Speaker diarization of spontaneous meeting room conversations

Thèse n. 6542 2015
présenté le 19 Jan 2015
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE L'IDIAP
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
pour l'obtention du grade de Docteur ès Sciences
par

Sree Harsha Yella

acceptée sur proposition du jury:

Prof. Anja Skrivervik, président du jury
Prof. Hervé Bourlard, directeur de thèse
Dr. Andreas Stolcke, rapporteur
Dr. Xavier Anguera, rapporteur
Dr. Jean-Marc Vesin, rapporteur

Lausanne, EPFL, 2015

To my parents

Acknowledgements

First and foremost I would like to thank my thesis advisors Fabio Valente and Hervé Bourlard for giving me the opportunity to do Ph.D. at Idiap. Working with Fabio in the early stages of my Ph.D. was a great experience. I am grateful for his time and patience in teaching me the basics of research. His dedication and passion towards research are contagious, and have been a constant source of motivation for me during some tough times in the pursuit of Ph.D.

I am also grateful to Hervé for his insights, comments and constructive criticism on my work. His hard work over two decades has resulted in turning Idiap into a truly world class research organization which enables a number of researchers and students like me to have their 'Eureka!' moments. I am thankful to him for encouraging me to collaborate with other researchers and allowing me to go on internships, which have given me an invaluable experience.

I thank the members of the thesis committee Dr. Xavier Anguera, Dr. Andreas Stolcke, Dr. Jean-Marc Vesin and Prof. Anja Skrivervik for their comments on the thesis. I am thankful to them for taking time out of their busy schedule and reviewing my thesis. I also thank Marc Ferras and Ashtosh Sapru for sharing ideas and spending their valuable time in many discussions on my work. This work has also benefitted from the discussions, comments and help from the researchers at the speech group, Petr Motlicek, Phil Garner and Mathew Magimai Doss. I would also like to thank members of the speech group past and present for their valuable feedback on the work. Special thanks to Pierre-Edward for his help in writing the French abstract of the thesis. I express my gratitude to Swiss National Science Foundation (SNSF) for funding my Ph.D.

I thank Malcolm and Andreas for their mentorship during my internship at Microsoft Research, CA. This gave me an opportunity to learn from and interact with world class researchers and the work done during this internship has also become a part of this thesis. I also thank Xavi and Jordi for giving me an opportunity to do an internship at Telefonica Research, Barcelona. I had a great time working with you and thoroughly enjoyed my stay in Barcelona.

I am greatly indebted to the system administrators at Idiap, Frank, Bastien, Norbert and Louie-Marie for helping with numerous system related issues and for providing excellent computational resources at Idiap. I would like to especially thank Louie-Marie for setting up a system with a GPU dedicated to me, which helped a lot in running experiments in the later stages of my thesis. I also thank the Idiap secretaries Nadine and Sylvie for taking care of many

issues related to permits, apartments and in general with most of the paperwork related to VISAs, travel and life in Swiss.

Doing Ph.D. at Idiap and EPFL gave me a chance to meet wonderful people who have made this journey memorable. Many thanks to Ashtosh, Deepu, Jagan, Hari, David, Venky, Abhilasha, Ramya, Murali, Dinesh, Kavitha, Lakshmi, Saheer, Mathew, Francina, Anindhya, Marco, Ivana, Laurent, James, Petr, Blaise, Pierre-Edward, Manuel, Srikanth, Pranay, Marc, Phil, Somi, Chary, Sujatha and Ganga for their friendship and camaraderie over the past four years.

Last but not the least, I would like to thank my parents Usha and Dileep for their unconditional love and care. I would not have made it this far without their support.

Martigny, 24th January, 2015 Sree Harsha Yella

Abstract

Speaker diarization is the task of identifying "who spoke when" in an audio stream containing multiple speakers. This is an unsupervised task as there is no a priori information about the speakers. Diagnostical studies on state-of-the-art diarization systems have isolated three main issues with the systems; overlapping speech, effects of background noise and speech/non-speech detection errors on clustering, and signficant performance variance between different systems. In this thesis we focuss on addressing these issues in diarization.

We propose new features based on structure of a conversation such as silence and speaker change statistics for overlap detection. The features are estimated from a long-term context (3-4 seconds) and are used to estimate the probability of overlap at a given instant. These probabilities are later incorporated into acoustic feature based overlap detector as prior probabilities. Experiments on several meeting corpora reveal that overlap detection is improved significantly by the proposed method and this consequently reduces the diarization error.

