, and where the binomial coefficient $\binom{N}{2}$ is the number of unique cluster pairs. $D_{\mathrm{KL2}}(C_n||C_m)$ is the symmetrical Kullback-Leibler (KL) distance between the phone distributions for clusters C_n and C_m, defined as:

$$D_{\mathrm{KL2}}(C_n||C_m) = \frac{1}{2}\Big(D_{\mathrm{KL}}(C_n||C_m) + D_{\mathrm{KL}}(C_m||C_n)\Big) \tag{6.10}$$

where $D_{\mathrm{KL}}(C_n||C_m)$ is the KL divergence of C_n from C_m. We note that the symmetrical KL metric has been used for the segmentation and clustering of broadcast news [Siegler et al., 1997].

In the case where clusters are well normalized against phone variation then the average inter-cluster distance is expected to be small, since the clusters should have similar phone distributions. Significant differences between distributions, however, indicate poor phone normalization and possibly a sub-optimal local maximum of (6.2). This latter case might reflect a higher degree of convergence toward phones, or other acoustic classes, rather than toward speakers.

The mean and the variance of the inter-cluster distances are presented in columns 2 and 3 of Table 6.1 for the RT'07 and RT'09 datasets respectively. For the baseline bottom-up system average inter-cluster distances of 0.17 and 0.14 are obtained. These fall to 0.13 and 0.12 with purification indicating improved normalization against phones. For the top-down system the average distances are 0.11 and 0.10. These fall to 0.07

Table 6.2: Average cluster purity and number of clusters.

	Cluster Purity (%)		No. Clusters	
System	**RT'07**	**RT'09**	**RT'07**	**RT'09**
Top-down	74.6	68.2	5.1	6.1
Top-down + Pur.	75.6	69.7	4.8	5.3
Bottom-up(I2R)	70.3	68.1	6.8	6.9
Bottom-up(I2R) + Pur.	71.4	66.4	5.8	6.9
Ground-truth	100.0	100.0	4.4	5.4

and 0.08 with purification and are significantly better than for the bottom-up system. Reassuringly, with combination the values remain stable at 0.07 and 0.07. Columns 4 and 5 of Table 6.1 show the corresponding variances in all cases and show a consistent decrease moving down the table: reductions in the mean are accompanied by reductions in the variation. These observations suggests that on average, and as predicted, the clusters identified with the bottom-up system are indeed less well normalized against phone variation than those identified with the top-down system and that combination preserves the normalization of the top-down system.

6.3.2 Cluster Purity

The observations reported above do not explain why, for the RT'09 dataset, the bottom-up system performance deteriorates with purification even though the phone normalization improves. To help explain this behavior we analyzed the average speaker purity in each system output. The cluster purity is the percentage of data in each cluster which are attributed to the most dominant speaker, as determined from the ground-truth reference. Average cluster purities are presented in columns 2 and 3 of Table 6.2. For the RT'07 dataset purification leads to marginal improvements: from 70.3% purity to 71.4% for the bottom-up system and from 74.6% to 75.6% for the top-down system. Different behavior is observed for the RT'09 dataset. Whereas purification gives an improvement from 68.2% to 69.7% for the top-down system it leads to a degradation from 68.1% to 66.4% for the bottom-up system.

Whilst a reduction in cluster purity may account for the decrease in diarization performance it is necessary to consider the number of clusters in the system output to properly interpret cluster purity and its impact on diarization performance. As explained

in Section 6.3.1 purification influences the number of identified clusters. A larger number of clusters may be associated with inherently higher purity (i.e. with a single cluster for each sample the purity is 100%) and so purity statistics alone do not fully reflect the effect of purification on diarization performance. The number of clusters detected in each system output is illustrated in columns 4 and 5 of Table 6.2 in which the last row shows the statistics for the ground-truth reference. All systems over-estimate the number of speakers and purification always reduces the number toward the number of true speakers. When coupled with increases in average purity, then improved diarization performance should be expected. For the bottom-up system and the RT'09 dataset there is no decrease in the number of clusters when purification is applied, whereas the purity also decreases. This can only result in poorer diarization performance.

6.4 Conclusion

Through a new theoretical framework, this chapter shows that top-down and bottom-up clusterings should theoretically be inconsequential on the speaker inventory and then should lead to the same optimal inventory. However, while ideally the models should be most discriminative for speakers and fully normalized across phones, the merging and splitting operations in the search process are likely to impact upon the discriminative power and phone-normalization of the intermediate and final speaker models, leading in practice to different behaviors and relative strengths and shortcomings. Indeed, our study shows that top-down systems are often better normalized toward phonemes and then more stable, but that they suffer from low speaker discrimination which explains that they are likely to benefit from purification. In contrast, bottom-up clusterings are more speaker discriminative, but as a consequence of their progressive merging scenario, they may be sensitive to phoneme variations which might lead the system to non-optimal, local maxima.

The distinct properties in terms of model reliability and discrimination of these two approaches suggest that there is some potential for system combination. The next chapter investigates this hypothesis and reports two possible approaches to combine top-down/bottom-up systems.

Chapter 7

System Combination

System combination is a popular and sometimes straightforward means of improving performance in many fields of statistical pattern classification, including speech and speaker recognition where combination or fusion strategies have led to significant leaps in performance e.g.[Burget et al., 2009]. However, due to its unsupervised nature, the combination or fusion of diarization systems is somehow troublesome. In fact, the variability of the number of detected speakers and the fact that systems are not standardized in terms of labeling, i.e. there is no natural correspondence between system output labels, make the task very challenging.

However, as outlined in Chapter 6, bottom-up and top-down clustering strategies have different weaknesses and are likely to behave differently toward phoneme effects, leading to some complementary diarization outputs. For these reasons we can expect to get some improvements in performance while combining or merging these two systems. The following work was published in [Bozonnet et al., 2010; Evans et al., 2012],[Bozonnet et al., 2010] and is organized as follows.

In Section 7.1 we present the possible strategies to combine or fuse two diarization systems. In Section 7.2 we introduce an integrated Top-Down Bottom-Up system, while in Section 7.3 a combination of the Top-Down and Bottom-Up system outputs is proposed.

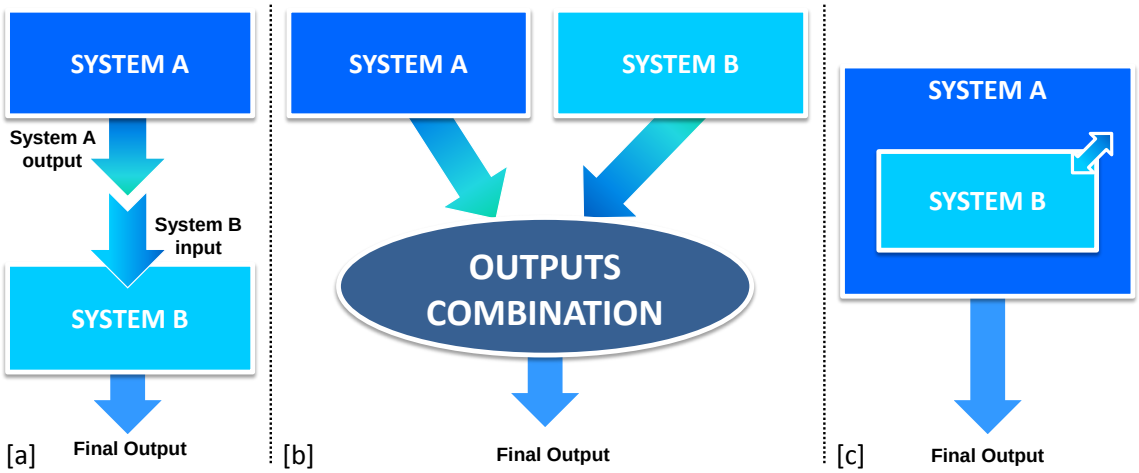


Figure 7.1: Three different scenarios for system combination: Piped System (a), Fused System (b) and Integrated System (c)

7.1 General Techniques for Diarization System Combination

System combination[1] is a popular way to harness the strengths of each system and thus to improve performance and stability. According to the work published in [Meignier et al., 2006] we propose to differentiate three ways to combine the system: they are the piped system, so-called hybridization strategy, the fused system (or merging strategy) and the integrated system as illustrated on Figure 7.1

7.1.1 Piped System - Hybridization Strategy

The piped system, or so-called hybridization strategy, as shown in Figure 7.1(a), involves the output of one system being used to initialize a second system. This scenario is certainly the easiest to implement but it may be sensitive to weaknesses of the first system applied since errors introduced first cannot be corrected by the second system. This strategy was used in [Meignier et al., 2006] where the output of a bottom-up system is applied to the input of a top-down system.

[1] Note that for clarity and consistency we keep the terminology 'System Fusion' for the 'Fused System' only while we designate by 'System Combination' the three techniques: 'Piped', 'Fused' and 'Integrated' Systems

7.1.2 Merging Strategy - Fused System

While the piped system aims to run iteratively one system in order to feed the second, the fused system first runs simultaneously and independently the two systems (Figure 7.1(b)) and then combines the outputs. The method may be more robust than the hybridization strategy in the case that one of the two systems gives poor performance.

This scenario is quite popular and can be used at the frame level, e.g. in [Meignier et al., 2006] labels are first merged and a resegmentation is made, but the process can also operate at the cluster level. In [Gupta et al., 2007], for example, the most relevant, common clusters of two system outputs are first identified. Then all segments which are not identified as belonging to the common clusters are labeled as misclassified. They are next reassigned through a new realignment based on the GMM models issued from the common clusters and a maximum likelihood based decision. Still operating at the cluster level [Tranter, 2005] proposed a cluster voting approach to combine the outputs of two different speaker diarization systems while [Huijbregts et al., 2009] perform a fusion at the show level and propose a segmentation voting approach in order to elect the best segmentation of each show.

7.1.3 Integrated System

Finally the integrated approach[1] aims to fuse the two systems at their heart (Figure 7.1(c)). The systems are not run sequentially as for the piped system, neither independently like for the merging strategy but simultaneously, one system calling the other as a subroutine during its execution. Due to difficulties in implementing such a system, only few works involve truly integrated approaches. They include [Vijayasenan et al., 2008] where one system based on an agglomerative Information Bottleneck (aIB) approach is combined with a sequential Information Bottleneck (sIB) approach or [El-Khoury et al., 2008] where two different hierarchical clustering systems are coupled.

[1]Note that in [Meignier et al., 2006] the Top-Down approach is described as Integrated due to the fact that it is based on an Evolutive Hidden Markov Modeling (E-HMM) where the number of speakers, their models and the segmentation are re-evaluated together at each step even if this system is not really comparable to the real integration of two different systems.

Among all the works reported in the literature, none of them involved an integrated system based on bottom-up and top-down cutting edge diarization systems. Moreover, the existing approaches for system fusion at the cluster level involve diarization systems of the same nature. In the following we investigate these two different approaches to combine the baseline systems presented in Section 3.4.

7.2 Integrated Bottom-up/Top-down System to Speaker Diarization

A way to take the benefit of each of the different system is to combine them at the heart of the segmentation and clustering stage, in an integrated approach. We propose a new system whose skeleton is based upon the LIA-EURECOM top-down system, described in Section 3.4.1, but where each speaker model is trained by following an integrated bottom-up approach with sequential EM training, as used in the I2R system [Nguyen et al., 2009] presented in Section 3.4.2.2.

7.2.1 System Description

As detailed in Section 3.4.1, and as illustrated in Figure 7.2, the first step involves the learning of a general model $S0$ which is tuned by EM using all the available speech segments. Then initialization with sequential EM as described in [Nguyen et al., 2009] is applied using all of the speech data assigned to model $S0$. However instead of splitting the data uniformly into 30 clusters as presented in [Nguyen et al., 2009], the speech segments assigned to model $S0$ are divided linearly into 30-second sub-clusters (3 in Figure 7.2 labeled A, B, and C). Our experiments show that this approach gives better results. Then the steps described in [Nguyen et al., 2009] are performed 10 times on the resulting sub-clusters: 25% of the data which best fits the corresponding model are considered as classified whereas other data are unlabeled. The models are updated using only the classified data and a decoding is performed where only a fraction of the newly classified data are reassigned to their nearest sub-clusters. Several steps of Viterbi realignment and adaptation are performed until all the data are classified. As illustrated in Figure 7.2 the sub-cluster which is assigned the greatest amount of speech data is used to introduce a new speaker $S1$ into the E-HMM system. The data in all

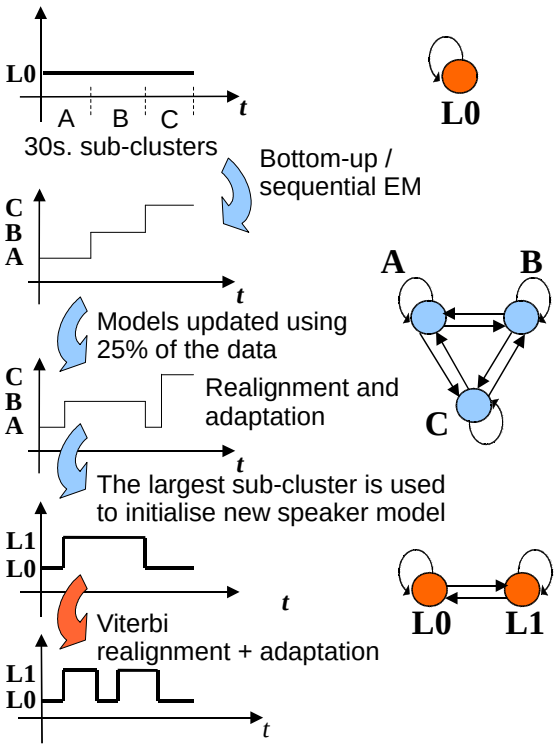


Figure 7.2: The integrated approach

other sub-clusters are assigned back to $S0$. Several iterations of Viterbi decoding and adaptation are performed with the E-HMM until the system is stable.

This process is repeated in exactly the same way to add additional speakers to the E-HMM until there is no longer sufficient data assigned to $S0$ with which to create a new speaker model. Thus in this approach we harness the better initialization provided by the bottom-up approach to initialize each new speaker model in the top-down approach.

7.2.2 Performance

Figure 7.3 shows the cluster purity of a collection of candidate clusters obtained by sequential EM training according to their size for RT'07 and RT'09 datasets. This chart clearly illustrates that the higher the amount of frames in the resulting cluster, the more chance we have to select a cluster with high purity. This behavior justifies the choice of the candidate cluster with the greatest amount of speech data to be introduced as a new speaker into the E-HMM as described above.

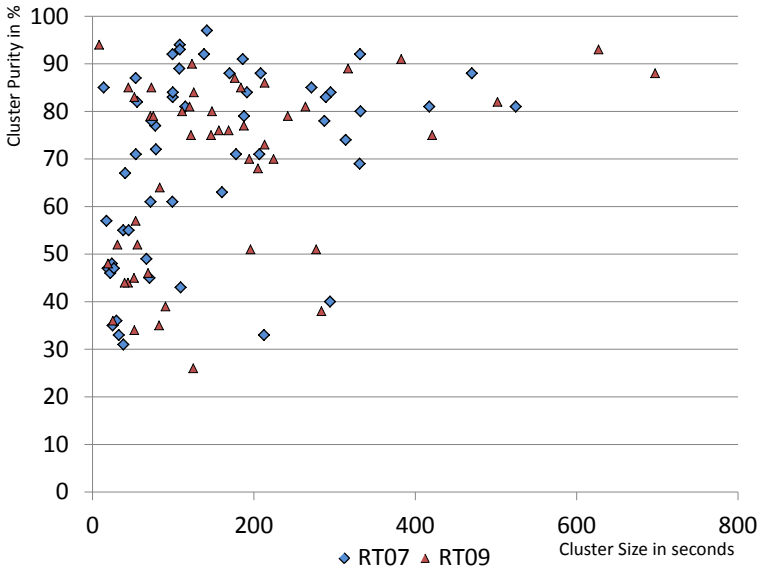


Figure 7.3: Purity rate of the clusters according to their size (seconds)

System	Dev. Set	RT07	RT09	GE
Top-down+Pur.	21.1/18.3	17.8/14.4	21.1/16.0	38.5/33.9
Bottom-Up (I2R)	21.7/18.9	23.8/20.8	19.1/13.5	33.7/29.0
Integrated System	17.3/14.3	16.5/13.0	23.8/18.6	30.9/26.3
Integrated System+Pur.	16.2/13.2	16.4/12.9	23.5/18.2	28.4/23.2

Table 7.1: % Speaker diarization performance in terms of DER with/without scoring the overlapped speech. Results illustrated without and with (+Pur.) purification for the Dev. Set and the RT‘07, RT‘09 and GE datasets.

Results for 4 different datasets including the TV-talk show dataset Grand Échiquier as introduced in Subsection 3.3.2, are presented in Table 7.1 where the DER is given with/without the scoring of overlapping speech. Since none of the systems assessed provide a means of detecting or labeling overlapping speech, we refer in the text to scores where overlapping speech is ignored. The first line of Table 7.1 presents the result with our top-down baseline system as described in Section 3.4.1 and the purification component of Section 5.1.

Upon comparison of results for the baseline system (row 2) and I2R bottom-up system (row 3), we see that the top-down system gives similar results to the bottom-up system for the development set (18.3% vs. 18.9%). For the RT'07 dataset the top-down system gives the best performance (14.4% vs. 20.8%) while for the RT'09 and GE datasets, the bottom-up system gives the best performance (13.5% vs. 16.0% and 29.0% vs. 33.9%).

Finally rows 4 and 5 of Table 7.1 show results for the new integrated system described in Subsection 7.2.1, with and without purification respectively. Referring first to results without purification and their comparison to results for the baseline system (2nd row), we observe largely consistent improvements in performance. Relative improvements of 22%, 10% and 22% are obtained for the development, RT'07 and GE datasets respectively. For the RT'09 dataset, however, performance is worse with the integrated approach (18.6% vs. 16.0%). Whilst this is disappointing we note that the RT'09 dataset has a particularly high degree of overlapping speech and very short speech segments. Other researchers have also reported difficulties with this particular dataset[1].We also note that the decrease in performance is concentrated on only two shows whereas for other shows performance improves.

Note that with added purification, small improvements in performance are obtained for the development and GE datasets (8% and 12% relative improvements respectively).

7.2.3 Stability

The box plots in Figure 7.4 depict performance and stability for each of the 3 systems: the baseline top-down system with purification, I2R's bottom-up system and the new integrated system. All plots illustrate the spread in performance across an entire dataset, first for meeting data and second for the TV-show data. The rectangular boxes show

[1]As related during NIST RT'09 workshop in Melbourne and illustrated in http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-SPKR-v3.pdf

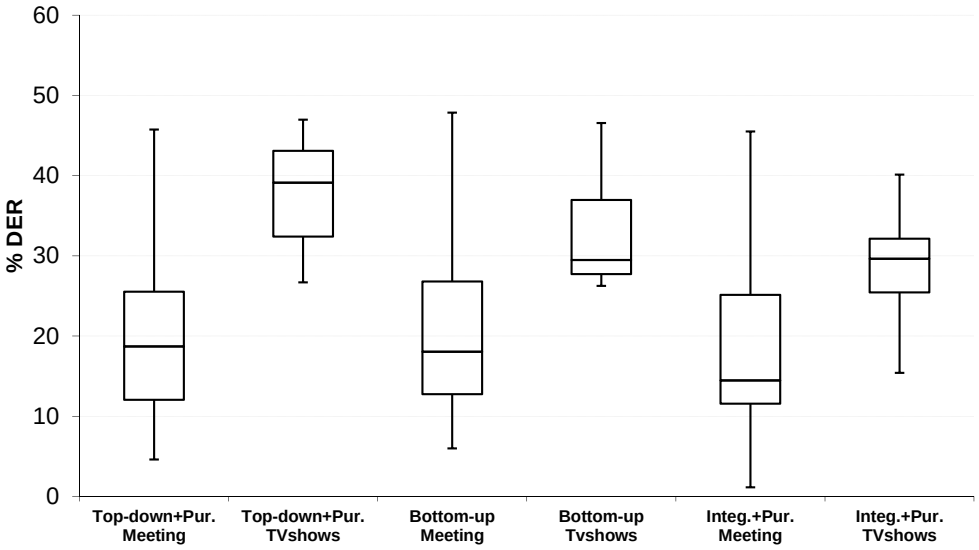


Figure 7.4: Box plot of the variation in DER for the three systems on 2 domains: meeting (averaged across the Dev. Set, RT'07 and RT'09 datasets) and TV-show (GE dataset). Systems are (left-to-right): the top-down baseline system with purification, I2R's bottom-up system and the integrated system with purification.

the inter-quartile range (IQR) and illustrate the intra-domain stability, while the middle line indicates the median performance. The comparison of any corresponding pair of box plots (one for meeting data, one for TV-show data) serves to illustrate the inter-domain stability.

The first two box plots illustrate performance for the baseline top-down system with purification, first for meetings and then for TV-show data. We observe that performance differs greatly between the two datasets. The third and fourth box plots illustrate comparative performance for the bottom-up system for meeting data and then TV-show data. In general there is a greater spread in performance for the bottom-up system than there is for the top-down system. This variation is partially accounted for by poor merging/stopping performance. For meetings the median performance of the bottom-up system is the same as for the baseline whereas for the TV-show data the bottom-up system achieves significantly better performance.

The last two box plots show performance for the new integrated system. Compared to the baseline the spread in performance with meeting data is unchanged whereas the median decreases noticeably. There is thus an overall improvement in performance,

however, whilst the best score also decreases, the worst score remains unchanged. The largest improvement is achieved for the TV-show data for which significant decreases in both the IQR and median performance are observed. We also notice that the difference between the box plots for meeting and TV-show data is less for the integrated system than it is for any other system. Thus the inter-domain stability is greatly improved with the new integrated approach.

7.3 Fused System to Speaker Diarization

In this section we combine into a fused system LIA-EURECOM's top-down approach with purification as described in Chapter B.2 and published in [Bozonnet et al., 2010] with a state-of-the-art bottom-up speaker diarization system. According to the last NIST evaluations [NIST, 2007, 2009], we can consider ICSI's bottom-up system [Wooters & Huijbregts, 2008] and I2R's system [Nguyen et al., 2009] as two state-of-the-art bottom-up approaches. It is important to remember that to achieve the fusion each system needs first to be run independently, then the outputs can be combined as illustrated in Figure 7.1(b).

All the following related to the fusion of LIA-EURECOM and ICSI systems is the result of collaborative work involving LIA, ICSI, Telefonica and EURECOM as published in [Bozonnet et al., 2010]. This collaboration allows us to work with ICSI's official system outputs[1]. In contrast, for the fusion of LIA-EURECOM and I2R systems, I2R's system outputs are issued from our own implementation of their system published in [Nguyen et al., 2009], using LIA-EURECOM's SAD technology and so cannot be compared directly to I2R's official outputs.

According to the results published in the most recent evaluation [NIST, 2009] we may expect the LIA-EURECOM/I2R combination to lead to better performance than the LIA-EURECOM/ICSI combination. However, if ICSI's outputs can be characterized as 'entirely independent' to LIA-EURECOM's outputs, our implementation of I2R's system can be characterized as 'less independent' due to the uses of similar technologies and/or configurations to the top-down system e.g. initialization of the EM algorithm, length of

[1]It should be noted that, in order to combine the systems, some of ICSI's standard optimizations had to be turned off for different technical reasons, i.e. here ICSI's system did not include a prosodic feature stream [Friedland et al., 2009] and no adaptive initialization [Imseng & Friedland, 2009].

the Viterbi buffer, model for MAP adaptation. In the following we hypothesize that our implementation of I2R's system, due to its different clustering nature is 'independent enough' in order to bring some complementary information to the top-down system.

Despite the use of some cutting edge systems, their outputs can still contain some errors or impurities such as some inaccurate segmentations or some duplicate clusters, for example due to a high intra-speaker variation. For this reason, we hypothesize that some speaker models may reliably represent specific, individual speakers, whereas others may be relatively unreliable. Key to the scenario is the identification of reliable models so that better diarization performance may be achieved by re-clustering the data assigned to the unreliable models.

Since the systems considered are run independently in a totally unsupervised way and since they are based on different technologies, we can expect them to give some significant variations in diarization performance. Indeed, differences in Speech Activity Detection (SAD) outputs and further down-stream dependent processes, such as speaker modeling and more general differences in the particular approach to speaker diarization, will all contribute to differences in the number of speaker boundaries, or turns, and different turn locations. However, while the independence of the systems can be considered as an advantage in terms of complementary information, several issues have to be solved to permit system fusion.

On the one hand, different segmentation outputs are generally not time-synchronized. This is particularly true if different SAD algorithms are used[1]. In this case, whilst one system might produce a speaker label, another may classify it as non-speech. On the other hand, no mapping is possible in terms of labeling and moreover the number of speakers detected may differ from one system to another. A preliminary matching algorithm is therefore necessary to identify speaker label pairs between two segmentation hypotheses.

In order to first highlight the potential for improved speaker diarization performance through system combination, we present in Section 7.3.1 a comparison of each system output on the RT'07 and RT'09 datasets. Then, to demonstrate the capacities while unifying and combining two systems we introduce in Section 7.3.2 an artificial experiment which aims to show the optimal reachable performance. More technical details about the

[1]Note that this is not true for our implementation of I2R's system which shares the same SAD algorithm than LIA-EURECOM's system

Source	Av. no. spkrs RT'07	Av. no. spkrs RT'09	Av. Err RT'07s	Av. Err RT'09s
Ground Truth	4.37	5.42	-	-
ICSI	6.62	5.28	2.25	1.86/1.33
I2R	6.75	7.29	2.88	3.29/3.33
LIA-EURECOM	4.75	5.28	0.87	1.28/0.66
Combined LIA-EUR/ICSI	4.62	5.28	0.65	1.28/0.66
Combined LIA-EUR/I2R	4.38	4.57	0.75	1.14/0.50

Table 7.2: Average number of speakers and average error for the ground-truth reference, the three individual systems and their combination, for RT'07 and RT'09 datasets. Results in column 5 illustrated with/without the inclusion of the *NIST_20080307-0955* show which is an outlier.

practical combination are introduced in Section 7.3.3 in order to obtain the performance of the different systems in Section 7.3.4.

7.3.1 System Output Comparison

In order to characterize the differences in the outputs generated by the three systems we propose to focus on two main features: the number of speakers which can vary a lot according to the system, the process being totally unsupervised and the segment sizes which may reveal the sensitivity of the system to detect short speaker turns.

7.3.1.1 Number of Speakers

Reliably estimating the number of speakers is both extremely challenging and crucial to the overall performance of any diarization system. In order to successfully combine the outputs of the different systems we first compared their clustering characteristics with respect to the number of detected speakers. Table 7.2 shows the number of speakers per show, averaged across the full RT'07 and RT'09 datasets in columns 2 and 3 respectively, for the ground-truth reference (row 1) and the segmentation hypotheses obtained from the ICSI, I2R and LIA-EURECOM systems (rows 2, 3 and 4 respectively).

In addition, shown in columns 4 and 5 of Table 7.2, is the error in the number of speakers detected by each system, also averaged across the full datasets. This is computed by averaging the absolute value of the difference between the real number of

Source	No. segments RT'07	No. segments RT'09	Av. seg. length (s) RT'07	Av. seg. length (s) RT'09
Ground Truth	676	882	2.0	1.8
ICSI	617	694	2.2	2.2
I2R	315	310	4.4	5.0
LIA-EURECOM	307	313	4.5	6.3
Combined LIA-EUR/ICSI	353	315	3.9	6.2
Combined LIA-EUR/I2R	355	314	3.9	4.9

Table 7.3: Average number of segments and average segment length in seconds for the ground-truth reference, each individual system and their combination for the RT'07 and RT'09 datasets.

speakers (i.e. that in the reference) and the number hypothesized by each system for each meeting.

For the RT'07 dataset all systems are shown to under-cluster, i.e. they produce more than a single cluster per speaker (results of 6.62, 6.75 and 4.75 speakers cf. 4.37). For the RT'09 dataset, however, LIA-EURECOM's and ICSI's systems over-cluster, i.e. some clusters correspond to more than a single speaker (results of 5.28 for both systems cf. 5.42), while I2R's system under-clusters (result of 7.29 vs 5.42). In both cases, the average error is lower for the LIA-EURECOM system than for any of the two bottom-up systems.

While combining two systems which under-cluster (e.g. LIA-EURECOM/I2R for RT'07) the robust matching of clusters identified by the two systems may give improved performance when their outputs are combined. Where both combined systems over-cluster (e.g. LIA-EURECOM/ICSI for RT'09) improvements may only be obtained if the clusters in each system which correspond to more than a single speaker do not overlap, i.e. we can find clusters in one system output that do not correspond to clusters in the other system output and hence introduce 'new' clusters into the combined output. This is likely to be more difficult.

7.3.1.2 Segment Sizes

Table 7.3 shows the average number of segments and segment length in seconds, for the ground-truth data (row 1) and for each system output (rows 2, 3 and 4). The number of

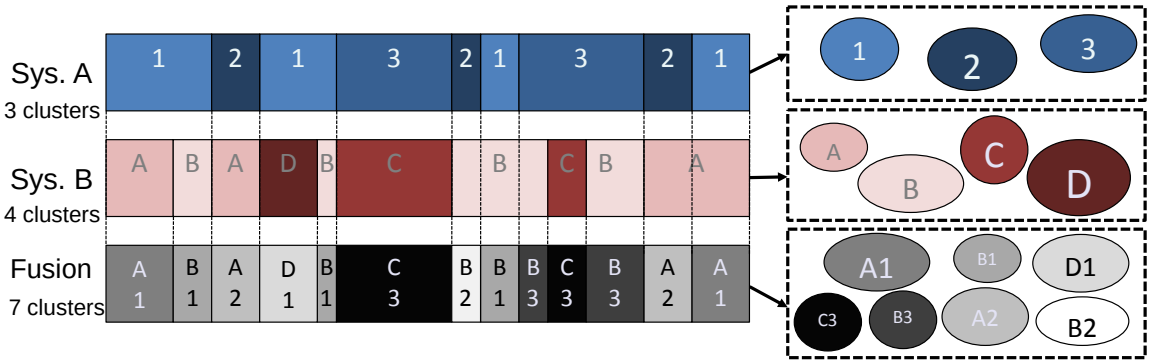


Figure 7.5: Artificial Experiment for Output Combination: System A with 3 clusters is fused artificially with System B containing 4 clusters to create 7 virtual clusters.

segments and their average length are comparable for LIA system and I2R system. This can be explained by the use of the same Viterbi decoding technology (both systems are implemented with ALIZE-MISTRAL library [Bonastre et al., 2005] and share the same Viterbi decoding. The ICSI system estimates the number of segments more reliably than LIA and I2R systems (617 and 307, 315 cf. 676). Similar results are obtained for the RT'09 dataset. The ICSI system also better reflects the average segment length (2.2s and 4.5s, 4.4s cf. 2.0s) and once again similar results are obtained for the RT'09 dataset.

When we focus on the differences between LIA-EURECOM and I2R system outputs , we observe that, despite their comparable average segment lengths, I2R system always under-clusters and provides a number of speakers higher in average than the LIA-EURECOM system. For this reason we may expect there is some potential for a robust cluster matching issued from the two systems.

While comparing LIA-EURECOM and ICSI system outputs, we note that, whilst one system better estimates the true number of speakers with a smaller average error, the other system better reflects the true number of segments and their average length. Should it be possible to exploit the beneficial characteristics of each system then this observation supports the hypothesis that a combined system has the potential to deliver better results.

7.3.2 Artificial Experiment

In order to estimate the optimal capacities of the possible combinations we propose to design the following oracle experiment. This particular work was made by ICSI in the

context of the collaborative work [Bozonnet et al., 2010]. The principle is to combine the outputs in an optimal manner using the ground-truth reference. With this end in view, segment boundaries (i.e. speaker turns) from both systems are merged and virtual clusters are defined by taking the product space of the clusters for each of the two systems. For example, if, for a given segment, $system1$ outputs label c_{1i} and $system2$ outputs label c_{2j}, then we attribute the virtual cluster assignment $c^V_{(i,j)}$. Thus, the resulting cardinality for our virtual cluster space becomes N_1 N_2 where N_i refers to the total number of clusters output by system i. Figure 7.5 shows the creation of 7 virtual clusters resulting from a System A made of 3 clusters and a System B with 4 clusters. Note that in this case, all the cluster pairs are not possible leading to a number of virtual clusters (7) smaller than the resulting cardinality of the virtual cluster space (21). The virtual clusters are then merged in an optimal manner in order to minimize the DER, without violating cluster groupings nor changing the segment boundaries. This is achieved with a dynamic programming search making use of the ground-truth data to find optimal many to one mappings.

Results are shown in Table 7.4 for the RT'07 corpus and respectively in Table 7.5 for the RT'09 corpus. For the optimal LIA-EURECOM/ICSI fused system, the overall DER reveals an improvement of 42% relative compared to the best system for RT'07, respectively 25% for RT'09. For the LIA-EURECOM/I2R system we note an improvement on the overall DER of 35% relative for RT'07, respectively 29% relative for RT'09. Thus there appears to be less scope for improvement on the RT'09 corpus, but it still shows a significant potential for fusion. In fact, according to the different shows in the corpus, the possible optimal improvement can reach up to 50% relative improvement for the RT'07 corpus, e.g.*CMU_20061115-1530*, using the LIA-EURECOM/ICSI combination, or 48% relative improvement for RT'09 using the LIA-EURECOM/I2R combination, e.g. *IDI_20090128-1600*.

RT07	ICSI	I2R	LIA-EU	Optimal LIA-EU/ICSI	Optimal LIA-EU/I2R	Combined LIA-EU/ICSI	Combined LIA-EU/I2R
CMU_20061115-1030	36.08	40.26	21.88	16.82	17.05	21.62	21.62
CMU_20061115-1530	19.65	20.83	35.15	9.65	10.20	19.87	19.42
EDI_20061113-1500	32.39	29.06	20.30	16.51	15.57	19.14	**21.31**
EDI_20061114-1500	22.73	34.08	29.96	12.72	19.78	**28.85**	24.33
NIST_20051104-1515	7.56	10.82	10.88	6.76	5.41	**11.09**	11.09
NIST_20060216-1347	9.34	11.03	9.72	6.81	9.64	10.31	9.96
VT_20050408-1500	16.92	26.79	4.60	4.26	4.49	4.53	5.01
VT_20050425-1000	27.31	18.04	11.34	9.14	10.56	9.84	**17.96**
Average	21.30	23.82	17.72	10.23	11.47	15.48	16.11

Table 7.4: Speaker diarization performance in DER for the RT‘07 dataset. Results illustrated for the three individual systems, and optimally (with reference) and practically combined (without reference) systems. All scores are given while scoring the overlapped speech

RT09	ICSI	I2R	LIA-EU	Optimal LIA-EU/ICSI	Optimal LIA-EU/I2R	Combined LIA-EU/ICSI	Combined LIA-EU/I2R
EDI_20071128-1000	20.34	14.65	10.00	9.38	9.85	10.01	9.86
EDI_20071128-1500	18.12	30.53	25.24	15.56	16.62	16.63	19.34
IDI_20090128-1600	18.94	8.84	11.64	6.03	6.49	10.40	6.75
IDI_20090129-1000	23.69	16.29	15.29	13.15	11.16	**17.49**	15.48
NIST_20080227-1501	45.09	16.24	17.69	13.46	13.03	**18.31**	**18.66**
NIST_20080307-0955	47.11	11.72	31.85	21.58	10.35	31.59	**17.38**
NIST_20080201-1405	65.79	51.12	51.66	45.06	38.47	46.89	**55.32**
Average	31.15	19.13	21.06	15.70	13.47	19.61	17.83

Table 7.5: As for Table 7.4 except for the RT‘09 dataset

7.3.3 Practical System Combination

For practical system combination without the ground-truth we performed cluster alignment using a cluster confusion matrix obtained from the output of both systems. The elements of the matrix contain the total speech time assigned to speaker C_i in the top-down system and speaker C_n in the bottom-up system.

Then, for each cluster in the top-down system C_i a candidate cluster contained in the bottom-up system C_n is chosen as a matching cluster if:

- they share a sufficient proportion of frames and the candidate cluster is that with the highest value in that column of the confusion matrix.
- among all other clusters contained in the bottom-up system C_n is the closest to C_i, where the inter-cluster distance is measured in terms of the Information Change Rate (ICR) [Han et al., 2008].

Each matched cluster pair is accepted as a reliable speaker and is retrained with only those frames that are common to both C_i and C_n. All frames which have mismatching labels are rejected during this stage. This set of reliable, matching clusters is denoted Ξ.

Note that in some cases the cluster pairing with the highest ICR is not the same as the pairing with the highest value in the confusion matrix and thus some clusters in the outputs of each system are not aligned through this process.

Having obtained an initial set of reliable clusters Ξ the following step uses one of two different alternatives depending on which bottom-up system we combine with the top-down in order to introduce the forgotten speakers.

- **LIA-EURECOM/ICSI combination:** As published in [Bozonnet et al., 2010] for each cluster C_i in the top-down system which does not have a paired cluster in the bottom-up system we retrain only a percentage $\alpha = 20\%$ of frames which best match the cluster C_i, according to those which have the highest likelihood. α is the only parameter which requires optimization.
- **LIA-EURECOM/I2R combination:** In contrast to the previous variant, as published in [Evans et al., 2012], the outputs of *both* the bottom-up and top-down systems are utilized in order to select frames for re-estimating new speaker models

in the case of unmatched clusters. This can be explained since the I2R system always under-clusters the data and is the system which outputs the biggest averaged amount of clusters (see Table 7.2).

All unreliable, or unmatched clusters are then compared to Ξ in order to identify additional reliable clusters, as follows:

$$\Xi \leftarrow C_m \tag{7.1}$$

if

$$\ell(C_m, \Xi) = \max_k \ell(C_k, \Xi) \quad C_k \notin \Xi \tag{7.2}$$

and

$$\ell(C_m, \Xi) > \theta \tag{7.3}$$

where θ is a tunable threshold determined empirically, and where ℓ is the minimum ICR distance defined by:

$$\ell(C_k, \Xi) = \min_t ICR(C_k, C_t) \quad C_k \notin \Xi, C_t \in \Xi. \tag{7.4}$$

Additionally there is no significant overlap between C_m and any of the clusters in set Ξ. This procedure is conducted iteratively until no further reliable clusters remain. For each new added cluster, the $\alpha = 50\%$ best-fitting frames (according to likelihood) are used to re-estimate a new speaker model.

Further purification is achieved by training models using only the best fitting data and thus better speaker diarization performance is expected.

Note that for each variant, the value of α is first optimized on the RT'07 dataset and then evaluated using the RT'09 dataset. Then the roles of the development and testing sets are inverted and α is optimized again. Experiments show that the optimized value of $\alpha\%$ can differ significantly from one dataset to the other but the resulting DER was in any case observed to be quite stable with α in the range of 20 to 60%. (variations in term of DER are lower than 0.5% absolute)

Finally, in all cases, the new hypothesis is then used to perform a finale resegmentation and Normalization as detailed in Section 3.4.1.

7.3.4 Experimental Work

	RT‘07		RT‘09	
System	**OV**	**NOV**	**OV**	**NOV**
Bottom-up (ICSI)	21.3	17.9	31.2	26.5
Bottom-up (I2R)	23.8	20.8	19.1	13.5
Top-down+Pur.	17.8	14.4	21.1	16.0
Combined LIA-EUR/ICSI	15.5	12.1	19.6	14.6
Combined LIA-EUR/I2R	16.1	12.8	17.8	12.3

Table 7.6: DERs with (OV) and without (NOV) the scoring of overlapping speech for bottom-up, top-down and combined systems with and without purification (Pur.).

The combination algorithm described above was each time optimized on the RT‘07 dataset and then applied to the RT‘09 dataset without modification. Results are illustrated in columns 7-8 of Tables 7.4 and 7.5 for each dataset and each combination: LIA-EURECOM/ICSI, LIA-EURECOM/I2R. In all but two cases for both the RT‘07 development set and RT‘09 evaluation set and for each of the two possible combinations, illustrated in bold in Tables 7.4 and 7.5 respectively, results for the combined systems are as good as, or better than the best results for either of the single systems. In the case of the LIA-EURECOM/ICSI combination, for the RT‘07 dataset, single system results of 21% and 18% fall to 15% when combined, a relative improvement of 13% over the best single system. For the RT‘09 evaluation set single system results of 31% and 21% fall to 20% which corresponds to a relative improvement of 7% over the best single system. While considering the LIA-EURECOM/I2R combination, we notice a relative improvement of 9% compared to the best standalone system for RT‘07 respectively 7% for RT‘09.

Comparative speaker statistics for the combined system are also illustrated in Table 7.2. We note that for the two different combinations, even though each time *both* systems over-estimate the number of speakers for the RT‘07 dataset, the combined system gives a more accurate estimate. Similar improvements are observed with the error in the number of detected speakers. For the RT‘09 dataset, in the case of the LIA-EURECOM/ICSI combination, both single systems estimate the same number of

speakers and no improvement is obtained with the combined system. For the LIA-EURECOM/I2R combination, the top-down system originally under-estimated the number of speakers for RT'09, while the I2R system showed a reversed trend, the fused system gives a better estimate of the number of speakers according to the averaged error.

When we compare the number of segments and their average length, as illustrated in Table 7.3, we notice consistent improvements over the LIA-EURECOM system and I2R system only. This behavior is to be expected for 2 reasons:

- During the last resegmentation in the fused system, a succession of adaptations and realignments are made with the same algorithms used in the LIA-EURECOM system.
- The ICSI system provides some outputs whose segment durations are closer to the ground-truth. However, in contrast to the LIA-EURECOM/I2R combination, only the outputs of the top-down system are utilized in order to select frames for re-estimating new speaker models in the case of unmatched clusters.