To address the issues arising from background noise, errors in speech/non-speech detection and capture speaker discriminative information in the signal, we propose two methods. In the first method, we propose Information Bottleneck with Side Information (IBSI) based diarization to supress artefacts of background noise and non-speech segments introduced into clustering. In the second method, we show that the phoneme transcript of a given recording carries useful information for speaker diarization. This obervation was used in estimation of phoneme background model which is used for diarization in Information Bottleneck (IB) framework. Both the methods achieve significant reduction in error on various meeting corpora.

We train different artificial neural network (ANN) architectures to extract speaker discriminant features and use these features as input to speaker diarization systems. The ANNs are trained to perform related tasks such as speaker comparison, speaker classification and auto encoding. The bottleneck layer activations from these networks are used as features for speaker diarization. Experiments on different meeting corpora revealed that combination of MFCCs and ANN features reduces the diarization error.

To address the issue of performance variations across different sytems, we propose feature level combination of HMM/GMM and IB diarization systems. The combination does not require any changes to the original systems. The output of IB system is used to generate features which when combined with MFCCs in a HMM/GMM system reduce diarization error.
Key Words: Speaker diarization, meeting room conversations, conversational speech, overlapping speech, side information, phoneme background model, artificial neural networks,

bottleneck features, system combination

Résumé

La diarization – aussi appelée segmentation et regroupement – de locuteur est la tâche d'identifier "qui a parlé quand" dans un flux audio contenant plusieurs locuteurs. Il s'agit d'une tâche non supervisée car il n'y a a priori aucune information sur les locuteurs. Des études diagnostiques sur l'état de l'art des systèmes de diarization ont isolé trois problèmes principaux pour ces systèmes : la parole se recouvrant (chevauchement), les effets du bruit ambiant et des erreurs de détection de parole/absence de parole sur le regroupement, et une variance des performances des différents systèmes significative. Dans cette thèse, nous nous intéressons à ces trois problèmes dans le contexte de la diarization.

Nous proposons de nouveaux paramètres basés sur la structure des conversations, comme les silences et les statistiques de changement de locuteur pour la détection de chevauchement (de locuteurs). Ces paramètres sont estimés en prenant en compte un contexte à long terme (3-4 secondes) et sont utilisés pour estimer la probabilité de chevauchement à un instant donné. Ces probabilités sont ensuites incorporées en tant que probabilités a priori dans un détecteur de chevauchement basé sur des paramètres acoustiques. Des expériences sur plusieurs corpus de réunions révèlent que la méthode proposée améliore significativement la détection de chevauchement et réduit par conséquent l'erreur de diarization.

Pour s'attaquer aux problèmes liés au bruit ambiant, erreurs de détection de parole/absence de parole et pour capturer les informations discriminant les locuteurs dans le signal, nous proposons deux méthodes. Dans la première approche, nous proposons un système de diarization basé sur le principe du goulot d'étranglement de l'information avec information additionnelle (IBSI pour Information Bottleneck with Side Information) pour supprimer les artefacts de bruit ambiant et de segments sans parole introduits dans le regroupement. Dans la seconde méthode, nous montrons que la transcription phonétique d'un enregistrement donné contient des informations utiles pour la diarization de locuteurs. Cette observation permet d'estimer le modèle de fond basé sur les phonèmes, lequel est utilisé pour la diarization dans le cadre du goulot d'étranglement de l'information (IB pour Information Bottleneck). Ces deux méthodes résultent en une réduction significative de l'erreur sur différents corpus de réunions.

Nous entraînons différentes architectures de réseaux de neurones artificiels (ANN pour Artificial Neural Network) pour extraire des paramètres discriminant les locuteurs. Ces paramètres sont alors utilisés comme données d'entrée pour les systèmes de diarization de locuteurs. Les ANNs sont entraînés pour effecteur des tâches liées comme la comparaison de locuteurs, la classification de locuteurs et l'auto encodage. Les activations des couches d'étranglement de

ces réseaux sont utilisées comme paramètres pour la diarization de locuteur. Des expériences sur différents corpus de réunions ont montré que la combinaison de coefficients spectraux MFCC (Mel-Frequency Cepstral Coefficients) avec les paramètres des ANN réduit l'erreur de diarization.
Pour répondre au problème des variations de performances entre les différents systèmes, nous proposons une combinaison des systèmes de diarization HMM/GMM et IB au niveau des paramètres. La combinaison se fait au niveau des paramètres et ne requiert aucun changement sur les systèmes originaux. La sortie du système IB est utilisée pour générer les paramètres qui, une fois combinés avec les MFCCs dans un système HMM/GMM, réduisent l'erreur de diarization.
Mots-clés : conversation en salle de réunion, parole conversationnelle, chevauchement de parole, regroupement avec information additionnelle, modèle à base de phonèmes, paramètres de réseaux de neurones artificiels, paramètres de goulot d'étranglement de l'information, combinaison de systèmes.