The comparison of columns 4 and 5 in Tables 7.4 and 7.5 shows how well the combination performs with respect to the optimum combination. We see that in many cases the combined system achieves performance very close to the optimum but also that there are plenty of examples where the combined system gives results which are far from the optimum and thus more work is required to improve practical combination performance. This is particularly true for the RT'09 dataset and can be explained since the degree of overlapping speech is particularly high on this dataset (13.6% cf. 7.6% for RT'07).

Speaker diarization performance of Table 7.6 in which results are presented with (OV) and without (NOV) the scoring of overlapping speech confirm this hypothesis. Of note is the large difference in performance with and without the scoring of overlapping speech on the RT'09 dataset: we can estimate a difference of 3.3% absolute DER with/without overlapped speech on the RT'07 dataset versus 5.2% on RT'09.

7.4 Discussion

This chapter introduces two different ways to combine a top-down and a bottom-up system. One aims to first run the systems individually in order to then combine their outputs while the second integrates bottom-up clustering in the heart of the top-down

Table 7.7: Average and variance of the inter-cluster phone distribution distance for each show in the RT'07 and RT'09 datasets. As in Table 6.1 but considering the combined systems

	Mean		**Variance**	
System	**RT'07**	**RT'09**	**RT'07**	**RT'09**
Top-down + Pur.	0.07	0.08	0.001	0.002
Bottom-up (I2R)	0.17	0.14	0.014	0.013
Bottom-up (ICSI)	0.16	0.23	0.005	0.017
Combination LIA-EURECOM/ICSI	0.09	0.10	0.001	0.001
Combination LIA-EURECOM/I2R	0.07	0.07	0.001	0.002
Integrated System	0.13	0.09	0.003	0.001

system. Chapter 6 highlights the difference of behaviors toward linguistic content of each standalone clustering approach and from this point of view it is interesting to make a similar comparison for the new fused and integrated resulting systems.

Similar to Table 6.1, Table 7.7 gives the average and the variance of the inter-cluster phone distribution KL2 distance, as defined in equations (6.9) and (6.10) for the integrated, the combined systems and the different standalone top-down and bottom-up systems. As mentioned in Section 6.3.1, when the clusters are well normalized toward lexical content, we can expect the KL2 distance between the distributions of two different clusters to be small. In contrast, in the case of a possible convergence to another acoustic class (i.e. phoneme) rather than toward speaker, we expect the phone distribution between clusters to be high.

In Table 7.7, the two first lines are given for reference only and are the same than in Table 6.1. Line 3 gives an estimate of the phone normalization for the output obtained with ICSI's bottom-up system. The mean is comparable to I2R's system for the RT'07 dataset, while it is worst for RT'09. In all cases the top-down system with purification provides better normalized outputs compared to the bottom-up system, i.e. average of 0.07 for RT'07 (resp. 0.08 for RT'09) for the top-down system, vs. 0.17/0.16 for RT'07 (resp. 0.14/0.23 for RT'09).

Lines 4 and 5 give an estimate of the phone normalization for the two fused systems. For the LIA-EURECOM/I2R combination the average inter-cluster phone distribution

distance is still low and comparable to the top-down system. However for the LIA-EURECOM/ICSI combined system, the average distance is slightly higher than for the top-down system but lower than the bottom-up. In both cases we note a positive improvement in terms of phone normalization while combining the output of two systems.

Finally, looking at the integrated system which embeds I2R's technology at the heart of the top-down system, we observe slightly improved performances in terms of phone normalization for RT'09 and RT'07 datasets, compared to I2R's system. The system integration seems efficient to improve the phone normalization, but the output combination shows even more efficiency.

In order to assess the real strength of each of the system combinations we compared the performance in terms of DER[1] shown in Table 7.6 and Table 7.1. While looking at the absolute final DER of RT'07 and RT'09 datasets for the integrated system and the fused system we observe that the best performance is always obtained with the fused system. This strengthens the trend in terms of phone normalization highlighted previously.

[1]DER including the scoring of the overlapped speech

Chapter 8

Linguistic Normalization

Chapter 7 presents different ways to combine top-down and bottom-up speaker diarization systems thereby exploiting the benefits of each approach and particular behaviour toward linguistic variability. In this chapter we propose an alternative approach to reduce the linguistic variation direction from within the feature space, with a phone normalization strategy. To apply such a technique, the transcription from an ASR system is required. Since there is little collaboration between the speaker diarization and speech recognition communities there is only little prior work in the literature in this direction. For example, [Chen et al., 2010] proposes to model speakers with some phonetic subspace mixture in a bottom-up diarization system, in order to introduce phonetic information to the ΔBIC distance measure, or [Žibert et al., 2006] presents a SAD component using the output of an ASR system. However each of these approaches uses lexical information only with a single system component (e.g. for cluster fusion, or SAD).

In this chapter we present our latest work involving a novel technique referred to as Phone Adaptive Training (PAT) by analogy to Speaker Adaptive Training (SAT) used in speech recognition. PAT aims to attenuate linguistic variation and leads to a more speaker discriminative feature space, and hence better diarization performance. Section 8.1 presents the SAT technique and introduces the PAT algorithm. In Section 8.2 we present some experimental work which aims to explain the behavior of the approach. Finally Section 8.3 gives some speaker diarization experimental results when PAT is combined with the speaker diarization systems described previously.

8.1 From Speaker Adaptive Training to Phone Adaptive Training

Speaker adaptive training (SAT), a technique developed by the automatic speech recognition (ASR) community, aims to adapt an acoustic space to suppress speaker variability, considered as noise in an ASR problem. At the same time it is essential to keep the wanted variability, in the case of ASR, the phonetic variation which contains the information sought in speech recognition. With the same analogy, we introduce Phone Adaptive Training (PAT) which aims to decouple speaker and phonetic variability in order that the latter is suppressed to provide a more speaker discriminative feature space for speaker diarization.

In Section 8.1.1 and Section 8.1.2, the MLLR and constrained MLLR algorithms are first introduced. Section 8.1.3 details the application of SAT. Finally, in Section 8.1.4, we present the PAT approach.

8.1.1 Maximum Likelihood Linear Regression - MLLR

Maximum Likelihood Linear Regression (MLLR) is a technique for model adaptation using a linear transformation. The algorithm computes transforms which reduce the mismatch between an initial model and an adaptation dataset. When the model is a GMM, the effect of this transformation results in shifting the mean and altering the variance so that the GMM is more likely to generate the adaptation data. In [Leggetter & P., 1995], the adaptation of the component mean is defined as:

$$\hat{\mu} = A\mu + b \tag{8.1}$$

where the transform is characterized by an $n \times n$ regression matrix A and the n-dimensional vector b (n being the dimension of the feature space). Both are optimized according to a standard expectation maximisation algorithm as presented in [Dempster et al., 1977] to maximize the likelihood of the model with respect to the adaptation data.

Let ξ be an (n+1)-dimensional, extended mean vector defined as follows:

$$\xi = [\ \omega\ \mu_1\ \mu_2\ \mu_3\ ...\ \mu_n\ ]^T \tag{8.2}$$

where ω is a bias offset whose value is usually set to 1. Equation (8.1) then becomes:

$$\hat{\mu} = W\xi \tag{8.3}$$

where W is the $n \times (n+1)$ transformation matrix including the bias:

$$W = [\ b\ A\ ] \tag{8.4}$$

A global W transform is sufficient to adapt the whole GMM, however, for more accuracy, when a sufficient amount of data is available, several transforms can be computed for different groups of similar components, allowing the use of a more specific transform. A way to cluster the components in classes according to their similarity involves the use of a regression tree [Leggetter & Woodland, 1995] in order to group together Gaussian components that are close together in acoustic space. MLLR is widely used in ASR when the size of adaptation data is often restrained and the whole speaker specific linguistic model has to be adapted, i.e. a model for each phoneme.

MLLR can also be applied to adapt the Gaussian covariance matrix Σ as explained in [Gales, 1998; Gales & Woodland, 1996]. In [Gales & Woodland, 1996] the transform to update the variance is defined as follows:

$$\hat{\Sigma} = BHB^T \tag{8.5}$$

where $\hat{\Sigma}$ is the new updated variance, H is the $n \times n$ linear transformation matrix to be estimated and B is the Choleski factor of the inverse covariance matrix Σ^{-1}:

$$\Sigma^{-1} = CC^T \tag{8.6}$$

and

$$B = C^{-1} \tag{8.7}$$

8.1.2 Constrained Maximum Likelihood Linear Regression - CMLLR

In contrast to the mean and covariance MLLR where two transforms are independently optimized, Digalakis et al. [1995] propose to update all the parameters i.e. mean and

covariance, with one joint transform. This technique is called Constrained Maximum Likelihood Linear Regression (CMLLR). Equations (8.1) and (8.5) then become :

$$\hat{\mu} = A_c\mu + b_c \tag{8.8}$$

$$\hat{\Sigma} = A_c H A_c^T \tag{8.9}$$

where A_c and b_c are the constrained transform matrix and bias vector respectively estimated in the maximum likelihood sense from the training data.

The constrained nature of CMLLR reduces the number of variables to be optimized during the estimation process and thus requires a smaller amount of training data as compared to separate mean and covariance MLLR matrices. Moreover, since the variance and the mean transforms are tied together, CMLLR can also be used in order to transform the input feature instead of transforming the model. In this case, the observation vectors o_t are transformed as follows:

$$\hat{o_t} = A_c^{-1} o_t + A_c^{-1} b_c \tag{8.10}$$

However, there is no tractable solution for the computation of the CMLLR transform and thus it is estimated through an iterative update process.

8.1.3 Speaker Adaptive Training - SAT

First introduced by Anastasakos et al. [1996], SAT aims to decouple speaker and phonetic variation, in order to reduce the speaker variation which is desirable for the ASR task. In order to decouple these variations, SAT jointly estimates a set of speaker transforms to capture speaker variations and a canonical speaker independent language model λ.

We consider a training set of R speakers ($r = 1, 2..., R$) which contribute to a transcribed observation sequence $O^{(r)} = (o_1^{(r)}, o_2^{(r)}...o_{T_r}^{(r)})$ of length T_r. If we hypothesize that all observations are produced by the same source, i.e. speaker characteristics, and channel conditions and noise levels are constant through the training set, then the optimal

acoustic model $\bar{\lambda}$ can be computed according to:

$$\bar{\lambda} = \underset{\lambda}{\operatorname{argmax}} P(O|\lambda) = \underset{\lambda}{\operatorname{argmax}} \prod_{r=1}^{R} P(O^{(r)}|\lambda) \tag{8.11}$$

where $P(O^{(r)}|\lambda)$ is the likelihood of the model λ with respect to the observations $O^{(r)}$.

However, speaker variation generally has a considerable impact and it is thus desirable to suppress it. The SAT paradigm involves the estimation of a speaker transform $G^{(r)}$ for each speaker r which captures the speaker component, while computing simultaneously a speaker independent model λ_{SAT} normalized across speakers which captures the phonetic characteristics. The SAT optimization task can be defined as follows.

$$(\bar{\lambda}_{SAT}, \bar{G}) = \underset{\lambda_{SAT}, G^{(r)}}{\operatorname{argmax}} \prod_{r=1}^{R} P(O^{(r)}|\lambda_{SAT}, G^{(r)}) \tag{8.12}$$

where $\bar{G} = (\bar{G}^{(1)}, \bar{G}^{(2)}, ...\bar{G}^{(R)})$ is the estimate of each speaker transform.

Practically, SAT parameters are estimated in an iterative process as detailed below:

1. Train a speaker independent model with all the available speech data from all the speakers.

2. From the speaker independent model obtained in (1), estimate a CMLLR transform for each of the speakers in the training set.

3. Normalize the feature vectors of each speaker according to their specific transform computed in (2).

4. Retrain the speaker independent model with the normalized feature vectors from (3).

5. Repeat steps (1) to (4) until likelihood scores converge. The final set of models correspond to the speaker normalized model set λ_{SAT}.

8.1.4 Phone Adaptive Training - PAT

By analogy to SAT, we present a novel approach which we refer to as Phone Adaptive Training (PAT) which aims to decouple speaker and phonetic variations in order to then remove the phonetic variations considered as noise while discriminating speakers.

With this end in view, PAT estimates a phoneme transformation $W^{(p)}$ for each phoneme (or acoustic class) p capturing the linguistic component. Simultaneously the algorithm trains iteratively a phoneme independent speaker model $\Lambda_{PAT} = (\lambda_{PAT}^{(1)}, \lambda_{PAT}^{(2)}, ...\lambda_{PAT}^{(S)})$ normalized across phonemes. The PAT problem can be defined as follows.

$$(\bar{\Lambda}_{PAT}, \bar{W}) = \underset{\Lambda_{PAT}, W}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{p=1}^{P} P(O^{(r,p)} | W^{(p)} \lambda_{PAT}^{(r)}) \tag{8.13}$$

where $\bar{W} = (\bar{W}^{(1)}, \bar{W}^{(2)}, ...\bar{W}^{(P)})$ is the estimate of each phoneme transform and P is the total number of phonemes (or acoustic classes).
Practically, PAT parameters can be estimated in an iterative process as detailed below:

1. Train a phoneme independent speech model for each speaker.
2. From the phoneme independent speech model obtained in (1), estimate a CMLLR transform for each of the phonemes in the training set.
3. Normalize the feature vectors corresponding to each phoneme according to their specific transform computed in (2)
4. Retrain each phoneme independent speech model for each speaker with the normalized feature vectors from (3)
5. Repeat steps (1) to (4) until likelihood scores converge. The final model corresponds to the phoneme normalized speech model λ_{PAT}.

Note that the CMLLR transform computed in step (2) for each phoneme is shared across all speakers. However due to data limitations, more general acoustic classes are sometimes preferred to phoneme classes.

In order to build the acoustic classes, a binary regression tree based on linguistic analysis can be used. It defines different groups of phonemes proposed by phoneticians

and reported in Appendix A. The regression tree is first initialized with one single acoustic class. Then iteratively each class is split into two sets in a top-down approach. The use of two sets rather than one, allows the likelihood of the training data to be increased and the split which maximizes this increase is selected for the first branch of the tree. The process is then repeated until the increase in likelihood achievable by any split at any node is less than a threshold.

8.2 Phone Adaptive Training: Preliminary Experiments

As discussed above, the use of the PAT requires a speech transcription and the speaker segmentation ground-truth. These requirements do not fit with the unsupervised nature of the diarization task, however, in this section we use an oracle setup which hypothesizes that this information is known in order to show experimentally the potential of PAT to improve speaker discrimination. We first compare in Section 8.2.1 different methods to evaluate speaker discrimination, while in Section 8.2.2 we describe the oracle experiment and results in terms of speaker discrimination and diarization performance.

8.2.1 Measure of the Speaker Discrimination

Speaker diarization is a task involving two joint challenges: speaker segmentation and speaker clustering. While improving the speaker discrimination of the feature space, we expect the system to better differentiate between speakers and so to provide better speaker segmentation. Higher quality speaker segmentation is then expected to improve speaker clustering and therefore provide a better overall DER.

However, since the DER depends on the performance of both speaker segmentation and speaker clustering, it cannot be used as a direct measure of speaker discrimination which leads us to use an alternative, better-suited metric. [Duda et al., 2000] proposes different measures for the distances between clusters, however we have to consider some additional constraints. First, we are working with 21 dimensions and, second, speaker clusters cannot be considered as a single Gaussian distribution but as a mixture of Gaussians. For these reasons, the Kullback Leibler (KL) divergence cannot be used since it relates to Gaussian distributions. While the KL divergence for GMMs can be estimated [Hershey & Olsen, 2007], it requires first the training of a GMM for each cluster, involving another source of potential error. Finally there is the Fisher metric

which measures the ratio of inter-class variability and intra-class variability. The Fisher metric was used in [Friedland et al., 2009] to measure the discrimination of prosodic and long term features. The Fisher metric can be defined as follows:

$$score_{Fisher} = \frac{\sum_{i=1}\sum_{j=1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\sum_{i=1}\sum_{j:y_j=i}(x_j - \mu_i)^2} \tag{8.14}$$

where x represents a sample feature value, μ is the mean value for the feature for a given speaker i, or j, and where y_j is the speaker index for the jth sample.

Additionally, instead of using a measure based on the inter-cluster distances, we can measure speaker discrimination as follows:

1. A 16-component GMM is trained with an EM algorithm for each speaker using only 50% of the total speaker time in the ground-truth segmentation.

2. A speaker is attributed by maximum likelihood to each frame of the speech available in the recording. (50% of these speech frames were not used for the training)

3. The overall speaker error is then computed by summing up the False Alarm and Missed Speaker decisions.

8.2.2 Oracle Experiment

In order to assess the potential of the PAT algorithm we introduce an oracle experiment and then analyze the effects on speaker discrimination and diarization performance. Note that these experiments assume knowledge of the ground-truth speech transcription. For this reason the development dataset presented in Section 3.3.1 is truncated, keeping only the 9 files shown in Table 8.1 for which the transcription is available.

Meetings ID	
AMI_20041210-1052_ci01_NONE	ICSI_20011113-1100_d02_NONE
AMI_20050204-1206_ci01_NONE	NIST_20050427-0939_d02_NONE
CMU_20050228-1615_d02_NONE	VT_20050304-1300_d01_NONE
CMU_20050301-1415_d02_NONE	VT_20050318-1430_d01_NONE
ICSI_20010531-1030_d05_NONE	

Table 8.1: Development set used for the PAT process

8.2.2.1 PAT Oracle Experiment

While using an oracle setup, all available information about the speakers and the text transcription can be used. Using this setup, the PAT scenario introduced in Section 8.1.4 can be applied directly to each recording. The signal is characterized by 20 unnormalized LFCCs plus energy coefficients computed every 10ms using a 20ms window. The independent speech models (16-component GMMs) of step (1) in Section 8.1.4 are MAP adapted for each speaker according to the ground-truth. The global process (steps 1 to 5) are repeated 20 times.

Due to the limited quantity of data present in one recording for each speaker and each phoneme, a regression tree is applied as described in Section 8.1.4. The number of acoustic classes which is dynamically controlled by the regression tree plays an important role for the performance. A trade-off has to be found in order to ensure sufficient data per class for training accurate CMLLR transforms, while ensuring enough acoustic classes so that phonetic variations are well modeled.

8.2.2.2 Effect on Speaker Discrimination

In order to assess the capacity of the PAT process, speaker discrimination is evaluated as explained in Section 8.2.1. The impact of two main parameters needs to be analyzed, namely the number of iterations required to reach convergence and the optimal number of acoustic classes.

Effects on speaker and phoneme discrimination using the Fisher metric are shown in Figure 8.1. First, algorithm convergence needs to be estimated in order to know when to stop the iterative process presented in Section 8.1.4. The red curve in Figure 8.1 illustrates phoneme discrimination which drops rapidly over the first 5 iterations. The big decrease of phoneme discrimination observed at the first iteration is due to the use of acoustic classes, tying together different phonemes i.e. a common CMLLR transform is computed for the set of phonemes of a same acoustic class. The phoneme discrimination stabilizes after 15 iterations. The blue curve in Figure 8.1 illustrates speaker discrimination. A significant increase is observed over the first 10 iterations. However, as mentioned in Section 8.2.1, the Fisher metric is not always robust in the case of complex data of high dimensionality. For this reason, results have to be corroborated with another criterion.