Contents

Acknowledgements v

Abstract vii

List of figures xiii

List of tables xvii

1 Introduction 1
- 1.1 Motivations . . . 2
- 1.2 Objectives . . . 2
- 1.3 Contributions of the thesis . . . 3
- 1.4 Organization of the thesis . . . 5

2 Speaker Diarization: Background and resources 7
- 2.1 Processing of multiple audio channels . . . 8
- 2.2 Features for speaker diarization . . . 10
  - 2.2.1 Short-term spectrum based features . . . 10
  - 2.2.2 Time delay of arrival . . . 10
  - 2.2.3 Long-term features . . . 11
- 2.3 Speech/non-speech detection . . . 12
  - 2.3.1 Energy based detectors . . . 12
  - 2.3.2 Model based detectors . . . 12
  - 2.3.3 Hybrid approaches . . . 13
- 2.4 Segmentation . . . 13
  - 2.4.1 Bayesian information criterion . . . 14
  - 2.4.2 Generalized likelihood ratio . . . 15
  - 2.4.3 KL divergence . . . 15
- 2.5 Clustering and re-alignment . . . 16
- 2.6 Speaker diarization systems . . . 16
  - 2.6.1 HMM/GMM system . . . 16
  - 2.6.2 Information bottleneck (IB) system . . . 18
  - 2.6.3 Other Approaches . . . 21
- 2.7 Resources for speaker diarization . . . 23

2.7.1 Multi-party meeting corpora . . . 24
2.7.2 Evaluation Metric . . . 24

3 Overlapping speech handling for Speaker diarization 27
3.1 Introduction . . . 27
3.2 Previous approaches for overlap detection . . . 27
3.3 Motivation for the proposed features . . . 29
3.4 Data-sets used . . . 30
3.5 Overlap speech diarization . . . 31
3.5.1 Overlap handling for speaker diarization . . . 31
3.5.2 Overlap detection using acoustic features . . . 31
3.6 Conversational features for overlap detection . . . 32
3.6.1 Silence statistics for overlap detection . . . 33
3.6.2 Speaker change statistics for overlap detection . . . 36
3.6.3 Combination of silence and speaker change statistics . . . 38
3.6.4 Combination of conversational and acoustic features . . . 40
3.7 Experiments and Results . . . 41
3.7.1 Experiments on overlap detection . . . 41
3.7.2 Speaker diarization with overlap handling . . . 45
3.8 Summary . . . 49

4 Auxiliary information for speaker diarization 51
4.1 Introduction . . . 51
4.2 Non-speech as side information for speaker diarization . . . 52
4.2.1 Agglomerative Information Bottleneck with Side Information (IBSI) . . . 52
4.2.2 Agglomerative IBSI for speaker diarization . . . 53
4.3 Phoneme background model for speaker diarization . . . 54
4.3.1 Diarization experiments using oracle annotations . . . 54
4.3.2 Estimation of a phoneme background model (PBM) . . . 56
4.4 Experiments and Results . . . 57
4.4.1 Diarization experiments using phoneme background model . . . 57
4.4.2 Diarization experiments in agglomerative IBSI framework . . . 59
4.5 Summary . . . 62

5 Artificial neural network features for speaker diarization 65
5.1 Introduction . . . 65
5.2 ANN features for speaker diarization . . . 66
5.2.1 ANN for speaker comparison . . . 66
5.2.2 ANN for speaker classification . . . 67
5.2.3 ANN as an auto-encoder . . . 68
5.3 Experiments and Results . . . 69
5.3.1 Datasets used in experiments . . . 69
5.3.2 Training of the ANNs . . . 70

5.3.3 Speaker diarization evaluation 70
5.4 Summary 74

6 Combination of different diarization systems **77**
6.1 Introduction 77
6.2 Previous approaches to system combination 77
6.3 Combination of HMM/GMM and IB systems 78
6.3.1 Motivation for the proposed method 78
6.3.2 Information Bottleneck features 79
6.4 Experiments and Results 80
6.5 Summary 84

7 Conclusion and Future Directions **85**
7.1 Conclusion 85
7.2 Future directions 86

Bibliography **101**

Curriculum Vitae **103**

List of Figures

2.1 Speaker diarization segments the given audio stream based on speaker turns. Segments of same colour belong to a single speaker. 7

2.2 Block diagram showing agglomerative clustering framwork for speaker diarization: The captured speech signals at Multiple Distant Microphones (MDMs) are enhanced by the delay-sum beamforming followed by feature extraction and speech/non-speech detection. Non-speech regions are ignored and speech portions are uniformly segmented and agglomerative clustering is initialized with each segment as a cluster. Multiple iterations of cluster merging (based on a distance measure) and re-alignment are performed until the stopping criterion is satisfied. 9

2.3 HMM/GMM speaker diarization system. 18

2.4 HMM/GMM speaker diarization system using multiple feature streams. 19

2.5 Agglomerative Information Bottleneck speaker diarization system. 21

2.6 Agglomerative Information Bottleneck (aIB) speaker diarization system with multiple feature streams. 21

3.1 Speaker vocalizations from a snippet of multi-party conversation. The fixed length segments (a) and (c) are in regions of speaker change and contain overlap as well as less amount of silence whereas segments (b) and (d) contain single speaker speech, are reasonably away from any speaker changes, and also contain more silence. 30

3.2 Block diagram of overlap handling for speaker diarization: Overlaps obtained from overlap detection are first excluded from the diarization process and later, appropriate speaker labels are assigned to the overlap segments based on methods described in Section 3.5.1 in the labelling stage, to generate the final diarization output. 31

3.3 *Probability of overlap based on silence duration obtained using ground truth speech/sil segmentation and automatic speech activity detector output.* 33

3.4 *Estimation of probabilities of single-speaker speech and overlapping speech states for a frame i based on duration of silence d_i present in the segment s_i centered around the frame i.* 34

3.5 Cross-entropy between estimates based on silence duration (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) NIST-RT 09 (c) ICSI. 35

3.6 *Probability distributions of number of occurrences of overlaps (O) for different numbers of speaker changes (c) obtained using ground-truth segmentation and diarization output.* 37

3.7 *Estimation of probabilities of single-speaker and overlapping speech classes for a frame i based on number of speaker changes c_i present in the segment s_i centered around the frame i.* 38

3.8 Cross entropy measure between estimates based on speaker changes (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) RT 09 (c) ICSI for various segment lengths. 38

3.9 Block diagram of the proposed method of overlap detection: Speech, silence statistics and speaker change statistics are obtained as a by product of initial pass of IB speaker diarization. These statistics are used to estimate the probability of occurrence of overlap as explained in Sections 3.6.1, 3.6.2 and 3.6.3. These probabilities are incorporated into the acoustic feature based overlap detector (Section 3.5.2) as prior probabilities as detailed in Section 3.6.4 41

3.10 Overlap detection evaluation on (a) AMI-test set (b) NIST-RT 09 set and (c) ICSI-test set. In each subplot (a), (b), (c), dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) in x-axis. 43

3.11 Overlap detection evaluation on laughter segments from ICSI-test set: dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) on x-axis. 44

3.12 Effect of diarization errors: (a) High error set (b) Low error set. Dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection on in percentage y-axis for various overlap insertion penalties (OIP) on x-axis. 45

3.13 Overlap labelling: Tuning for optimal overlap insertion penalty (OIP) on AMI development set. 48

4.1 Agglomerative IBSI algorithm 55

4.2 *Comparison of oracle UBMs: Spkr-UBM is estimated with the knowledge of speaker labels, Spkr-phone-UBM is estimated with the knowledge of both the speaker and phoneme being spoken and Plain-UBM is the regular UBM.* 56

4.3 *Block diagram of the proposed diarization method using PBM.* 57

4.4 *Meeting wise speaker error values for baseline aIB diarization system and aIBSI system.* 60

4.5 *Speaker error for various values of γ on development set of meetings from RT05 eval set.* 61