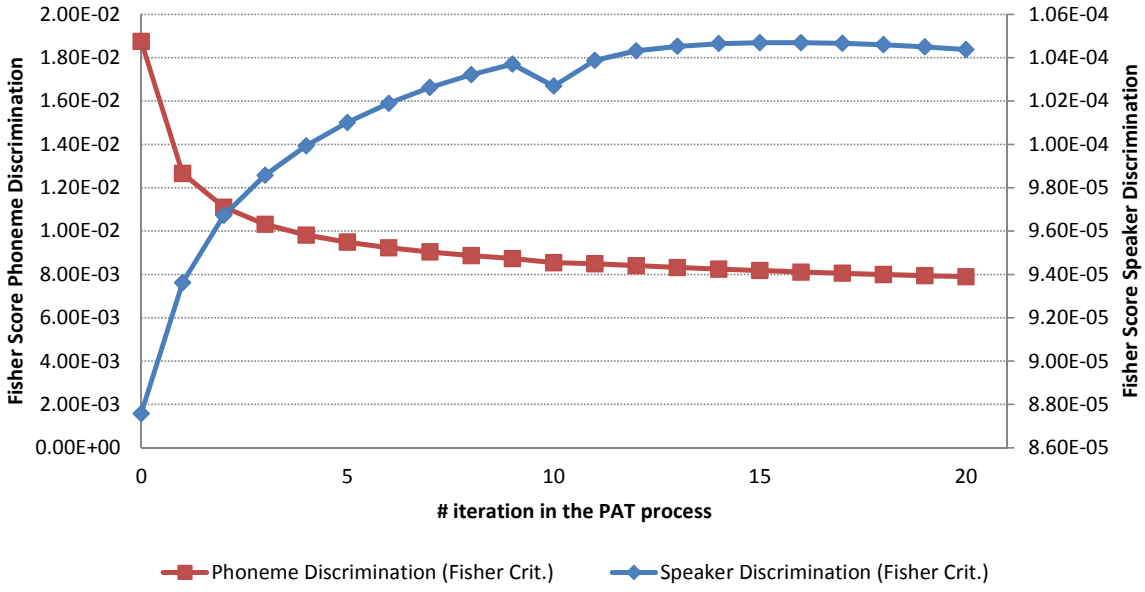


Figure 8.1: Evolution Fisher criterion

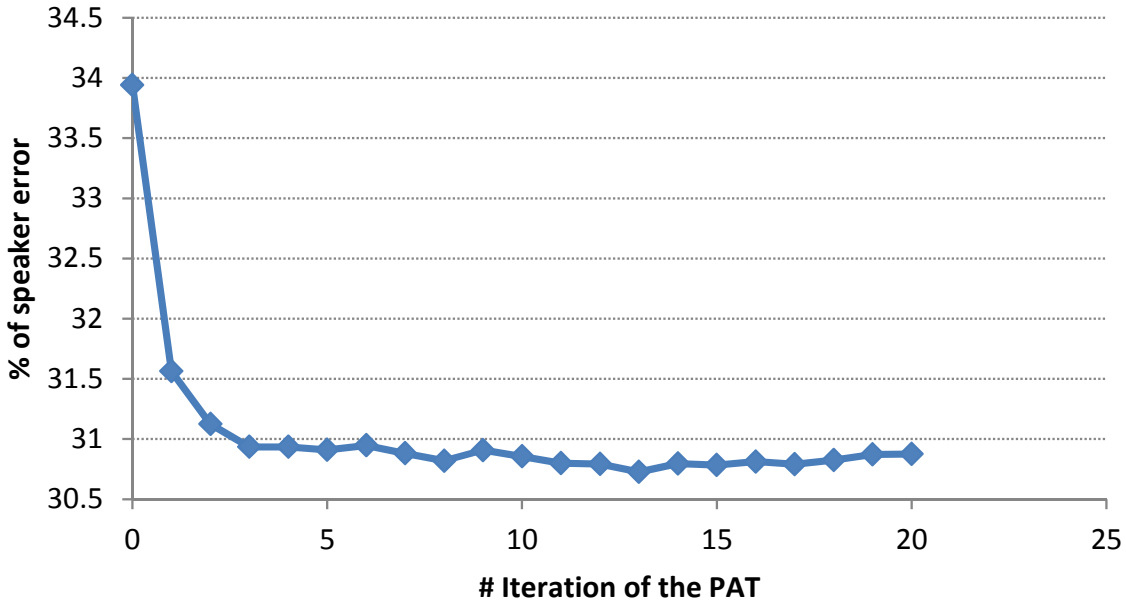


Figure 8.2: Convergence of the Speaker Error across iterations

Figure 8.2 shows speaker discrimination as a percentage of speaker error according to the process described in Section 8.2.1. The first 20 iterations of the PAT algorithm are illustrated for an averaged number of 25 acoustic classes[1]. The profile confirms the tendency observed with the Fisher metric. We observe in Figure 8.2 a significant decrease in speaker error of 9% relative within the first five iterations.

An improvement of 9% relative is quite significant insofar as we did not use any time and duration models to assess a speaker label for each frame i.e. the decision for the frame at time t is taken independently of previous and subsequent frames.

The number of acoustic classes generated by the regression tree has to be optimized in order to get a compromise between the accuracy of the CMLLR transforms (ideally one per phoneme) and the quantity of data available for their training. Figure 8.3 shows speaker discrimination for different numbers of acoustic classes. According to Figure 8.2 the biggest change in speaker discrimination is observed for the first iteration of PAT and so all points in Figure 8.3 relate to a single iteration.

When the number of acoustic classes is too small to reliably capture phonetic variation (around 10 classes), the improvement in speaker error is low (5% relative improvement). With a greater number of acoustic classes we can expect phonetic variation to be better modeled thereby leading to an increase of 7% relative improvement with 24 classes. However, increasing the number of acoustic classes further leads to a decrease in performance due to the limited amount of data which does not permit the reliable estimation of CMMLR transforms for each class. In all the following we choose an average of 25 acoustic classes, which gives an acceptable trade-off between CMLLR training reliability given the quantity of data available.

8.2.2.3 Effect on Diarization Performance

Through the oracle experiment, we show that speaker discrimination is improved with PAT. We now investigate the effect of phone normalization on diarization performance. The Top-down system described in Section 3.4.1 is fed with the normalized features and the segmentation and resegmentation steps are performed in exactly the same way as before (without further optimization). Note that the resegmentation step requires the use of a UBM. Originally the UBM was trained on a speaker recognition corpus as

[1]Due to the dynamic aspect of the regression tree defining the acoustic classes, the number of classes may be different for each of the recordings.

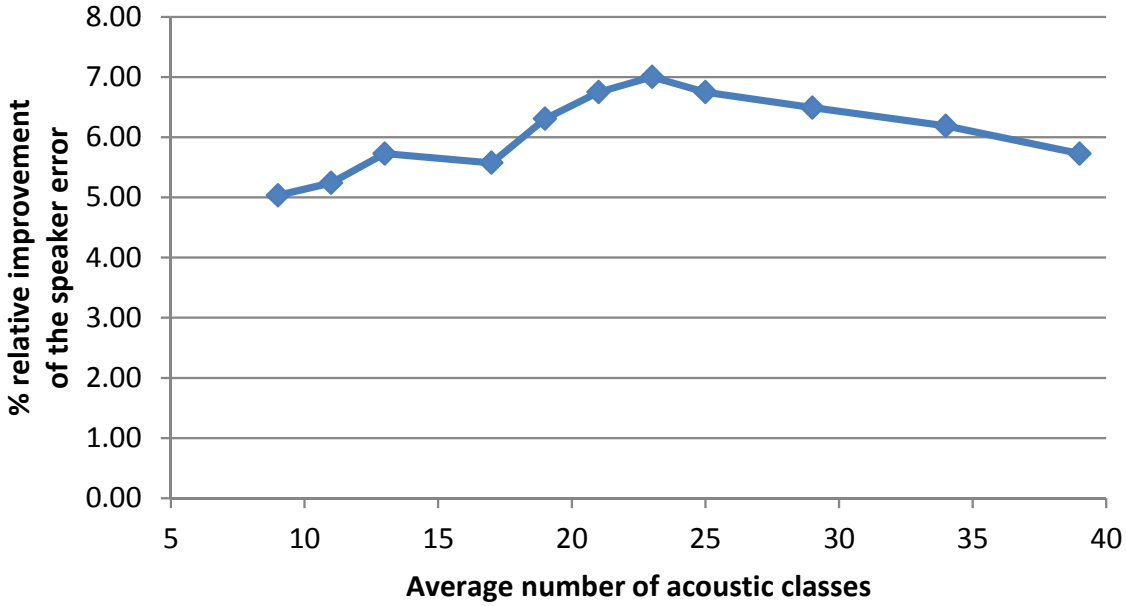


Figure 8.3: Influence of the number of acoustic classes on speaker discrimination

mentioned in Section 8.2.2.1. However, in this case, the UBM like has to be trained on a normalized feature space. Due to the unavailability of the text transcription on the speaker recognition corpus[1], the choice of another training set is necessary. The NIST RT'04 dataset composed of 14 files shown in Table 8.2 is chosen to train the new UBM. Note that this dataset is totally independent of the development set and any of the evaluation sets used in Section 8.3.

Meetings ID	
CMU_20020319-1400_d01_NONE	LDC_20011116-1400_d06_NONE
CMU_20020320-1500_d01_NONE	LDC_20011116-1500_d07_NONE
CMU_20030109-1530_d01_NONE	LDC_20011121-1700_d02_NONE
CMU_20030109-1600_d01_NONE	LDC_20011207-1800_d04_NONE
ICSI_20000807-1000_d05_NONE	NIST_20020214-1148_d01_NONE
ICSI_20010208-1430_d05_NONE	NIST_20020305-1007_d01_NONE
ICSI_20010322-1450_d05_NONE	NIST_20030623-1409_d03_NONE
ICSI_20011030-1030_d02_NONE	NIST_20030925-1517_d03_NONE

Table 8.2: Dataset used for the training of a phoneme normalized UBM (NIST RT04 dataset, SDM conditions)

[1]SRE04 in this case

Table 8.3 presents diarization performance in terms of DER for 3 datasets: the development datasets, composed of the files shown in Table 8.1, and the NIST RT'07 and RT'09 datasets. The second column presents the respective baseline performance obtained with the top-down system and the standard feature space detailed in Section 3.4.1. For consistency, the UBM used for the resegmentation is trained on the standard feature space and the dataset of Table 8.2.

The third column of Table 8.3 shows performance in DER for the top-down system applied in the new feature space obtained by the oracle setup. On the development set, the oracle setup shows a relative improvement of 33 %. While a similar improvement is observed on the RT'07 dataset (25% relative improvement), only 10% relative improvement is shown on the RT'09 dataset. The lower performance on the RT'09 dataset can be explained by the high degree of overlapping speech which brings some artifacts in the captured phonetic component and on another hand, the increased number of speakers which leads to less training data for each speaker.

The average improvement obtained with the oracle experiment over the three datasets is 23% relative, showing there is some potential for diarization performance improvement.

	BASELINE	ORACLE	EXPERIMENTAL
dev Set	23.90	16.07	18.95
RT07	17.13	12.88	15.88
RT09	22.56	20.21	21.45

Table 8.3: Baseline results, oracle experiments and experimental results for the development set detailed in Table 8.1, NIST RT'07 and RT'09 datasets. Results for SDM conditions, without scoring the overlapping speech

8.3 Experimental Results

The oracle experiment presented above confirms the potential of PAT to improve speaker discrimination and to help speaker diarization. However, we previously considered the speaker ground-truth to be known while this is the final objective of the diarization task.

In this section we propose an experimental system based on the output of the baseline speaker segmentation which is used in place of the speaker ground-truth in the PAT process.

The initial speaker segmentation is based on the sequential EM algorithm introduced in Section 3.4.2.2 and uses the standard feature space (20 unnormalized LFCCs plus energy coefficients). The agglomerative hierarchical clustering which originally follows this step is not performed since while the PAT process is almost insensitive to under-clustering, it would be strongly affected in the case of over-clustering. Indeed, in the case where several clusters represent the same speaker, CMLLR training is not directly affected, except, eventually, by a smaller quantity of data being available for each cluster[1]

Finally, the segmentation and resegmentation of the top-down system of Section 3.4.1 is performed on the phone normalized feature space. The last column in Table 8.3 shows performance for this experimental system. For the development set, experimental results shows a significant improvement of 14% relative DER over the baseline leading to a DER close to the optimal oracle performance. Performance improvements on the NIST RT'07 and RT'09 datasets are less significant (7% and 5% relative improvement respectively) but show consistent behavior. While comparing the experimental performance with the optimal oracle DER, we hypothesize that there is still some potential to further improve the experimental setup for the NIST RT'07 and RT'09 datasets. The average improvement over the three datasets is 11% relative.

8.4 Conclusion

This chapter introduces a new phone adaptive training approach to attenuate phonetic variation. An oracle experiment shows that the use of such a process can lead to a new phone normalized feature space which is more speaker discriminative. When performing speaker diarization on the new features obtained through an oracle setup, experiments show some potential for significant improvement.

A more practical experimental setup is also reported where the speaker ground-truth is replaced with an automatically derived segmentation. The feature space produced by the PAT process is then used to feed the top-down baseline system which is not further modified. Results shows an average improvement of 11% relative over 3 datasets.

We have to admit however that the use of the meeting transcript is paradoxical to the unsupervised nature of the speaker diarization task and to come across this problem

[1] in the case of an experimental speaker segmentation, we consider each cluster of the diarization outputs, however they do not always map to a real speaker, and may suffer from under/over- clustering. For this reason we prefer to speak about 'clusters' than 'speakers'.

an automatic speech transcription system would be required. We note however, that the detection of precise phonemes is not fundamental to the proposed approach and imprecise transcriptions may not necessarily lead to inferior performance. There is clear potential in the PAT approach which requires further work to fully optimize in the context of a fully practical diarization system.

Chapter 9

Summary & Conclusions

Speaker diarization is an important step for data analysis, indexation and content structuring. This thesis presents an original framework for the speaker diarization problem and the first thorough comparison of two state-of-the-art approaches, namely bottom-up and top-down, and presents new contributions in purification, system combination and linguistic normalization. While the literature highlights the dominance of bottom-up approaches, we show through different insights that the top-down approach is not without merit and has some specific advantages. A summary of results is given in Section 9.1, while future work is introduced in Section 9.2.

9.1 Summary of Results

Experimental results based on original, state-of-the-art top-down and bottom-up systems show that bottom-up approaches often lead to better performance. However, in this thesis we report a novel purification algorithm which brings an improvement of 15% relative DER over the top-down baseline system. This new baseline system leads to competitive performance and illustrates that none of the approaches is consistently superior.

While, theoretically, the clustering direction should be inconsequential on the speaker inventory which should lead to the same optimal segmentation, we show that the two different approaches exhibit different behaviors toward linguistic variation. Indeed, while ideally the models should be most discriminative for speakers and fully normalized across phones, we show that the merging and splitting operations in the search process are likely

to impact upon the discriminative power and phone-normalization of the intermediate and final speaker models, leading in practice to different behaviors and relative strengths and shortcomings. Our study shows that top-down systems are often better normalized toward phonemes and then more stable, but that they suffer from low speaker discrimination. In contrast, bottom-up clusterings are more speaker discriminative, but as a consequence of their progressive merging scenario, they may be sensitive to phoneme variations which might lead the system to non-optimal, local maxima.

The behavioral differences of the two approaches suggest that there is some potential for system combination. This thesis reports new integrated and combined systems. Experimental results show that system combination is effective in addressing linguistic variation and gives up to 32% relative improvement in diarization performance.

Finally, we show that system performance can be increased by reducing linguistic variation in the feature space. We introduce the first such approach in the context of speaker diarization which we refer to as Phone Adaptive Training (PAT). While the approach is equally relevant to bottom-up and top-down approaches, it is shown to deliver a 10% relative improvement in DER for our own top-down system.

9.2 Future Works

This thesis shows the impact of linguistic variation on top-down and bottom-up approaches to speaker diarization. It highlights specific differences in behavior and demonstrates solutions through purification, system fusion and phoneme normalization through PAT. Future research should extend this work to fully address linguistic variation and thus to further improve performance. In particular, further work is required to address the following:

- **System combination:** The combined system reported in this thesis is based on a 'hard decision' model, i.e. a decision taken during any given iteration cannot be changed in subsequent iterations. This leads to the risk of taking some decision too early while decisions regarding cluster mapping, for example, may be more accurate at the end of the process, if it improves cluster quality. To avoid such drawbacks, the use of a fully Bayesian combination system should be investigated. This solution should have the potential to consider every hypothetical decision

(cluster mapping, cluster fusion) in a probabilistic manner, without 'hard decision' restrictions, and should reconsider each decision up until the final fusion step.

- **Use of speech transcription in PAT:** The PAT approach proposed in this thesis is based on the use of the speech transcription. Although the ground-truth speaker segmentation is not used, we acknowledge that the use of an automatic speech transcription system would be more inline with the unsupervised nature of the diarization task. Future work should thus investigate automatically derived transcriptions. We note that, since the speech transcription only plays a role in the training of CMLLR transforms, the PAT approach should not be overly sensitive to inaccuracies in speech transcription.

- **Data limitation in PAT:** A weakness of the PAT approach proposed in thesis is the amount of data available for the training of the CMLLR transforms which may be impractical in some scenarios. Indeed, CMLLR transforms are trained for each recording and each phoneme, however, with an average duration of only 20 minutes for each NIST show, there is sometimes only little training data for some phonemes which occur rarely. One way to tackle this drawback would involve joint CMLLR transforms across a set of files. However, in this case there is a risk of capturing inter-channel effects. Depending on the recording conditions, the channel effect can differ from show to show and thus transforms learned in this way may lead to be less effective. Furthermore, for the speaker diarization task, the channel effect can be considered as relevant information to distinguish the speakers and should not be removed, e.g. for telephone conversations, since we can expect each speaker to use different telephones, channel information can actually help to track different speakers and thus further work is required to develop the potential of PAT in this context.

- **Overlapping Speech:** Finally, another challenge in the context of speaker diarization (though not addressed in this thesis) involves the handling of overlapping speech. Overlapping speech is known to degrade speaker diarization performance with impacts on both speech activity detection, speaker clustering and segmentation (speaker error) and Anguera et al. [2011]; Huijbregts & Wooters [2007] has shown that overlapping speech can be a dominant source of error. These problems

have attracted increasing attention in recent years and various approaches to detect and attribute intervals of overlap have been proposed. While important advances have been made Boakye et al. [2008]; Huijbregts et al. [2009], the problem remains largely unsolved. We have recently started new work in overlap handling based on convolutive, non-negative matrix factorization with sparse coding constraints. This work is relatively new but there is a large potential to further improve performance and the robustness of speaker diarization to overlapping speech. This is likely to be an area of active research in coming years.

Appendices

Appendix A

Acoustic Group of Phonemes

Name of the Group	Phonemes
Stop	p,pd,b,t,td,d,dd,k,kd,g
Nasal	m,n,en,ng
Fricative	s,sh,z,f,v,ch,jh,th,dh
Liquid	l,el,r,w,y,hh
Vowel	eh,ih,ao,aa,uw,ah,ax,er,ay,oy,ey,iy,ow
C-Front	p,pd,b,m,f,v,w
C-Central	t,td,d,dd,en,n,s,z,sh,th,dh,l,el,r
C-Back	sh,ch,jh,y,k,kd,g,ng,hh
V-Front	iy,ih,eh
V-Central	eh,aa,er,ao
V-Back	uw,aa,ax,uh
Front	p,pd,b,m,f,v,w,iy,ih,eh
Central	t,td,d,dd,en,n,s,z,sh,th,dh,l,el,r,eh,aa,er,ao
Back	sh,ch,jh,y,k,kd,g,ng,hh,aa,uw,ax,uh
Fortis	p,pd,t,td,k,kd,f,th,s,sh,ch
Lenis	b,d,dd,g,v,dh,z,sh,jh
UnFortLenis	m,n,en,ng,hh,l,el,r,y,w
Coronal	t,td,d,dd,n,en,th,dh,s,z,sh,ch,jh,l,el,r
NonCoronal	p,pd,b,m,k,kd,g,ng,f,v,hh,y,w
Anterior	p,pd,b,m,t,td,d,dd,n,en,f,v,th,dh,s,z,l,el,w
NonAnterior	k,kd,g,ng,sh,hh,ch,jh,r,y
Continuent	m,n,en,ng,f,v,th,dh,s,z,sh,hh,l,el,r,y,w
NonContinuent	p,pd,b,t,td,d,dd,k,kd,g,ch,jh
Strident	s,z,sh,ch,jh
NonStrident	f,v,th,dh,hh
UnStrident	p,pd,b,m,t,td,d,dd,n,en,k,kd,g,ng,l,el,r,y,w
Glide	hh,l,el,r,y,w
Syllabic	en,m,l,el,er

Unvoiced-Cons	p,pd,t,td,k,kd,s,sh,f,th,hh,ch
Voiced-Cons	jh,b,d,dd,dh,g,y,l,el,m,n,en,ng,r,v,w,z
Unvoiced-All	p,pd,t,td,k,kd,s,sh,f,th,hh,ch,sil
Long	iy,aa,ow,ao,uw,en,m,l,el
Short	eh,ey,aa,ih,ay,oy,ah,ax,uh
Dipthong	ey,ay,oy,aa,er,en,m,l,el
Front-Start	ey,aa,er
Fronting	ay,ey,oy
High	ih,uw,aa,ax,iy
Medium	ey,er,aa,ax,eh,en,m,l,el
Low	eh,ay,aa,aw,ao,oy
Rounded	ao,uw,aa,ax,oy,w
Unrounded	eh,ih,aa,er,ay,ey,iy,aw,ah,ax,en,m,hh,l,el,r,y
NonAffricate	s,sh,z,f,v,th,dh
Affricate	ch,jh
IVowel	ih,iy
EVowel	eh,ey
AVowel	eh,aa,er,ay,aw
OVowel	ao,oy,aa
UVowel	aa,ax,en,m,l,el,uw
Voiced-Stop	b,d,dd,g
Unvoiced-Stop	p,pd,t,td,k,kd
Front-Stop	p,pd,b
Central-Stop	t,td,d,dd
Back-Stop	k,kd,g
Voiced-Fric	z,sh,dh,ch,v
Unvoiced-Fric	s,sh,th,f,ch
Front-Fric	f,v
Central-Fric	s,z,th,dh
Back-Fric	sh,ch,jh

Table A.1: Group of phonemes for the construction of a regression tree.

Appendix B

French Summary

B.1 Introduction

B.1.1 Motivations

Depuis le $20^{ème}$ siècle, la quantité de données multimédia s'est accrue exponentiellement. Courant 2011-2012, les statistiques[1] montre qu'une moyenne de 60 heures de vidéo est uploadée sur le site *YouTube* chaque minute ou l'équivalent d'une heure de vidéo chaque seconde. 4 milliards de vidéos sont regardées chaque jour. Comme illustré sur la Figure 1.1, ceci représente deux fois plus de données qu'en 2010 et l'on peut s'attendre à ce que ces chiffres augmentent encore dans les années à venir comme le suggère le profile de la courbe.