5.1 *An ANN architecture to classify two given speech segments as belonging to same or different speakers. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.* 67
5.2 *An ANN architecture to classify a given input speech segments into one of speakers in the output layer. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.* 68
5.3 *An ANN as an auto-encoder: The network reconstructs the center frame of the input (center frame + context) at the output layer. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.* 69
5.4 Input context length tuning using ICSI-dev set for various bottleneck features in HMM/GMM based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder. 71
5.5 Input context length tuning using ICSI-dev set for various bottleneck features in information bottleneck based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder. 71
5.6 Feature stream weight tuning using ICSI-dev set for combination of MFCC features with bottleneck features obtained from different ANNs in HMM/GMM based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder. 73
5.7 Feature stream weight tuning using ICSI-dev set for combination of MFCC features with bottleneck features obtained from different ANNs in information bottleneck based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder. 73

6.1 *Block diagram of the proposed method.* 80
6.2 *Meeting wise speaker error values for baseline HMM/GMM diarization system and IB_multistr.* 82
6.3 *Speaker error after each merge for baseline and IB_multistr systems for meeting EDI_20061113-1500* 83

List of Tables

2.1 *Details of meeting room conversations corpora.* 24

3.1 *Number of meetings used as part of train, development and test sets in each data set.* 30

3.2 Cross entropy measure on AMI development set of meetings for various combination strategies. 40

3.3 *Overlap exclusion on the AMI-test, NIST-RT 09 and ICSI-test data sets: SpNsp represents error in speech/non-speech detection, Spkr denotes speaker error (clustering error), DER is the total diarization error rate and f-measure of the respective overlap detection systems at the operating point (OIP=0) used for exclusion.* . . . 47

3.4 *Overlap labelling on AMI-test and NIST-RT 09 and ICSI-test data sets: missed (Miss) and false-alarm (FA) errors in speech/non-speech detection and total speech/non-speech error (SpNsp), speaker error (Spkr), diarization error rate (DER) and f-measures of the respective overlap detection systems at the operating point (OIP=90) chosen to do labelling.* 49

3.5 *DERs (with relative improvements over baseline diarization within parenthesis) obtained while performing overlap handling by using overlaps detected by proposed method (+Conv-priors) and acoustic feature based overlap detector (Acoustic).* 49

4.1 *Diarization experiments on development sets: Speaker error for different systems on development set of meetings from AMI (20 meetings) and NIST-RT (RT-05,06) corpora.* 58

4.2 *Diarization experiments on 100 test meetings from AMI corpus: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems.* 59

4.3 *Diarization experiments on NIST-RT test set: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems on NIST-RT 07, 09 data sets.* 59

4.4 *Speech/non-speech errors for RT 06, 07 and 09 sets.* 61

4.5 *Speaker error for aIB, aIBSI diarization systems (with relative improvements over aIB baseline in parenthesis) on RT 06, 07 and 09 sets. We also report the performance of HMM/GMM system for comparison.* 62

5.1 *Meeting corpus statistics as used in experiments, including numbers of distinct speakers, meeting room sites, and number of meetings used as part of train, development and test sets.* 69
5.2 *Speaker errors on test data sets for various bottleneck features in HMM/GMM and IB speaker diarization system.* 72
5.3 *Speaker errors on test data sets after combining different bottleneck features with MFCC features in HMM/GMM and IB speaker diarization systems.* 74

6.1 Main differences between the HMM/GMM and the IB diarization systems. . . . 78
6.2 *Speech/non-speech error rate in terms of missed speech (Miss), false alarm speech (FA) and the total error (SpNsp) in the evaluation data sets.* 81
6.3 *Total speaker error with relative improvement over baseline in parenthesis on the evaluation data sets (RT06, RT07, RT09 combined) for various diarization systems.* 81
6.4 *Data-set wise speaker error on RT06, RT07, RT09 data sets for various diarization systems.* 81
6.5 *Combination of IBSI and HMM/GMM systems: The IBSI system uses phoneme background model (PBM) as relevance variable set and components of GMM estimated from non-speech regions as irrelevant variable set. Speaker error is reported for various test sets NIST-RT 06/07/09.* 83

1 Introduction

We are witnessing an exponential growth in audio and video recordings being produced and archived due to the availability of cheap and efficient storage and processing power. This exponential growth in the audio and video data has created a necessity for technologies that can facilitate efficient access to information present in these recordings. This data carries various rich information sources such as lexical content (what is being spoken), speaker (who is speaking), background scene and environment etc.,. Automatic methods to extract this information pave way for large scale indexing of these data streams which facilitates efficient browsing and information retrieval.

A significant portion of these recordings consist of television broadcast shows, telephone calls, lectures and conference style meeting recordings. These recordings typically feature multiple speakers and knowing who is speaking when is an useful information in this scenario. The goal of speaker diarization is to provide such meta-information by identifying 'who is speaking when' in a given audio recording containing multiple speakers [Tranter and Reynolds, 2006]. It involves segmenting a given audio stream into speaker homogeneous regions and attributing each region to the respective speaker by clustering. It is an unsupervised task as there is no a-priori information about the speakers or the number of speakers present in a recording.

One of the main applications of speaker diarization is that it is used as a preprocessing step before performing automatic transcription of multi-party conversations. In this application, speaker diarization output is used to adapt the speaker independent acoustic models to increase the accuracy of transcription. Apart from this, speaker diarization output adds structure to a multi-party conversation by identifying the beginning and end times of different speakers (speaker turns) in the conversation. This structure is very helpful in meaningful indexing and is also useful in analyzing the conversation to infer various high level phenomena such as group dynamics, co-operation and participation [Jayagopi et al., 2009b, Jayagopi et al., 2009a, Jayagopi and Gatica-Perez, 2010].

1.1 Motivations

The three main domains of application of speaker diarization which are broadcast shows, telephone calls and meeting recordings vary in the nature of speech, recording equipment used to capture the audio signal, and the typical number of speakers involved in a recording [Tranter and Reynolds, 2006, Anguera et al., 2012].

In broadcast shows the speech is usually read or prepared, has very few disfluencies and the signal is usually captured by close talking microphone in a studio environment. These shows might also contain background music and advertisements.

Telephone calls are usually between two people and contain predominantly spontaneous speech which might be corrupted by background noise or channel distortions.

Among all these domains, diarization of meeting recordings is the most challenging task as it usually contains spontaneous speech with simultaneous speakers and short speaker turns. Also the audio which is typically captured by distant microphones is often corrupted by room reverberation and background noise. Due to this reason the most recent NIST-RT evaluation campaigns have concentrated on diarization of meeting room conversations. Studies [Huijbregts and Wooters, 2007, Shriberg et al., 2001, Knox et al., 2012, Huijbregts et al., 2012, Sinclair and King, 2013] done on the state of the art speaker diarization systems in NIST-RT evaluation campaigns have highlighted three main issues with these systems:

1. **Overlapping speech:** Systems are unable to handle simultaneous speakers resulting in overlapping speech. Overlapping speech corrupts speaker clusters and also results in missed speaker error when all the speakers in the overlapping region are not labelled correctly. Studies on meeting recordings have shown that the error can be reduced by around 40% relative by effective overlap handling.

2. **Background noise and errors in speech/non-speech detection:** Errors in speech/non-speech detection introduce non-speech segments into speaker clusters which corrupt the models and background noise effects the clustering in speech regions with low signal to noise ratio.

3. **Variations among different systems:** The performance of the systems varies significantly on the same set of data i.e., there is no single best system on all the recordings. This shows that systems are modelling complementary information to each other and diarization performance can be improved by combining different systems.

1.2 Objectives

The main objective of this thesis is to address the above mentioned shortcomings observed in the state of the art speaker diarization systems. The objectives can be summarized as follows:

- Propose novel features and methods for overlapping speech detection and diarization.
- Increase the robustness of the diarization system to effects of errors made in speech/non-speech detection and background noise by proposing modifications to the clustering framework and by the use of new features.
- Propose a novel combination method for different diarization systems such that the combination exploits the complementary nature of the two diarization systems being combined and does not require any modifications to the original systems.

Previous approaches in the literature have proposed methods using different features such as time delay of arrival (TDOA) for the problem of overlapping speech diarization [Zelenák et al., 2012] and to improve the robustness of clustering [Pardo et al., 2007, Anguera, 2006]. In this thesis we mainly focus on leveraging the information present in the acoustic signal and the structure of conversation to address these issues.