Face à un nombre de données multimédia toujours croissant, l'indexation et l'analyse automatique des données se sont révélées être la seule stratégie. Différentes approches existent déjà, principalement basées sur l'analyse de contenu vidéo [Truong & Venkatesh, 2007]. Cependant, les vidéos présentes sur les sites de partage proviennent généralement de différents supports notamment webcams, téléphone mobiles, caméras haute résolution, ou encore clips vidéo amateurs utilisant une piste audio et vidéo originalement non enregistrées simultanément : par exemple, la vidéo peut correspondre à un diaporama et ne peut alors pas être considérée comme une 'vraie' vidéo.

Une façon d'analyser la structure des données et de pouvoir annoter différents types de vidéo est d'en extraire l'information disponible dans la piste audio dans le but, éventuellement, de permettre d'associer par la suite cette information à un système de

[1] source : `http://www.youtube.com/t/press_timeline`

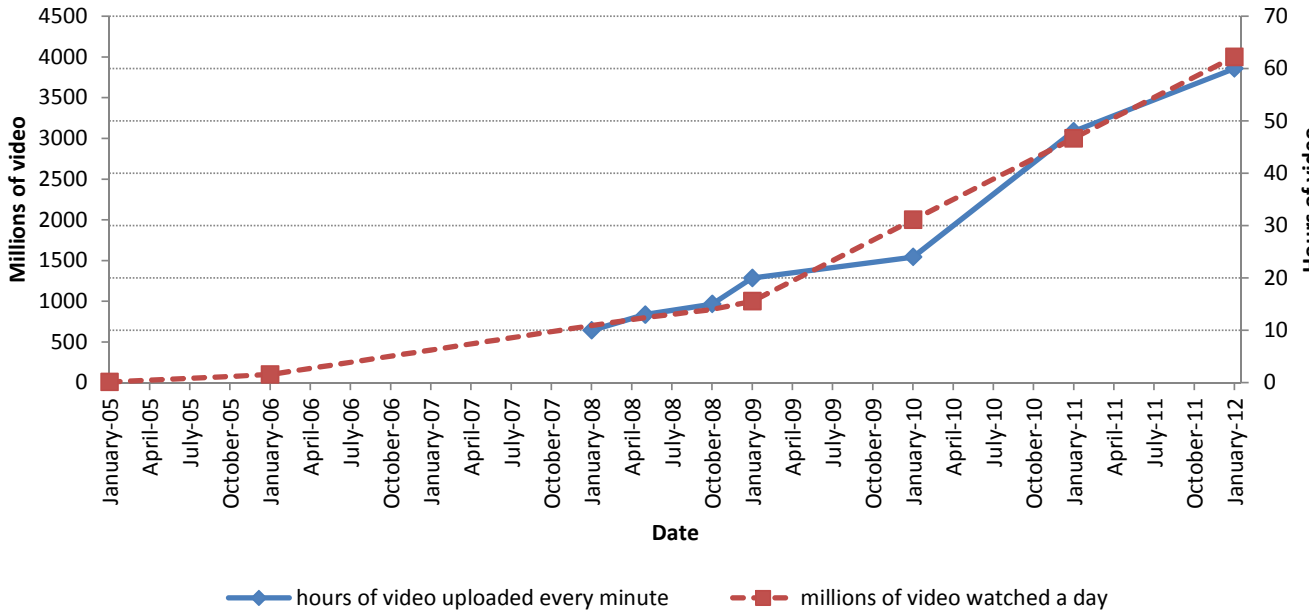


Figure B.1: Evolution du nombre d'heures de vidéo chargées sur *YouTube* de 2005 à 2012 (trait plein), et de la quantit de vidéo regardées par jour en millions (pointillés). Statistiques provenant de : `http://www.youtube.com/t/press_timeline`. Notons qu' aucune donnée n'est disponible de 2005 à 2007 concernant la quantité de vidéo uploadées chaque minute.

reconnaissance vidéo. Tout un ensemble de technologies a pour but d'extraire les informations audio, parmi celles-ci on peut citer la reconnaissance des émotions, la détection d'événements acoustiques, la reconnaissance du locuteur, la détection du langage, la reconnaissance de la parole ou encore la segmentation et le regroupement en locuteur. Alors que la reconnaissance de la parole et du locuteur se rapportent à la reconnaissance de l'identité d'une personne spécifique ou la transcription de ses propos, la segmentation et le regroupement en locuteurs se rapporte au problème "Qui parle quand". Plus formellement, cela requiert l'identification non supervisée de chaque locuteur dans les données audio ainsi que les différents intervalles de temps pendant lesquels chaque locuteur est actif.

Contrairement à la musique ou les événement acoustiques, la parole, de part sa nature sémantique et l'une des composantes les plus informatives du contenu audio. En effet, la transcription de la parole nous renseigne sur des informations clées sur le thème de la discussion, alors que la reconnaissance du locuteur et/ou la segmentation et le

regroupement en locuteurs nous révèle l'identité du locuteur[1] grâce aux caractéristiques issues de la voix. De par sa nature non supervisée, la segmentation et le regroupement en locuteurs a trouvé son utilité dans un bon nombre d'applications ou plusieurs locuteurs peuvent être attendus et qui ont émergé comme un important domaine de recherche en traitement de la parole.

En effet, la segmentation et le regroupement en locuteurs permettent tout d'abord d'indexer et d'extraire les locuteurs de la bande audio dans le but de récupérer l'information essentielle. De plus, lorsque des informations a priori sont connues sur les différents locuteurs, la segmentation et le regroupement en locuteurs peuvent être utilisés comme un pré-traitement pour la reconnaissance du locuteur afin de déterminer l'identité absolue des locuteurs.

De plus, la tache de segmentation et regroupement en locuteurs est considérée comme une étape de pré-traitement importante pour la reconnaissance automatique de la parole dans la mesure ou l'information relative au locuteur facilite l'adaptation du modèle acoustique au locuteur spécifique, comme par exemple la normalisation selon la longueur du conduit vocal ou encore le "Speaker Adaptative Training" (SAT). Dans ce cas, les modèles spécifiques de locuteurs fournissent des retranscriptions plus précises.

La tache de segmentation et de regroupement en locuteurs est donc un pré-requis pour l'indexation audio, l'analyse de contenu, l'annotation automatique, ou encore plus généralement la "Rich Transcription (RT)". Elle fournit ainsi une information directe concernant la structure et l'indexation des locuteurs et peut être utilisée comme une étape de pré-traitement pour la reconnaissance de la parole et du locuteur.

B.1.2 Objectifs de la thèse

La segmentation et le regroupement en locuteurs n'est pas une thématique nouvelle et les recherches dans le domaines ont débuté courant 2002. Comme nous pouvons l'observer sur la Figure B.2, le nombre de publications dans le domaine de la segmentation et le regroupement en locuteurs n'a cessé d'augmenter d'années en années montrant l'intérêt croissant de la communauté et l'importance du domaine. Parmi les différents challenges étudiés par la communauté, quatre principaux domaines ont été abordés. Début 2000, la communauté s'est d'abord concentrée sur les discussions téléphoniques (voir Figure B.3),

[1] ou tout du moins son identité relative dans le cas du problème non supervisé de la segmentation et du regroupement en locuteur

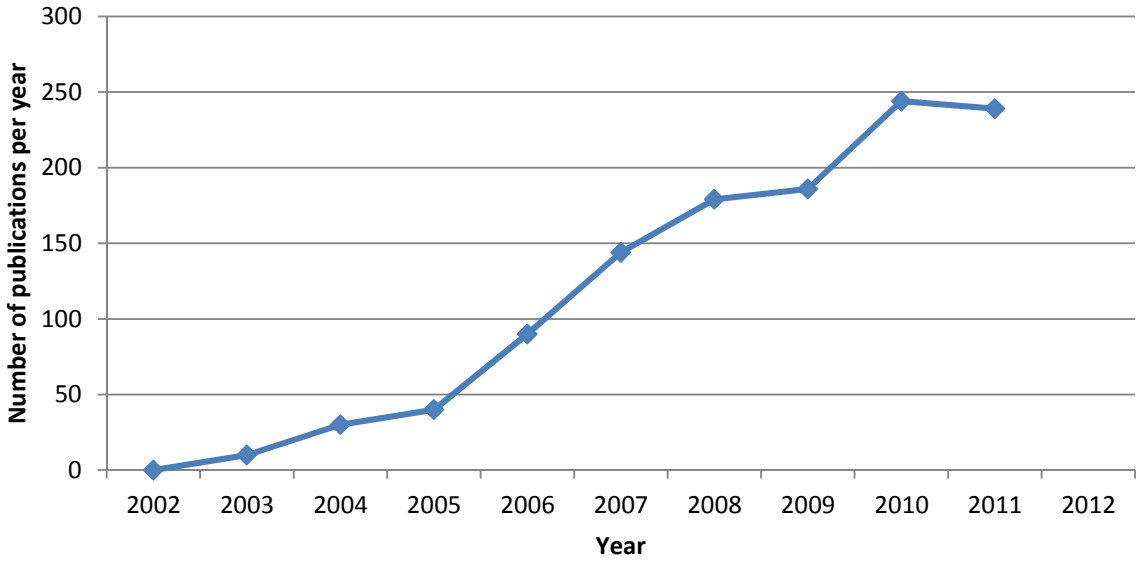


Figure B.2: Nombre de citations par année dans le domaine de la segmentation et du regroupement en locuteurs. Source : *Google Scholar*

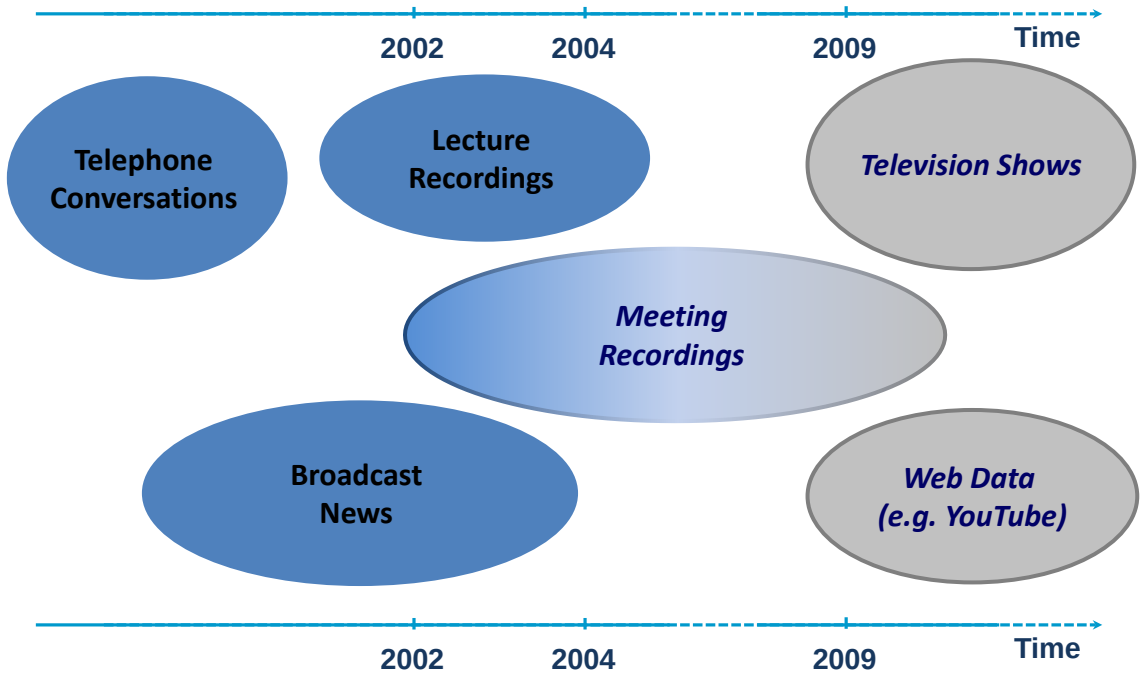


Figure B.3: Les différents domaines d'application de la segmentation et du regroupement en locuteurs

lesquelles correspondent à un problème bien spécifique dans la mesure ou le nombre de locuteurs est connu d'avance. Puis la communauté s'est orientée vers les émissions de type journal télévisé, incluant généralement un locuteur dominant et quelques locuteurs minoritaires. Courant 2002–2004, l'intérêt de la communauté se tourne vers les enregistrements de conférences puis les enregistrements de réunions. L'enregistrement de réunions, de par son nombre généralement important de locuteurs et ses prises de paroles très spontanées (par comparaison aux émissions TV où le scénario repose souvent sur un script préparé) s'est démarqué comme l'une des taches les plus délicates devenant le principal centre d'intérêt de la communauté depuis 2004. Il existe cependant d'autres domaines méritant d'être étudiés, parmi ceux-ci on peut citer : les shows tlviss ou d'une manière plus générale les données présentes sur les sites d'échange de vidéo tels que *YouTube*.

Cette thèse se rapporte à la segmentation et au regroupement en locuteurs pour l'enregistrement de réunions, domaine d'application où la recherche est très active, objet des dernières compétitions internationales NIST. Ces dernières permettent notamment une comparaison rigoureuse des performances avec d'autres systèmes de l'état de l'art. De plus, il est important de mentionner que les enregistrements de réunions comportent certaines caractéristiques en termes de nombre de locuteurs and de spontanéité de prise de parole comparable au données disponibles sur le web.

D'importants progrès ont été réalisés dans le domaine ces dernières années principalement menés par les différentes compétitions organisées par le NIST où deux principales approches se sont démarquées : l'approche ascendante (bottom-up) ainsi que l'approche descendante (top-down). L'approche ascendante est de loin la plus utilisée alors que seulement quelques systèmes sont basés sur l'approche descendante.

Bien que les systèmes les plus performants ces dernières années ont toujours été de type ascendant, nous voulons montrer dans cette thèse que l'approche descendante n'est pas sans mérite et que chaque approche a ses propres avantages. L'objectif de cette thèse peut être formulé comme il suit :

- Peut on considérer que l'une des deux approches ascendante ou descendante soit supérieure à l'autre ?
- Comment leurs comportements diffèrent ?

- Quelles sont leurs principales faiblesses ?
- Comment peut on tirer bénéfice de leurs différents comportements ?

B.1.3 Contributions

Les principales contributions de cette thèse se divisent en quatre points et peuvent être résumées comme il suit :

(i) une nouvelle composante de purification laquelle, appliquée au système descendant, apporte des améliorations notables au système de segmentation et regroupement en locuteurs et rend ainsi l'approche descendante comparable à l'approche ascendante en termes de performance (DER).

(ii) une étude comparative ayant pour but de montrer les différences en termes de comportement entre l'approche ascendante et descendante sur la base d'un cadre de référence commun et une série d'expériences oracle.

(iii) un système de fusion et un système d'intégration descendant/ascendant lesquels confirment que, compte tenu de leurs natures différentes, la combinaison des systèmes descendant/ascendant apporte des améliorations et se traduit par des performances supérieures aux systèmes d'origine.

(iv) Une nouvelle méthode de normalisation à l'échelle des phonèmes permettant d'améliorer les performances du système de segmentation et regroupement en locuteurs.

De plus amples détails sur ces contributions sont donnés ci-après :

(i) Nouvelle composante de purification pour l'approche descendante de segmentation et regroupement en locuteurs

La purification de clusters n'est pas un sujet nouveau dans le domaine de la segmentation et du regroupement en locuteurs, cependant, les travaux antérieurs se rapportent à la purification des clusters pour les systèmes ascendants. Aussi, la première contribution de cette thèse propose une nouvelle composante de purification ajoutée au système descendant. Grâce à celle-ci, la stabilité des performances sur cinq jeux de données du NIST en ressort renforcée et on peut constater une amélioration de 15% relatifs sur l'erreur DER.

Ce travail a été présenté à la conférence : International Conference on Acoustics, Speech, and Signal Processing (ICASSP) en 2010 [Bozonnet et al., 2011].

(ii) Étude comparative des systèmes descendant et ascendant

La deuxième contribution de cette thèse est une analyse détaillée des deux systèmes ascendant et descendant. En effet, les résultats expérimentaux montrent que la nouvelle composante de purification présentée dans la première contribution entraîne des performances incohérentes lorsqu'elle est appliquée au système ascendant. Ceci nous laisse supposer que chaque système a un comportement qui lui est propre imposé par sa nature spécifique. Dans le but de réaliser une analyse complète et rigoureuse, deux type d'étude sont menés : une étude de type Oracle, laquelle souligne les faiblesses de chaque système ainsi qu'une seconde étude détaillant d'avantage les différences en termes de convergence dues aux différents scénarii de clustering. Cette étude aide ainsi à comprendre l'effet négatif causé par l'algorithme de purification lorsqu'il est appliqué sur le système ascendant.

- **Expériences de type Oracle**

  Avec l'aide d'une série d'expériences de type oracle, la sensibilité et la robustesse des différentes composantes de l'approche descendante de référence est analysée dans le but d'identifier leurs possibles faiblesses. Une méthode similaire est réalisée pour le système ascendant. Les résultats expérimentaux montrent que, malgré des faiblesses communes aux deux système dues notamment à la détection de la parole (SAD) et aux traitement des passages multilocuteurs (overlapping speech), les deux algorithmes présentent des lacunes spécifiques. En effet, alors que la méthode ascendantes est quasiment indépendante de son initialisation, elle s'avère très sensible lors de sa phase de fusion des clusters ainsi que pour son critère contrôlant l'arrêt du processus de fusion (stopping criterion), notamment lors de la présence d'impureté dans les clusters (mélange de plusieurs locuteurs par exemple). Au contraire, le scénario du système descendant est principalement sensible à son initialisation et à la qualité de ses modèles initiaux lesquels influencent directement les capacités du système à discriminer les locuteurs.

- **Analyse comportementale et différences en termes de convergence**

  La seconde partie de cette analyse se rapporte aux effets de la direction du clustering (ascendante/descendante). Un cadre théorique incluant une définition formelle de la tache de segmentation et regroupement en locuteurs

ainsi qu'une analyse des challenges qui doivent être résolues sont tout d'abord développés, nous menant à croire que, théoriquement, le résultat final devrait être indépendant de la direction du clustering.

Cependant, nous avons montré qu'alors idéalement les modèles d'un système de regroupement et segmentation en locuteurs devraient être principalement discriminant pour les locuteurs et indépendant des variations acoustiques non désirées telles les phonèmes, il est vraisemblable que les étapes de fusion ou de division des clusters tout au long du processus aient un impacte sur les facultés du système à discriminer les locuteurs et normaliser le contenu phonétique, menant alors en pratique à différents comportements des systèmes avec différents atouts et défauts. En effet, notre étude montre que les systèmes descendants sont souvent mieux normalisés vis à vis des phonèmes and ainsi plus stables, toutefois, ils souffrent souvent d'une faible discrimination inter-locuteurs. Ceci permet d'expliquer pourquoi les systèmes descendants sont améliorés grâce à la composante de purification. Au contraire, les systèmes de type ascendants sont d'avantage discriminants à l'égard des locuteurs cependant, la fusion progressive des clusters les rend plus sensibles aux variations des phonèmes, menant alors à un maximum local non optimal de la fonction de coût.

Ce travail a été présenté à la conférence : International Conference on Acoustics, Speech, and Signal Processing (ICASSP) en 2011 [Bozonnet et al., 2011]. Une version approfondie de ce travail, incluant notamment une analyse plus complète a été publiée dans le journal : the IEEE Transactions on Audio Speech and Language Processing (TALSP), special issue on New Frontiers in Rich Transcription en 2012 [Evans et al., 2012].

(iii) Combinaison de système ascendant/descendant

La contribution précédente souligne les propriétés distinctes en termes de fiabilité des modèles et de discrimination des méthodes ascendantes et descendantes. Ces comportements spécifiques suggèrent ainsi un potentiel pour la combinaison de ces deux systèmes. La troisième contribution de cette thèse présente ainsi de nouvelles

méthodes afin de combiner la méthode ascendante et descendante, bénéficiant ainsi des atouts de chacune d'entre elles, améliorant ainsi les performances et la stabilité. Deux types de combinaison des systèmes sont ainsi étudiées :

- **Système de fusion**

  Le système de fusion a pour but de lancer simultanément et indépendamment les systèmes ascendants et descendants dans le but de combiner leurs sorties. Nous proposons une nouvelle approche qui dans un premier temps couple les clusters de chacun des deux systèmes selon des contraintes imposées sur la matrice de confusion et le contenu acoustique. Grâce à cette association de clusters, une première sélection de clusters est réalisée. Les clusters restants sont alors introduits par la suite selon leur distance acoustique aux clusters déjà sélectionnés. Seules les frames les plus vraisemblables sont conservées. Un ré-alignement final est ensuite réalisé afin d'associer les frames non classées. Grâce à ce scénario une amélioration de 13% relatifs (DER) est obtenue sur les performance du regroupement et de la segmentation en locuteurs.

  Ce travail a été présenté à la conférence : Annual Conference of the International Speech Communication Association (Interspeech) en 2010 [Bozonnet et al., 2010], une version plus approfondie de ce travail sur les effets du système de fusion a été publié dans le journal : the IEEE Transactions on Audio Speech and Language Processing (TALSP) , special issue on New Frontiers in Rich Transcription en 2012 [Evans et al., 2012].

- **Système d'intégration**

  Une approche alternative consiste á fusionner les deux systèmes en leur coeur, on la nomme : approche intégrée. Les systèmes sont lancés simultanément, le système descendant appelant le système ascendant telle une fonction durant son exécution. dans le but d'améliorer la qualité des nouveaux modèles introduits. Les résultats expérimentaux montrent une amélioration sur trois différents jeux de données incluant des enregistrements de réunions et des shows télévisés avec des améliorations atteignant jusqu'à 32% relatifs (DER).

  Ce travail a été présenté à la conférence : Annual Conference of the International Speech Communication Association (Interspeech) en 2010 [Bozonnet et al., 2010].

(iv) Normalisation à l'échelle des phonèmes pour la segmentation et le regroupement en locuteurs

La dernière contribution de cette thèse se rapporte à une nouvelle technologie ayant la capacité de limiter l'influence du contenu linguistique, considéré comme une importante source de nuisance dans notre étude comparative pouvant biaiser la convergence du système de segmentation et regroupement en locuteurs. Par comparaison au Speaker Adaptive Training (SAT), nous proposons d'une manière tout à fait analogue de réduire la composante linguistique dans les caractéristiques acoustiques. Notre approche est appelée Phone Adaptive Training (PAT). Cette technique se base sur une régression linéaire contrainte par maximum de vraisemblance (Constraint Maximum Likelihood Linear Regression CMLLR) laquelle a pour but de supprimer les composantes non désirées grâce à une transformation linéaire des caractéristiques. Les résultats expérimentaux montrent une amélioration de 11% relatifs pour le système de regroupement en locuteurs.

B.1.4 Organisation

Cette thèse se décompose en 8 chapitres comme il suit :

Une étude de l'état de l'art est présentée dans le chapitre 2 incluant les progrès dans le domaine, les principales approches, leurs spécificités ainsi que les principaux domaines étudiés par la communauté.

Le chapitre 3 introduit les métriques, les jeux de données et les procotoles officiels définis par le NIST dans le but de décrire les deux systèmes de référence de l'état de l'art : les méthodes ascendantes (bottom-up) et descendantes (top-down) ainsi que leur performances respectives.

Dans le chapitre 4, une étude Oracle est présentée ayant pour but d'évaluer la sensibilité et la robustesse des différentes composantes du système ascendant et descendant afin d'en comparer leurs points faibles.

Le chapitre 5 introduit une nouvelle composante ayant pour but d'améliorer la qualité des clusters des systèmes en les purifiant. Après une première description de l'algorithme, la nouvelle composante de purification est intégrée dans le système ascendant et descendant et une analyse des performances est menée.

Une étude comparative des méthodes ascendantes et descendantes est détaillée dans le Chapitre 6 incluant tout d'abord une formalisation de la tache de segmentation et

regroupement en locuteurs. Une comparaison qualitative et expérimentale est menée ensuite, montrant les différences de comportement des deux systèmes à l'égard des variations nuisibles telles que le contenu lexical.

Enfin, le chapitre 7 introduit une combinaison des deux systèmes permettant d'exploiter le bénéfice de chacun d'eux afin d'obtenir un système résultant plus performant. Deux scénarii sont considérés et leurs performances respectives sont examinées.

Pour finir, le chapitre 8 introduit une nouvelle technique afin de normaliser les caractéristiques extraites du signal audio, appelée Phone Adaptive Training (PAT). Cette dernière a pour but d'atténuer les effets dus au contenu lexical considérés comme la principale nuisance face à la discrimination des locuteurs. Une description du processus est d'abord introduite, elle est suivie par un jeu d'expériences.

Les conclusions de ce travail sont données dans le chapitre 9 résumant les contributions majeures ainsi que les résultats obtenus dans cette thèse et évoquant différentes perspectives pour des travaux futurs.

B.2 Protocoles & Système de Référence

D'importants progrès ont été réalisés dans le domaine de la segmentation et du regroupement en locuteurs principalement conduits par le NIST (National Institute of Standards and Technology) notamment au travers différentes évaluations (Rich Transcription (RT) evaluations) [NIST, 2002, 2003, 2004, 2006, 2007, 2009]. Tout au long de ces différentes compétitions, deux principales approches ont émergé, elles sont l'approche ascendante (ou encore bottom-up) et l'approche descendante (top-down). Bien que les meilleures performances ont toujours été atteintes par l'approche ascendante ces dernières années, nous pensons que l'approche descendante n'est pas pour autant sans mérite. En effet, les résultats de la dernière évaluation NIST RT'09 montre que l'approche descendante produit des résultats compétitifs[1] et d'une complexité moindre en termes de calculs.

Dans ce chapitre nous décrivons tout d'abord le protocole officiel et les métriques proposés par l'institut NIST et nous introduisons ensuite les différents jeux de données

[1] pour les conditions SDM et MDM (même si dans ce cas, contrairement aux autres systèmes, le retard entre les canaux n'est utilisé que pour effectuer un beamforming et donc non considéré comme une caractéristiques supplémentaires pour distinguer les locuteurs)

utilisés lors des différentes évaluations. Un corpus de shows télévisés est également introduit afin de tester la robustesse des différents systèmes. Enfin, des détails sur les systèmes de référence de chacune des deux approches sont présentés.

B.2.1 Protocoles

Depuis 2004, l'institut NIST a organisé une série de compétitions internationales au travers de la campagne de Rich Transcription (RT)[1]. Ces évaluations, lesquelles incluent la tache de segmentation et regroupement en locuteurs, ont pour but de faciliter la tache des technologies de transcription et d'annotation des données. Grâce à son caractère international, les évaluations RT ont eu un rôle important dans l'évaluation de l'état de l'art en proposant des protocoles d'évaluation standardisés, incluant différentes métriques pour comparer les performances et des jeux de données communs. Une importante caractéristique de ces évaluations est qu'aucune information a priori n'est fournie aux participants (par exemple, le nombre de locuteurs, leurs identités, etc.) à l'exception de la nature des enregistrements (par exemple, les réunions, les journaux télévisés, etc.) et le langage (Anglais). Des formats standards pour les données d'entrées et de sortie des systèmes sont également définis et les participants peuvent utiliser des données externes pour créer un modèle de monde et/ou dans le but de normaliser les données.

Tout d'abord centrées sur les journaux télévisés, les conférences ou encore les pauses café, les évaluations NIST les plus récentes ont porté sur les enregistrements de réunions, un domaine particulièrement délicat pour la segmentation et au regroupement en locuteurs notamment du à la spontanéité de la parole. Pour cette raison, le travail présenté au sein de cette thèse se concentre également sur les enregistrements de réunions. Les enregistrements fournis par le NIST lors des évaluations étaient enregistrés à l'aide de plusieurs microphones de différents types et de différentes qualités lesquels sont positionnés sur les participants (par exemple les micro-cravate) ou dans différents endroits de la salle de réunion. En groupant ces microphones en différentes classes, l'institut NIST propose différents types de conditions pour les évaluations. Parmi celles-ci on peut noter : l'usage d'un unique micro-casque (IHM), un unique microphone distant (SDM), une multitude de microphones distants (MDM), ou encore un ensemble microphones de

[1] Voir http ://nist.gov/speech/tests/rt.

type mark III (MM3A[1]), ou enfin l'ensemble des microphones distants (ADM).

Les conditions MDM sont définies comme le coeur du challenge et sont requises pour tous-les participants. Les participants ont alors la possibilité d'utiliser les données enregistrées simultanément sur les différents canaux issues des différents microphones situés le plus souvent sur la table de la réunion. Pour la plupart des systèmes, un beamforming [Anguera, 2006] est alors réalisé dans le but de créer un pseudo canal, et les retards inter-canaux (ICD) [Anguera et al., 2005; Ellis & Liu, 2004; Evans et al., 2009] peuvent être ajouté aux caractéristiques audio classiques et peuvent ainsi mener à de meilleures performances pour la segmentation et le regroupement en locuteurs [Anguera et al., 2005].

Au contraire, les conditions SDM n'autorisent que l'usage de l'enregistrement issu d'un seul et unique microphone (le plus souvent celui situé le plus au centre) et ainsi ne peuvent pas utiliser de beamforming pour l'amélioration de la qualité du signal ni les ICD comme caractéristiques du locuteur. Dans cette thèse nous nous intéresserons principalement aux résultats pour les conditions SDM car nous les considérons comme étant plus représentatives des équipements les plus courants dans les salles de réunion.

B.2.2 Métriques

L'institut NIST définit un standard pour les sorties des systèmes de segmentation et regroupement en locuteurs, celles-ci doivent contenir une hypothèse sur l'activité de chaque locuteur incluant le temps de début et de fin de chaque segments de parole. Les étiquettes des locuteurs sont utilisées uniquement afin d'identifier les interventions multiples de chaque locuteur mais ne donnent pas d'information quant à leurs réelles identités. Afin d'estimer la qualité de l'hypothèse, les sorties sont comparées à la vérité terrain dans le but d'obtenir un score global : Diarization Error Rate (DER) également défini par le NIST. Cette métrique peut être définie comme la somme temporellement pondérée de trois sources d'erreurs :

- **Parole Manquée - Missed Speech** (MS) : pourcentage de parole mentionnée dans la vérité terrain mais non présente dans l'hypothèse de segmentation et regroupement en locuteurs ;

[1] Les microphones MM3A sont des types de microphones exclusivement produits et fournis par l'institut NIST. Ils ne sont habituellement pas inclus dans les conditions MDM, mais dans les conditions ADM.

- **Fausse Alerte - False Alarm** (FA) : pourcentage de parole mentionnée dans l'hypothèse de segmentation et regroupement en locuteurs mais non présente dans la vérité terrain ;

- **Erreur de Locuteur - Speaker Error** (*SpkErr*) : pourcentage de parole assignée au mauvais locuteur (en ignorant les passages avec plusieurs locuteurs simultanés)

L'erreur DER peut être déterminée avec ou sans l'inclusion des segments comportant plusieurs locuteurs (overlapping speech). Quand les segments comportant plusieurs locuteurs sont évalués, l'erreur DER reflète alors l'erreur dans l'estimation du nombre de locuteurs parlant simultanément (dans l'évaluation NIST RT on peut compter jusqu'à 4 locuteurs simultanés) ainsi que les erreurs dues à l'attribution de la parole à chacun des locuteurs. Les erreurs sur l'estimation du nombre de locuteurs mènent à une augmentation du taux de parole manquée (MS) lorsque moins de locuteurs sont détectés par rapport à leur vrai nombre, ou à une augmentation du taux de fausse alerte (FA) lorsque trop de locuteurs sont détectés. Dans le cas de l'erreur de locuteur (Speaker Error - *SpkErr*), l'erreur respective de chacun des locuteurs prise individuellement est considérée.

L'erreur DER est déterminée à l'aide de l'Equation B.1

$$DER = SAD_{error} + SpkErr = \underbrace{MS + FA}_{\text{SAD Error}} + SpkErr \tag{B.1}$$

Plus précisément, l'erreur DER est calculée comme la fraction du temps de chaque locuteur qui n'est pas correctement attribuée, basé sur l'association optimale des locuteurs de l' hypothèse/vérité terrains. Cette dernière est établie grâce a un algorithme dynamique défini par l'institut NIST. L'erreur DER peut être définie plus formellement comme il suit :

$$DER = \frac{\sum_{\forall i}\{D_i^R \cdot (\max(N_i^R, N_i^S) - N_i^C)\}}{\sum_{\forall i}\{D_i^R \cdot N_i^R\}} \tag{B.2}$$

où D_i^R représente la durée du i-ème segment de référence, et où N_i^R et N_i^S correspondent respectivement au nombre de locuteurs dans la vérité terrain et le nombre de locuteurs dans l'hypothèse de segmentation et regroupement en locuteurs. N_i^C est le nombre de

locuteurs qui est correctement identifié par le système de segmentation et regroupement en locuteurs. Il est toutefois important d'observer qu'en incluant les passages multi-locuteurs, N_i^R,N_i^S et N_i^C peuvent être supérieurs à 1.

Comme nous pouvons le voir dans l'Equation B.2, l'erreur DER est pondérée avec le temps, c'est à dire que moins d'importance est accordée aux locuteurs dont le temps de parole est court. De plus, un paramètre appelé 'collier' définit une zone de 250ms de part et d'autre de chaque segment mentionné dans la vérité terrain qui n'est alors pas évaluée lors du calcul de l'erreur DER. Ceci permet de s'affranchir d'éventuelles erreurs de temps de début/fin de segment lors de l'annotation manuelle de la base de donnée. Pour les shows télévisés pour lesquels un locuteur dominant et de multiple locuteurs relativement inactifs, l'erreur DER n'est pas toujours une bonne métrique de référence dans la mesure ou elle peut être très faible même si un unique et même locuteur est détecté.

On peut observer que depuis 2006, la métrique primaire pour les évaluations RT inclue l'évaluation des passages multi-locuteurs. Cependant, comme les systèmes reportés dans cette thèse ne permettent pas de détecter et de prendre en compte l'annotation multi-locuteurs, nous nous référons ci-après à la métrique sans évaluer les passages multi-locuteurs. Dans ce cas, N_i^R,N_i^S et N_i^C prennent alors les valeurs 0 ou 1. Selon les possibilités, nous indiquerons cependant les scores avec et sans l'évaluation des passages multi-locuteurs.

B.2.3 Jeux de Données

Dans la majeure partie de ce manuscrit, un corpus d'enregistrements de réunions est utilisé à titre expérimental, notamment utilisant le corpus RT de NIST. Cependant, dans le but d'estimer la robustesse des différents systèmes, un jeu de données supplémentaire, regroupant des shows télévisés Français et nommé : Grand Échiquier est présenté dans la Section B.2.3.2.

B.2.3.1 Corpus de Réunions RT

Pour chacune des évaluations NIST RT depuis 2004, un nouveau set d'enregistrement annotés de réunions a été collecté[1] Un total de cinq jeux de données est ainsi disponible.

[1]La vérité terrain est disponible après chaque évaluation afin d'être utilisée par la suite par la communauté pour la recherche et le développement indépendamment des évaluations officielles NIST.

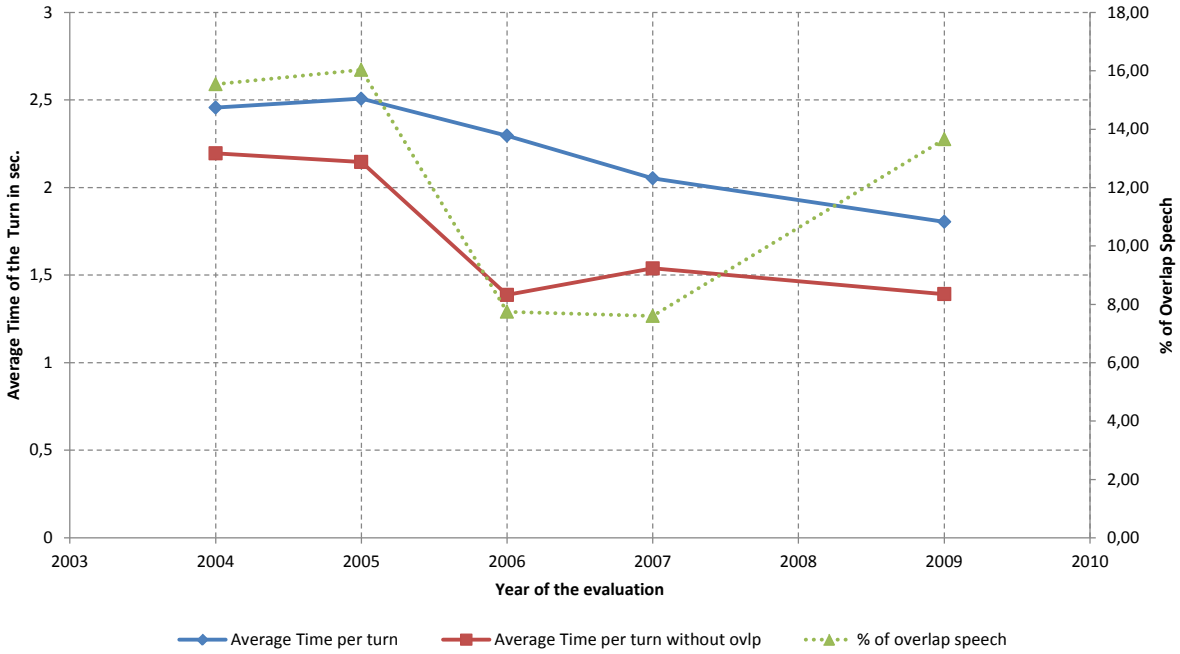


Figure B.4: Analyse des pourcentages de parole multi-locuteurs et de la durée moyenne des changements de locuteurs pour chacun des 5 jeux de données NIST RT. Les pourcentages de parole multi-locuteurs sont donnés en fonction de le temps total de parole.

La Figure B.4 montre les différences entre les différents jeux de données NIST RT en termes de pourcentage de parole multi-locuteurs et de durée de changement de locuteurs. Pour les jeux de données RT'04, RT'05 et RT'09 on peut voir un pourcentage de parole multi-locuteurs d'environs 15%, alors que les jeux de données de 2006 et 2007 contiennent 8% de parole multi-locuteurs. Lorsque l'on s'intéresse à la durée moyenne des changements de locuteurs, laquelle peut être définie comme la durée moyenne durant laquelle aucun changement concernant le(s) locuteur(s) n'a lieu (même locuteurs, mêmes conditions : avec/sans parole multi-locuteurs), on peut observer que les trois dernières évaluations : RT'06, '07 et '09 comportent en moyenne des changements de locuteurs plus fréquents, même si l'on ne considère par la parole multi-locuteurs. Ceci permet de souligner que la parole présente dans les trois dernières évaluations peut être considérée comme plus spontanée et plus interactive, menant ainsi à des changements de locuteurs plus fréquents. En accord avec ces remarques on peut s'attendre à une difficulté accrue pour les jeux de données RT'06,'07 et RT '09.

Par souci de cohérence avec les travaux antérieurs[Fredouille & Evans, 2008;

Fredouille et al., 2004], tous les systèmes expérimentaux présentés dans cette thèse ont été optimisés sur un set de développement de 23 enregistrements de réunions issus des évaluations NIST RT'04, '05 et '06. Les performances ont été alors confortées grâce à deux jeux de données indépendant : RT'07 et RT'09 évaluations. Il est important de préciser qu'aucun recouvrement n'est présent entre le set de développement et d'évaluation même si ces derniers peuvent contenir des enregistrements provenant de sites similaires et éventuellement de locuteurs identiques.

B.2.3.2 Corpus de shows télévisés GE

Par l'intermédiaire d'autres travaux [Bozonnet et al., 2010] nous avons également expérimentés nos systèmes sur une base de données de shows télévisés tel que 'Grand Échiquier' (GE). Dans la mesure où ces résultats nous ont permis d'évaluer la robustesse des systèmes de segmentation et regroupement en locuteurs (par exemple, dû notamment aux variations des temps de paroles des locuteurs, à la présence d'un locuteur dominant, etc...), ce jeu de données est décrit ci-après. Les résultats de référence concernant cette base de données figurent en Section 3.5.

Ce corpus comprend plus de 50 shows télévisés Français issus des années 1970 – 1980 et a été rendu populaire grâce à différents projets nationaux et Européens comme par exemple : the European K-Space network of excellence [K-Space, K-Space]. Chaque show est centré sur un invité principal et quelques autres invités minoritaires, tous sont interviewés par un présentateur. Les interviews sont ponctués par des intermèdes musicaux, des extraits de films, les applaudissements du publique ou encore des rires. Hormis ceci, les silences entre les changements de locuteurs peuvent être très brefs ou quasiment négligeable. Comparés aux enregistrement de réunions ou les locuteurs souvent s'arrête un instant pour réfléchir avant de répondre à une question, ou pour réorganiser leurs idées, les shows télévisés sont beaucoup plus fluide et parfois quasiment écrits. Ceci est probablement dû au fait que les principaux thèmes et discussions abordés sont souvent préparés à l'avance et connus des locuteurs.

Le Tableau 3.1 souligne d'une manière plus quantitative les différences entre les réunions issues de la base de données NIST RT'09 et 07 shows télévisés issus de la base de données GE, laquelle a été annotée manuellement avec le strict respect du protocole NIST [NIST, 2009]. En comparant les trois premières lignes du Tableau 3.1 on observe que les shows télévisés sont en moyenne plus longs que les réunions(147 minutes contre

Attribute	GE	NIST RT'09
No. of shows	7	7
Avg. Evaluation time	147 min.	25 min.
Total speech	50 min.	21 min.
Avg. No. of segments	1033	882
Avg. segment length	3 sec.	2 sec.
Avg. Overlap	5 min.	3 min.
Avg. % Overlap / Total speech	10 %	14 %
Avg. No. speakers	13	5
most active	1476 sec.	535 sec.
least active	7 sec.	146 sec.

Table B.1: Comparaison des caractéristiques issues des bases de données Échiquier (GE) et NIST RT'09

25 minutes) et en supprimant le bruit (par exemple les applaudissements) et la musique, la quantité de parole est deux fois celle que l'on peut trouver dans la base de données RT (50 minutes contre 21 minutes). Notons cependant que la durée moyenne des segments est légèrement plus petite pour RT'09 que pour GE (2 sec. contre 3 sec.). Ces premières investigations peuvent suggérer que les shows télévisés présentent d'avantage qu'un simple challenge dû à leur plus variabilité au sein d'un même locuteur plus important tout au long d'un même show.

De plus, les différences en termes de statistiques des locuteurs doivent être considérées également. En effet, le nombre moyen de locuteurs et la durée du locuteur le plus et le moins actif dans chaque show ne sont pas comparables dans chacun des deux domaines. On note ainsi une moyenne de 13 locuteurs pour les show télévisé contre 5 pour les réunions. Ceci pouvait être attendu donné la durée moyenne importante des shows télévisés. En considérant un nombre important de locuteurs, on peut s'attendre à des différences inter-locuteurs plus réduites que pour les réunions et ainsi augmentant les difficultés pour la tache de segmentation et regroupement ne locuteurs.

De plus, on peut remarquer que la durée minimale des locuteurs est beaucoup plus disparate pour le corpus GE que pour la base de données RT'09. La durée moyenne de parole pour le locuteur le plus actif est de 1476 secondes pour GE (contre 535 secondes pour RT'09) et correspond au présentateur dans chacun des cas. La durée moyenne de parole pour le locuteur le moins actif est 7 secondes (cf. 146 secondes pour RT'09) et correspond à l'un des invités minoritaires. Les locuteurs avec si peu de données sont

extrêmement difficiles à détecter, ainsi il est probable que cet aspect du show télévisé amène des difficultés majeures pour la segmentation et le regroupement en locuteurs.

Notons cependant que l'erreur DER globale n'est que peu sensible à ce types de locuteurs dans la mesure où la contribution de chaque locuteur est pondérée en fonction de son temps de parole. De plus, la présence d'un ou deux locuteurs dominants entraîne que d'avantage de locuteurs seront comparativement plus difficiles à détecter, même s' ils ont également un temps de parole significatif.

Finalement, la quantité de parole multi-locuteurs (moyenne de 5 minutes cf. 3 minutes par show), ou 10% (GE) cf. 14% (RT'09) lorsque l'on considère le pourcentage relatif à la quantité totale de parole, montre qu'il y a proportionnellement plutôt moins de parole multi-locuteurs dans la base de données GE que dans le jeu de données RT'09. Comparés à d'autres jeux de données RT, le pourcentage de parole multi-locuteurs peut toujours être considéré comme assez élevé.

Même s'ils comportent moins de parole multi-locuteurs, la nature des shows télévisés présente un challenge unique jusqu'alors jamais vu dans les corpus de réunions. Celui-ci repose principalement sur la présence de musique et d'autres bruits de fond autre que la parole, mais aussi sur une importante disparité dans le temps de parole de chacun des locuteurs, un nombre de locuteurs plus importants avec des silences plus brefs.

B.2.4 Description des Systèmes de Référence

Le système descendant est basé sur le travail du LIA [Fredouille & Evans, 2008], alors que le système ascendant se rapporte au travail d'ICSI [Wooters & Huijbregts, 2008] et plus récemment I2R [Nguyen et al., 2009].

B.2.4.1 Système Ascendant (Top-Down Système)

Le système ascendant décrit ci-après correspond au système officiel utilisé pour la participation LIA-EURECOM lors de la dernière évaluation NIST RT'09 [Fredouille et al., 2009] et a été entièrement développé grâce à la librairie open-source ALIZE [Bonastre et al., 2005]. Le système peut être décomposé en 5 étapes incluant une étape de pré-traitement, puis un processus de détection de présence de parole (Speech Activity Detection - SAD), une étape de segmentation et regroupement en clusters, puis une resegmentation et normalisation des données. Parmi les modifications

réalisées comparées au système utilisé pour l'évaluation NIST RT'07 [Fredouille & Evans, 2008], on peut noter l'utilisation d'un beamforming pour les sets de données MDM (multiple distant microphone) mais aussi des changements significatifs dans l'algorithme de segmentation notamment en termes d'initialisation et de modélisation des locuteurs lesquels sont détaillés dans ce qui suit.

1. **Pré-Traitement**

   Tous-les fichiers audio sont tout d'abord traités avec un filtre de Wiener afin d'en réduire le bruit [Adami et al., 2002b]. Puis, quand plusieurs canaux sont présents (MDM condition), un canal virtuel est créé pour chaque show en utilisant le toolkit BeamformIt v2 [Anguera, 2006; Anguera et al., 2007] avec une fenêtre d'analyse de 500ms capturée toutes les 250ms. Cette étape n'est pas nécessaire pour les conditions SDM. Notons cependant que ceci représente la seule différence pour le système utilisé dans les conditions SDM et MDM et qu'aucun délais inter-canal n'est utilisé pour les autres étapes du processus.

2. **Détection de la Parole - Speech Activity Detection (SAD)**

   Après l'étape de pré-traitement, une étape de détection de la parole (SAD) est accomplie dans le but d'isoler la parole utile dans les données. Ce processus repose sur une chaine de Markov caché (HMM) ou chaque état est associé avec une GMM de 32 composantes entrainée avec un algorithm EM/ML sur une quantité de données importante de speech et non speech provenant des évaluations RT'04 et RT'05[1].

   Le système utilise 12 LFCCs et l'énergie auxquelles sont ajoutées leurs dérivées premières et secondes extraites toutes les 10ms en utilisant une fenêtre de 20ms.

   Tout d'abord, une première segmentation parole/non-parole est menée grâce à un décodage Viterbi utilisant des probabilités équiprobables entre les états de la chaîne de Markov cachée ainsi qu'un buffer[2] de 30 frames. Ensuite les modèles sont adaptés par Maximum A Posteriori (MAP) puis un décodage Viterbi est réalisé à nouveau. Ces deux étapes sont réalisés jusqu'à 10 fois prenant fin en cas d'absence

[1] Notons que ces deux jeux de données utilisés pour l'entrainement des modèles sont totalement indépendants des données de développement ou d'évaluation utilisées par la suite.

[2] Le buffer Viterbi permet de fixer un temps minimum pendant lequel un état est assigné et rend ainsi le système plus stable.

de changement entre deux segmentations consécutives. Finalement, une étape de lissage est exécutée basée sur des règles heuristiques afin de supprimer les rapides transitions speech/non-speech.

3. **Segmentation et Regroupement en Clusters**
   Exécutée directement sur les sorties du détecteur de parole (SAD), l'étape de segmentation et regroupement en clusters peut être considérée comme le cur du programme. Cette dernière se base sur une chaîne de Markov cachée évolutive (E-HMM) [Meignier et al., 2000, 2006] où chaque état a pour but de représenter un locuteur et où les transitions correspondent aux changements de locuteurs. Tous les changements de locuteurs sont envisageables et un buffer Viterbi[2] de 30 frames est utilisé. Ici le signal est caractérisé par 20 LFCCs non normalisées et son énergie calculés toutes les 10ms en utilisant une fenêtre de 20ms.

   Le processus de segmentation et regroupement en locuteur pour chaque show peut être défini comme il suit :

   (a) **Initialisation :** La chaîne de Markov cachée évolutive (E-HMM) a seulement un seul et unique état S0 comme le montre l'Etape 1 de la Figure B.5. Un modèle de monde de 16 Gaussiennes est alors entraîné par EM sur l'ensemble de données de parole. Un processus itératif débute alors, introduisant à chaque itération un nouveau locuteur.

   (b) **Ajout de locuteur :** A la n^{eme} itération un nouveau locuteur Sn est ajouté dans la E-HMM : le segment de parole le plus long d'une durée minimale de 6 secondes est sélectionné parmi l'ensemble des segments associés à S0. Le segment sélectionné est attribué à Sn et est utilisé afin d'estimer une nouvelle GMM par EM.

   (c) **Adaptation/Boucle de Décodage :** L'objectif est de détecter tous les segments appartenant au nouveau locuteur Sn. Tous les modèles de locuteur sont ré-estimés à l'aide d'un ré-alignement Viterbi et d'une ré-estimation des modèles par EM selon la segmentation donnée. Une nouvelle segmentation est alors obtenue. Cette boucle de ré-alignement/apprentissage de modèles est répétée tant qu'un nombre significatifs de changement sont observés dans la segmentation entre deux itérations successives.

(d) **Validation des Modèles de locuteurs et Critère d'Arrêt :** La segmentation actuelle est analysée dans le but de déterminer si le dernier locuteur introduit Sn est pertinent, basé sur des règles heuristiques sur la durée de temps totale assignée au locuteur Sn. Le temps minimum autorisé pour un locuteur est de 10 secondes. Le critère d'arrêt est atteint s'il n'y a plus de segment de parole d'un minimum de 6 secondes associé à S0, sinon le processus retourne à l'étape (b).

La Figure B.5 illustre les 4 étapes décrites précédemment lors de l'addition du modèle de locuteur S1 et S2 (étapes 2 and 3).

4. **Resegmentation**

   L'étape de segmentation et regroupement en locuteur est suivi d'une étape de resegmentation laquelle a pour but d'affiner la segmentation et de supprimer les locuteurs non pertinents (par exemple les locuteurs avec quelques segments seulement). Une nouvelle HMM est générée à partir de la dernière segmentation et une boucle apprentissage des modèles/Décodage Viterbi est lancée. Par rapport à l'étape précédente, ici les modèles de locuteurs sont appris par adaptation MAP à partir d'un modèle de monde universel (UBM) entraîné sur un corpus de reconnaissance du locuteur (Speaker Recognition)[1]. Notons que durant la phase de resegmentation, toutes les frontières (sauf celles correspondantes au parole/non parole) et les étiquettes de chaque segments sont réévaluées.

5. **Normalisation et Resegmentation**

   Finalement une étape de normalisation et reségmentation est réalisée utilisant des vecteur de caractéristiques intégrant 16 LFCCs, l'énergie ainsi que les dérivées premières. Celles-ci sont extraites toutes les 10ms utilisant une fenêtre de 20ms. Les vecteurs sont normalisés, segment de parole par segment de parole afin d'obtenir une moyenne égale à zéro est un écart type d'une unité puis une reségmentation finale est effectuée.

[1] Comparé au données utilisées pour la segmentation et le regroupement en locuteurs, ce corpus utilisé pour la reconnaissance du locuteur contient beaucoup plus de locuteurs (environs 400)

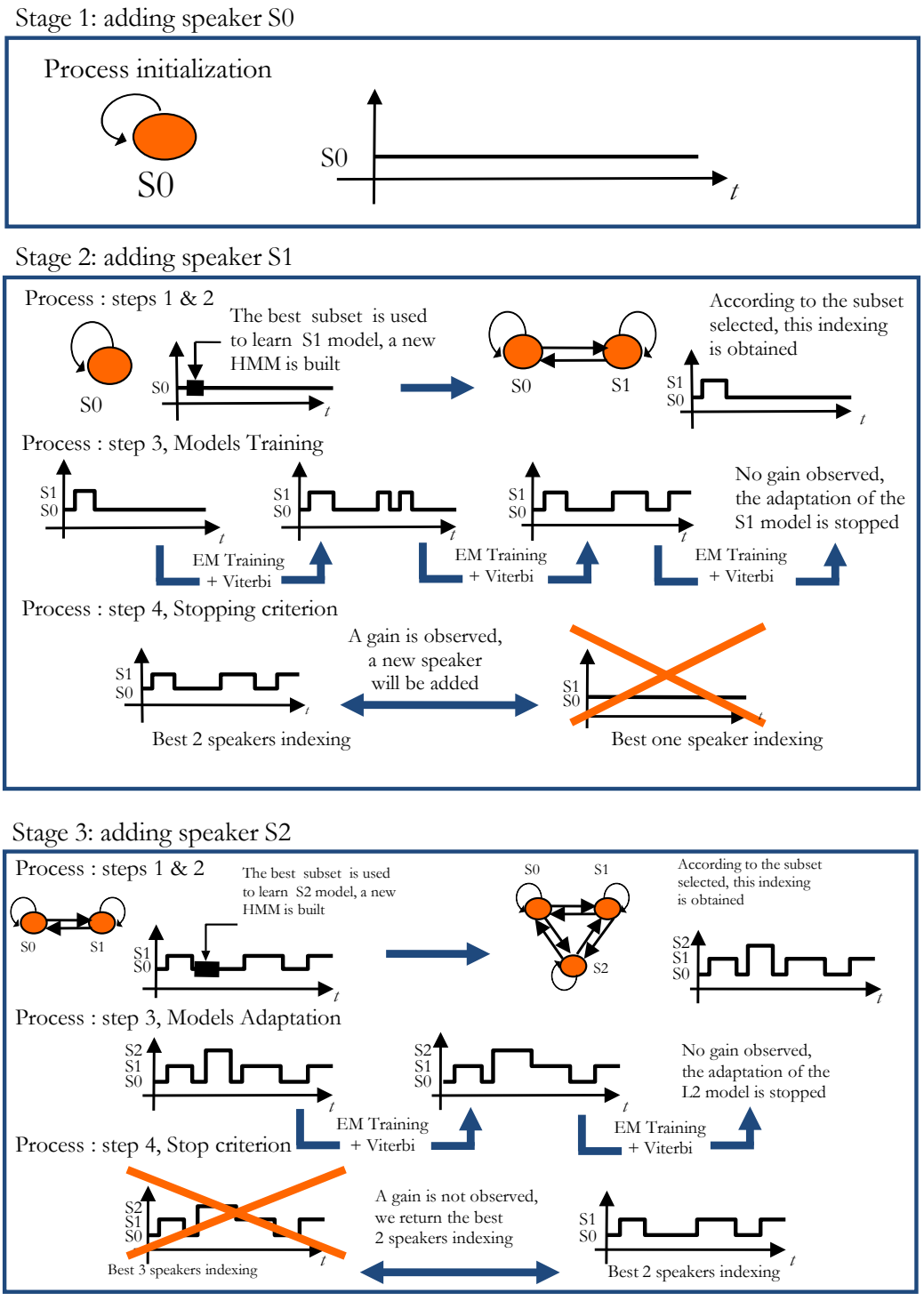


Figure B.5: Système ascendant de segmentation et regroupement en locuteur : cas de 2 locuteurs, image publiée avec l'aimable autorisation de Sylvain Meignier (LIUM) et Corinne Fredouille (LIA)

References

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., & Sivadas, S. (2002a). Qualcomm-ICSI-OGI features for ASR. In *Proc. ICSLP*, volume 1, (pp. 4–7). 18

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., & Sivadas, S. (2002b). Qualcomm-ICSI-OGI features for ASR. In *Proc. ICSLP*, (pp. 21–24). 40, 148

Ajmera, J. (2003). A robust speaker clustering algorithm. In *Proc. ASRU*, (pp. 411–416). 17, 23

Ajmera, J., Lathoud, G., & McCowan, L. (2004). Clustering and segmenting speakers and their locations in meetings. In *Proc. ICASSP*, volume 1, (pp. 605–8). 25

Ajmera, J., McCowan, I., & Bourlard, H. (2004). Robust speaker change detection. *IEEE Signal Process. Letters*, *11*, 649–651. 21, 47

Anastasakos, T., Mcdonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In *in Proc. ICSLP*, (pp. 1137–1140). 108

Anguera, X. (2006). BeamformIt (the fast and robust acoustic beamformer). http://www.xavieranguera.com/beamformit/. 18, 34, 40, 141, 148

Anguera, X. & Bonastre, J. (2010). A novel speaker binary key derived from anchor models. In *Proc. Interspeech*. 16

Anguera, X. & Bonastre, J. (2011). Fast speaker diarization based on binary keys. In *Proc. ICASSP*. 16

Anguera, X., Bozonnet, S., Evans, N. W. D., Fredouille, C., Friedland, G., & Vinyals, O. (2011). Speaker diarization : A review of recent research. *"IEEE Transactions On Audio, Speech, and Language Processing" (TASLP), special issue on "New Frontiers in Rich Transcription", February 2012, Volume 20, NÂ2, ISSN: 1558-7916*. 11, 51, 123

Anguera, X. & Hernando, J. (2004). Evolutive speaker segmentation using a repository system. In *Proc. Interspeech*. 21

Anguera, X., Wooters, C., Anguilo, M., & Nadeu, C. (2006). Hybrid speech/non-speech detector applied to speaker diarization of meetings. In *Speaker Odyssey workshop*, Puerto Rico, USA. 20

Anguera, X., Wooters, C., & Hernando, J. (2005). Speaker diarization for multi-party meetings using acoustic fusion. In *Proc. ASRU*, (pp. 426–431). 21, 34, 141

Anguera, X., Wooters, C., & Hernando, J. (2006a). Friends and enemies: A novel initialization for speaker diarization. In *Proc. ICSLP*, Pittsburgh, USA. 17

Anguera, X., Wooters, C., & Hernando, J. (2006b). Purity algorithms for speaker diarization of meetings data. In *Proc. ICASSP*. 17, 24, 63

Anguera, X., Wooters, C., & Hernando, J. (2006c). Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *Proc. ICSLP*, Pittsburgh, USA. 14, 17

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *IEEE TASLP*, *15*(7), 2011–2023. 18, 25, 40, 148

Anguera, X., Wooters, C., Peskin, B., & Aguilo, M. (2005). Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh. Springer. 18, 20, 23

Aronowitz, H. (2007). Trainable speaker diarization. In *Proc. Interspeech*, (pp. 1861–4). 22

Barras, C., Zhu, X., Meignier, S., & Gauvain, J. (2004). Improving speaker diarisation. In *Proc. DARPA RT04*. 22

Ben, M., Betser, M., Bimbot, F., & Gravier, G. (2004). Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *Proc. ICSLP, Jeju Island, Korea*. 21, 22

Boakye, K. (2008). *Audio Segmentation for Meetings Speech Processing*. PhD thesis, University of California at Berkeley. 27

Boakye, K., Trueba-Hornero, B., Vinyals, O., & Friedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. *Proc. ICASSP*, 4353–4356. 27, 124

Bonastre, J.-F., Wils, F., & Meignier, S. (2005). ALIZE, a free toolkit for speaker recognition. In *Proc. ICASSP'05*, volume 1, (pp. 737–740)., Philadelphia, USA. 40, 44, 95, 147

Bozonnet, S., Evans, N., Anguera, X., Vinyals, O., Friedland, G., & Fredouille, C. (2010). System output combination for improved speaker diarization. In *Proc. Interspeech*. 8, 83, 91, 96, 98, 137

Bozonnet, S., Evans, N., Fredouille, C., Wang, D., & Troncy, R. (2010). An integrated top-down/bottom-up approach to speaker diarization. In *Proc. Interspeech*. 8, 83, 137

Bozonnet, S., Evans, N. W. D., & Fredouille, C. (2010). The LIA-EURECOM RT'09 Speaker Diarization System: enhancements in speaker modelling and cluster purification. In *Proc. ICASSP*, Dallas, Texas, USA. xiv, 15, 17, 63, 67, 91

Bozonnet, S., Vallet, F., Evans, N. W. D., Essid, S., Richard, G., & Carrive, J. (2010). A Multimodal approach to initialisation for top-down speaker diarization of television shows. In *EUPSICO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark*. 38, 145

Bozonnet, S., Wang, D., Evans, N. W. D., & Troncy, R. (2011). Linguistic influences on bottom-up and top-down clustering for speaker diarization. In *ICASSP 2011, 36th International Conference on Acoustics, Speech and Signal Processing, May 22-27, 2011, Prague, Czech Republic*, Prague, CZECH REPUBLIC. 6, 7, 134, 136

Burget, L., Fapso, M., Hubeika, V., Glembek, O., Karafit, M., Kockmann, M., Matejka, P., Schwarz, P., & Cernock, J. (2009). But system for nist 2008 speaker recognition evaluation. In *Proc. Interspeech 2009*, number 9, (pp. 2335–2338). International Speech Communication Association. 83

Campbell, N. & Suzuki, N. (2006). Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus. In *Workshop Programme*, volume 10. 29

Çetin, O. & Shriberg, E. (2006). Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. In *Proc. ICASSP*, (pp. 357–360). Toulouse, France. 27

Chen, I., Cheng, S., & Wang, H. (2010). Phonetic subspace mixture model for speaker diarization. In Kobayashi, T., Hirose, K., & Nakamura, S. (Eds.), *INTERSPEECH*, (pp. 2298–2301). ISCA. 105

Chen, S. S. & Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, (pp. 127–132)., Lansdowne, Virginia, USA. 17, 21, 46, 47

Chen, T. & Rao, R. R. (1996). Cross-modal Prediction in Audio-visual Communication. In *Proc. ICASSP*, volume 4, (pp. 2056–2059). 28

Delacourt, P. & Wellekens, C. (2000). DISTBIC : a speaker-based segmentation for audio data indexing. *Speech Communication*, 111–126. 21

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm. In *Journal of Acoustical Society of America JASA*, volume 39, (pp. 1–38). 106

Digalakis, V., Rtischev, D., Neumeyer, L., & Sa, E. (1995). Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, *3*, 357–366. 107

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification*. Wiley. 111

El-Khoury, E., Senac, C., & Meignier, S. (2008). Speaker diarization: combination of the LIUM and IRIT systems. In *Internal report*. 30, 85

El-Khoury, E., Senac, C., & Pinquier, J. (2009). Improved speaker diarization system for meetings. In *Proc. ICASSP*, Taipei, Taiwan. 20

Ellis, D. P. W. & Liu, J. C. (2004). Speaker turn detection based on between-channels differences. In *Proc. ICASSP*. 25, 34, 141

Evans, N., Bozonnet, S., Wang, D., Fredouille, C., & Troncy, R. (2012). A comparative study of bottom-up and top-down approaches to speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(2), 382 –392. 7, 8, 73, 83, 98, 136, 137

Evans, N. W. D., Fredouille, C., & Bonastre, J.-F. (2009). Speaker diarization using unsupervised discriminant analysis of inter-channel delay features. In *Proc. ICASSP*, (pp. 4061–4064). 26, 34, 141

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, *1*(2), 209–230. 30

Fisher, J. W. & Darrell, T. (2004). Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, *6*(3), 406–413. 28

Fisher, J. W., Darrell, T., Freeman, W. T., & Viola, P. A. (2000). Learning joint statistical models for audio-visual fusion and segregation. In *Proc. NIPS*, (pp. 772–778). 28

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2008). An HDP-HMM for systems with state persistence. In *Proc. ICML*. 30

Fredouille, C., Bozonnet, S., & Evans, N. W. D. (2009). The LIA-EURECOM RT'09 Speaker Diarization System. In *RT'09, NIST Rich Transcription Workshop*. 15, 17, 23, 40, 147

Fredouille, C. & Evans, N. (2007). The influence of speech activity detection and overlap on the speaker diarizationfor meeting room recordings. In *Proc. Interspeech'07*. 57

Fredouille, C. & Evans, N. W. D. (2008). The LIA RT07 speaker diarization system. In Stiefelhagen, Bowers, F. (Ed.), *Lecture notes in Computer Science - Multimodal Technologies for Perception of Humans*, volume 4625/2008, (pp. 520–532). Springer. 15, 17, 20, 38, 40, 41, 42, 144, 147, 148

Fredouille, C., Moraru, D., Meignier, S., Besacier, L., & Bonastre, J.-F. (2004). The NIST 2004 spring Rich Transcription evaluation: Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation. In *NIST 2004 Spring Rich Transcrition Evaluation Workshop*, Montreal, Canada. 17, 38, 145

Fredouille, C. & Senay, G. (2006). Technical Improvements of the E-HMM Based Speaker Diarization System for Meeting Records. In *Proc. MLMI Third International Workshop, Bethesda, MD, USA, revised selected paper*, (pp. 359–370)., Berlin, Heidelberg. Springer-Verlag. 20

Friedland, G., Ching, J., & Janin, A. (2010). Parallelizing speaker-attributed speech recognition for meeting browsing. In *Proc. IEEE International Symposium on Multimedia*, Taichung, Taiwan. 16

Friedland, G., Hung, H., & Yeo, C. (2009). Multimodal speaker diarization of real-world meetings using compressed-domain video features. In *Proc. ICASSP*, (pp. 4069–4072). 29

Friedland, G. & Vinyals, O. (2008). Live speaker identification in conversations. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, (pp. 1017–1018)., New York, NY, USA. ACM. 23

Friedland, G., Vinyals, O., Huang, Y., & Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *IEEE TASLP*, *17*(5), 985–993. 23, 26, 78, 91, 112

Friedland, G., Yeo, C., & Hung, H. (2009). Visual speaker localization aided by acoustic models. In *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, (pp. 195–202)., New York, NY, USA. ACM. 29

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, *12*, 75–98. 107

Gales, M. & Woodland, P. (1996). Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, *10*, 249–264. 107

Gangadharaiah, R., Narayanaswamy, B., & Balakrishnan, N. (2004). A novel method for two speaker segmentation. In *Proc. ICSLP*, Jeju S. Korea. 21

Ghahramani, Z. & Jordan, M. I. (1997). Factorial hidden markov models. *Mach. Learn.*, *29*, 245–273. 28

Gish, H. & Schmidt, M. (1994). Text independent speaker identification. In *IEEE Signal Processing Magazine*, (pp. 18–32). 17

Griffiths, L. J. & Jim, C. W. (1982). An alternative approach to linearly con- strained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, *AP-30*, 27–34. 18

Gupta, V., Kenny, P., Ouellet, P., Boulianne, G., & Dumouchel, P. (2007). Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations. In *Signal Processing letters, IEEE*, (pp. 1040–1043). 30, 85

Han, K. J., Kim, S., & Narayanan, S. S. (2008). Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE TASLP*, *16*(8), 1590–1601. 48, 54, 61, 98

Han, K. J. & Narayanan, S. S. (2008). Novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering. In *Proc. ICASSP*, (pp. 4373–4376). 22

Hershey, J. & Olsen, P. (2007). Approximating the kullback leibler divergence between gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, (pp. IV–317 – IV–320). 111

Hinton, G. E. & van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, COLT '93, (pp. 5–13)., New York, NY, USA. ACM. 16

Huang, Y., Vinyals, O., Friedland, G., Muller, C., Mirghafori, N., & Wooters, C. (2007). A fast-match approach for robust, faster than real-time speaker diarization. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, (pp. 693–698)., Kyoto, Japan. 16

Huijbregts, M., van Leeuwen, D., & Wooters, C. (2012). Speaker diarization error analysis using oracle components. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(2), 393 –403. xiii, 54, 55, 57, 59

Huijbregts, M., Van Leeuwen, D. A., & de Jong, F. (2009). Speech overlap detection in a two-pass speaker diarization system. In *in ISCA 2009*, (pp. 1063–66). 124

Huijbregts, M., van Leeuwen, D. A., & de Jong, F. M. G. (2009). The majority wins: a method for combining speaker diarization systems. In *INTERSPEECH*, (pp. 924–927). ISCA. 85

Huijbregts, M. & Wooters, C. (2007). The blame game: performance analysis of speaker diarization system components. In *INTERSPEECH*, (pp. 1857–60). xiii, 53, 54, 60, 123

Hung, H. & Friedland, G. (2008). Towards Audio-Visual Online Diarization Of Participants In Group Meetings. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2*, Marseille France. 23, 29

Hung, H., Huang, Y., Yeo, C., & Gatica-Perez, D. (2008). Associating audio-visual activity cues in a dominance estimation framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Human Communicative Behavior*, Ankorage, Alaska. 29

Imseng, D. & Friedland, G. (2009). Robust speaker diarization for short speech recordings. In *Proc. of the IEEE workshop on Automatic Speech Recognition and Understanding*, (pp. 432–437). 27, 91

Imseng, D. & Friedland, G. (2010). Tuning-Robust Initialization Methods for Speaker Diarization. *IEEE TALSP*. 26

Istrate, D., Fredouille, C., Meignier, S., Besacier, L., & Bonastre, J. (2005). NIST RT05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings. In *NIST 2005 Spring Rich Transcrition Evaluation Workshop*, Edinburgh, UK. 18

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., & Wrede, B. (2004). The ICSI meeting project: Resources and research. In *Proc. ICASSP Meeting Recognition Workshop*. 17

Jin, Q., Laskowski, K., Schultz, T., & Waibel, A. (2004). Speaker segmentation and clustering in meetings. In *Proc. ICSLP*, Jeju S. Korea. 17, 21

Jothilakshmi, S., Ramalingam, V., & Palanivel, S. (2009). Speaker diarization using autoassociative neural networks. *Engineering Applications of Artificial Intelligence*, *22(4-5)*. 12, 23

K-Space. The European K-Space Network Of Excellence. http://www.k-space.eu/. 38, 145

Kenny, P. (2008). Bayesian analysis of speaker diarization with eigenvoice priors. Technical report, CRIM, Montreal. 16

Kingsbury, B. E. D., Morgan, N., & Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. *Speech Commun.*, *25*(1-3), 117–132. 27

Kotti, M., Benetos, E., & Kotropoulos, C. (2008). Computationally efficient and robust bic-based speaker segmentation. *IEEE TASLP*, *16(5)*. 12, 23

Lathoud, G. & Cowan, I. A. M. (2003). Location based speaker segmentation. In *Proc. ICASSP*, volume 1, (pp. 176–179). 25

Leggetter, C. & P., W. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, 171–185. 106

Leggetter, C. & Woodland, P. (1995). Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. ARPA Spoken Language Technology Workshop*, (pp. 110–115). Morgan Kaufmann. 107

Li, R. & Schultz, T. (2009). Improving speaker segmentation via speaker identification and text segmentation. 21

Liu, D. & Kubala, F. (September 1999). Fast speaker change detection for broadcast news transcription and indexing. In *Proc. EuroSpeech-99*, (pp. 1031–1034). 22

Lu, L. & Zhang, H. (2002). Real-time unsupervised speaker change detection. In *16th International Conference on Pattern Recognition*, volume 2, (pp. 358–361 vol.2). 21

Lu, L., Zhang, H., & Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE TSAP*, *10*, 504–516. 21

Luque, J., Anguera, X., Temko, A., & Hernando, J. (2008). Speaker diarization for conference room: The UPC RT07s evaluation system. In *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, (pp. 543–553)., Berlin, Heidelberg. Springer-Verlag. 23

Malegaonkar, A., Ariyaeeinia, A., Sivakumaran, P., & Fortuna, J. (2006). Unsupervised speaker change detection using probabilistic pattern matching. *Signal Processing Letters, IEEE*, *13*(8), 509–512. 21

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus. In *Proc. Measuring Behavior*. 17

McEachern, S. (1994). Estimating normal means with a conjugate style dirichlet process prior. In *Communications in Statistics: Simulation and Computation*, volume 23, (pp. 727–741). 16

McNeill, D. (2000). *Language and Gesture*. Cambridge University Press New York. 28

Meignier, S., Bonastre, J.-F., Fredouille, C., & Merlin, T. (2000). Evolutive HMM for speaker tracking system. In *ICASSP'00*, Istanbul, Turkey. 41, 149

Meignier, S., Bonastre, J.-F., & Igounet, S. (2001). E-HMM approach for learning and adapting sound models for speaker indexing. In *Proc. Odyssey Speaker and Language Recognition Workshop*, (pp. 175–180)., Chania, Creete. 15, 23

Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.-F., & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization. In *CSL, selected papers from the Speaker and Language Recognition Workshop (Odyssey'04)*, (pp. 303–330). 29, 41, 79, 84, 85, 149

Moraru, D., Ben, M., & Gravier, G. (2005). Experiments on speaker tracking and segmentation in radio broadcast news. In *Proc. ICSLP*. 22

Mori, K. & Nakagawa, S. (2001). Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition. In *Proc. ICASSP*, (pp. 413–416). 21

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., & Rochet, C. (2007). *The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms*, volume 41. Language Resources and Evaluation. 17

Nguyen, T. H., Chng, E. S., & Li, H. (2008). T-test distance and clustering criterion for speaker diarization. In *Proc. Interspeech*, Brisbane, Australia. 14, 48

Nguyen, T. H., Sun, H., Zhao, S. K., Khine, S. Z. K., Tran, H. D., Ma, T. L. N., Ma, B., Chng, E. S., & Li, H. (2009). The IIR-NTU Speaker Diarization Systems for RT 2009. In *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*. 14, 17, 40, 44, 47, 48, 63, 64, 86, 91, 147

Ning, H., Xu, W., Gong, Y., & Huang, T. (2006). Improving speaker diarization by cross em refinement. *Multimedia and Expo, IEEE International Conference on*, *0*, 1901–1904. 24

NIST (2002). The NIST year 2002 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf. 33, 139

NIST (2003). The rich transcription spring 2003 (RT-03S) evaluation plan. http://www.nist.gov/speech/tests/rt/rt2003/spring/docs/rt03-spring-eval-plan-v4.pdf. (Version 4, Updated 02/25/2003). 33, 139

NIST (2004). Spring 2004 (RT-04S) Rich Transcription meeting recognition evaluation-plan. http://www.itl.nist.gov/iad/894.01/tests/rt/rt2004/spring/documents/ rt04s-meeting-eval-plan-v1.pdf. 33, 139

NIST (2006). Spring 2006 (RT'06S) Rich Transcription meeting recognition evaluationplan. http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf. 33, 139

NIST (2007). Spring 2007 (RT'07S) Rich Transcription meeting recognition evaluationplan. http://nist.gov/speech/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf. 33, 44, 91, 139

NIST (2009). The NIST Rich Transcription 2009 (RT'09) evaluation. http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf. 33, 38, 44, 63, 70, 74, 91, 139, 145

Nock, H. J., Iyengar, G., & Neti, C. (2003). Speaker Localisation Using Audio-Visual Synchrony: An Empirical Study. *Lecture Notes in Computer Science*, *2728*, 565–570. 28

Noulas, A. & Krose, B. J. A. (2007). On-line multi-modal speaker diarization. In *ICMI '07: Proceedings of the ninth international conference on Multimodal interfaces*, (pp. 350–357)., New York, NY, USA. ACM. 28

Noulas, A. K., Englebienne, G., & Krose, B. J. A. (2009). Multimodal speaker diarization. *Computer Vision and Image Understanding*. 28

Nwe, T. L., Sun, H., Li, H., & Rahardja, S. (2009). Speaker diarization in meeting audio. In *Proc. ICASSP*, Taipei, Taiwan. 20

Otterson, S. & Ostendorf, M. (2007). Efficient use of overlap information in speaker diarization. In *Proc. ASRU*, (pp. 686–6)., Kyoto, Japan. 27

Pardo, J., Anguera, X., & Wooters, C. (2006a). Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features And Inter-Channel Time Differences. *Proceedings of Interspeech*. 23, 25

Pardo, J., Anguera, X., & Wooters, C. (2007). Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transaction on Computers*, *56*(9), 1212–1224. 25

Pardo, J. M., Anguera, X., & Wooters, C. (2006b). Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features And Inter-Channel Time Differences. In *Proceedings of Interspeech*. 25

Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Proc. ICASSP*, (pp. 2017–2020). 28

Ramirez, J., Girriz, J. M., & Segura, J. C. (2007). Voice activity detection. fundamentals and speech recognition system robustness. In Grimm, M. & Kroschel, K. (Eds.), *Robust Speech Recognition and Understanding*, (pp. 460)., Vienna, Austria. 19

Rao, R. & Chen, T. (1996). Exploiting audio-visual correlation in coding of talking head sequences. *International Picture Coding Symposium*. 28

Rentzeperis, A., Stergious, A., Boukis, C., Pnevmatikakis, A., & Polymenakos, L. (2006). The 2006 Athens Information Technology Speech activity detection and speaker diarization systems. In *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA, revised selected paper*, (pp. 385–395)., Berlin, Heidelberg. Springer-Verlag. 20

Reynolds, D., Kenny, P., & Castaldo, F. (2009). A study of new approaches to speaker diarization. In *Proc. Interspeech*. ISCA. 16

Rougui, J. E., Rziza, M., Aboutajdine, D., Gelgon, M., & Martinez, J. (2006). Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In *Proc. ICASSP*, volume 5. 14, 21, 22

Seltzer, M. L., Raj, B., & Stern, R. M. (2004). Likelihood maximizing beam- forming for robust hands-free speech recognition. *IEEE Transactions on Speech and Audio Processing*, *12*, 489–498. 18

Shriberg, E. (2007). Higher-Level Features in Speaker Recognition. In C. Müller (Ed.), *Speaker Classification I*, volume 4343 of *Lecture Notes in Artificial Intelligence*. Heidelberg: Springer. 26

Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversations. In *Proc. Eurospeech 2001*, (pp. 1359–1362). Aalborg, Denmark. 27

Shrikanth, S. H. & Narayanan, K. J. (2008). Agglomerative hierarchical speaker clustering using incremental gaussian mixture cluster modeling. In *Interspeech'08, Brisbane, Australia*, (pp. 20–23). 22

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA Speech Recognition Workshop*, (pp. 97–99). 22, 80

Siracusa, M. R. & Fisher, J. W. (2007). Dynamic dependency tests for audio-visual speaker association. In *Proc. ICASSP*. 28

Siu, M.-H., Yu, G., & Gish, H. (1991). Segregation of speakers for speech recognition and speaker identification. In *ICASSP 91*. 21

Sun, H., Nwe, T. L., Ma, B., & Li, H. (2009). Speaker diarization for meeting room audio. In *Proc. Interspeech'09*. 20

Tamura, S., Iwano, K., & Furui, S. (2004). Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images. *Real World Speech Processing*. 28

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581. 30

Temko, A., Macho, D., & Nadeu, C. (2007). Enhanced SVM Training for Robust Speech Activity Detection. In *Proc. ICASSP*, Hawai, USA. 20

Tranter, S. E. (2005). Two-way cluster voting to improve speaker diarisation performance. In *in Proc. ICASSP*, (pp. 753–756). 85

Trueba-Hornero, B. (2008). Handling overlapped speech in speaker diarization. Master's thesis, Universitat Politecnica de Catalunya. 27

Truong, B. T. & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *3*. 1, 129

Tsai, W., Cheng, S., & Wang, H. (2004). Speaker clustering of speech utterances using a voice characteristic reference space. In *Proc. ICSLP*. 14

Vajaria, H., Islam, T., Sarkar, S., Sankar, R., & Kasturi, R. (2006). Audio segmentation and speaker localization in meeting videos. *18th International Conference on Pattern Recognition (ICPR'06)*, *2*, 1150–1153. 29

Valente, F. (2005). *Variational Bayesian methods for audio indexing*. PhD thesis, Thesis. 16

Valente, F. (2006). Infinite models for speaker clustering. In *International Conference on Spoken Language Processing*. IDIAP-RR 06-19. 30

van Leeuwen, D. A. & Huijbregts, M. (2007). The ami speaker diarization system for nist rt06s meeting data. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, (pp. 371–384)., Berlin, Germany. Springer Verlag. 21

Van Leeuwen, D. A. D. A. & Konečný, M. (2008). Progress in the AMIDA Speaker Diarization System for Meeting Data. In *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, (pp. 475–483)., Berlin, Heidelberg. Springer-Verlag. 20, 23

Vandecatseye, A., Martens, J., Neto, J., Meinedo, H., Garcia-Mateo, C., Dieguez, J., Mihelic, F., Zibert, J., Nouza, J., David, P., Pleva, M., Cizmar, A., Papageorgiou, H., & Alexandris, C. (2004). The COST278 pan-european broadcast news database. In *Proc. LREC*, volume 4, (pp. 873–876)., Lisbon, Portugal. European Language Resources Association (ELRA). 21

Vijayasenan, D., Valente, F., & Boulard, H. (2009). An information theoretic approach to speaker diarization of meeting data. *IEEE TASLP*, *17*, 1382–1393. 15

Vijayasenan, D., Valente, F., & Bourlard, H. (2007). Agglomerative information bottleneck for speaker diarization of meetings data. In *Proc. ASRU*, (pp. 250–255). 15, 22

Vijayasenan, D., Valente, F., & Bourlard, H. (2008). Combination of agglomerative and sequential clustering for speaker diarization. In *Proc. ICASSP*, (pp. 4361–4364)., Las Vegas, USA. 29, 85

Žibert, J., Pavesić, N., & Mihelič, F. (2006). Speech/non-speech segmentation based on phoneme recognition features. *EURASIP J. Appl. Signal Process.*, *2006*, 47–47. 105

Wainwright, M. J. & Jordan, M. I. (2003). Variational inference in graphical models: The view from the marginal polytope. In *Forty-first Annual Allerton Conference on Communication, Control, and Computing, Urbana-Champaign, IL*. 16

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley. 18

Woelfel, M. & McDonough, J. (2009). *Distant Speech Recognition*. Wiley Eds. 18

Wölfel, M., Yang, Q., Jin, Q., & Schultz, T. (2009). Speaker Identification using Warped MVDR Cepstral Features. In *Proc. of Interspeech*. 26

Wooters, C., Fung, J., Peskin, B., & Anguera, X. (2004). Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY. 19, 20

Wooters, C. & Huijbregts, M. (2008). The ICSI RT07s speaker diarization system. *Multimodal Technologies for Perception of Humans*, *4625*, 509–519. 14, 20, 23, 40, 44, 91, 147

Zhang, C., Yin, P., Rui, Y., Cutler, R., & Viola, P. (2006). Boosting-Based Multimodal Speaker Detection for Distributed Meetings. *IEEE International Workshop on Multimedia Signal Processing (MMSP)*. 28

Zhu, X., Barras, C., Lamel, L., & Gauvain, J.-L. (2006). Speaker diarization: From broadcast news to lectures. In *MLMI*, (pp. 396–406). 22

Zhu, X., Barras, C., Lamel, L., & Gauvain, J.-L. (2008). Multi-stage Speaker Diarization for Conference and Lecture Meetings. In *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, (pp. 533–542)., Berlin, Heidelberg. Springer-Verlag. 12, 20, 23

Zochová, P. & Radová, V. (2005). Modified DISTBIC algorithm for speaker change detection. In *Proc. 9th Eur. Conf. Speech Commun. Technol.*, (pp. 3073–3076)., Bonn. Universitat Bonn. 22

May 14th 2012
version 5