1.3 Contributions of the thesis

The contributions of this thesis can be summarized as follows:

- **Overlapping speech detection to improve speaker diarization.**
  - We propose long-term conversational features derived from the structure of a conversation to improve overlap detection. The features include silence and speaker change statistics computed over a long-term window of 3–4 seconds. These features are used to estimate the probability of occurrence of overlap in the window. The estimated probabilities are used as prior probabilities in an acoustic feature based classifier. Experimental results show that the proposed method performs equally well on meetings from different corpora and significantly reduces the diarization error rate (DER). Parts of this work have been published in [Yella and Valente, 2012, Yella and Bourlard, 2013] and a comprehensive journal article has been published in [Yella and Bourlard, 2014a].
- **Use of auxiliary information sources for speaker diarization.**
  - To suppress the effects of background noise and errors due to speech/non-speech detection, we propose Information Bottleneck with Side Information (IBSI) framework for speaker diarization. The IBSI is a non-parametric clustering framework that maximizes mutual information between the final clustering output and a relevant variable set while minimizing the mutual information with irrelevant variable set. In this framework, the relevant variable set is denoted by the components of a Gaussian Mixture Model (GMM) based background model estimated from the speech regions of a given recording and the irrelevant variable set is represented

by the set of components of GMM estimated from the non-speech regions of a given recording. Experiments on meetings from several corpora have shown that the IBSI framework significantly reduces the diarization error. The findings of this work have been published in [Yella and Bourlard, 2014b]

  - We show that information from phoneme transcripts of a given recording are useful for speaker diarization by performing experiments using oracle annotations. Based on these observations, we propose a method to use the information from phoneme transcripts to improve speaker diarization. This is achieved by estimating a phoneme background model utilizing the information from phoneme transcript of a given recording. Experiments on several meeting corpora have shown that the use of PBM in the place of a normal background model improves speaker diarization output. This work has been published in [Yella et al., 2014a].

- **Artificial neural network (ANN) features for speaker diarization**

  - We propose the use of features extracted from three different ANN architectures trained on related tasks to speaker diarization as features for speaker diarization. The first network is trained to detect if two given input speech segments belong to the same or different speakers. The second network is trained to classify the input speech segment into one of the pre-determined set of speakers used while training the network. The third network is an auto-encoder which is trained to reconstruct the input at the output with as low reconstruction error as possible. The hidden layer activations from a bottleneck layer of these networks are used as features for speaker diarization. The rationale behind using the hidden layer activations as features is that the initial layers of ANN would transform the input features into a space more conducive for speaker discrimination. The neural network features are evaluated on their own and in combination with traditional Mel frequency cepstral coefficients (MFCCs) which are typically used as features for speaker diarization. Experimental results have shown that, though the ANN features do not decrease the diarization error when used on their own, they provide complementary information to MFCCs and improve the diarization output when combined with MFCCs. Part of this work has been published in [Yella et al., 2014b]

- **Combination of diarization systems to capture their complementary characteristics**

  - We address the issue of performance variance between different diarization systems by proposing a new combination strategy for combining a HMM/GMM based speaker diarization system and an Information Bottleneck (IB) based speaker diarization system. These systems vary in several aspects such as modelling and distance measure used in clustering. Hence the combination of the two systems captures the complementary nature. The combination takes place at the feature level, where the output of IB system is used to generate features for HMM/GMM system. The combination does not require any changes to the original system. This work has been published in [Yella and Valente, 2011].

1.4 Organization of the thesis

The thesis is organized as follows:

- Chapter 2 (Background and resources): This chapter provides a brief overview of the main components in a state of the art speaker diarization systems. It briefly explains state-of-the-art speaker diarization frameworks based on HMM/GMM and IB which are used in the current work. It also gives details of the various input features used for speaker diarization along with the evaluation metrics used to evaluate the speaker diarization output. Finally, it provides details of various databases of multi-party meeting recordings which are used in the current thesis.
- Chapter 3 (Overlapping speech detection and diarization): This chapter provides details on the proposed long-term conversational features for overlapping speech detection. It reports experiments done on several meeting corpora to evaluate the effectiveness of the proposed features. It also explains the ways in which the detected overlaps can be used to improve speaker diarization and reports speaker diarization experiments on different meeting corpora.
- Chapter 4 (Auxiliary information sources for diarization): This chapter explains the ways to use auxiliary information sources such as non-speech portions and phoneme transcripts to help improve speaker diarization. It explains the IBSI based speaker diarization where non-speech regions are used as side information for clustering. It also provides details on how to use the phoneme transcripts to improve speaker diarization by explaining the estimation of phoneme background model which is used in IB diarization system.
- Chapter 5 (Artificial neural network features for diarization): This chapter explains various neural network architectures explored to extract features for speaker diarization. It provides details on the network training, feature generation and reports speaker diarization experiments on HMM/GMM and information bottleneck systems.
- Chapter 6 (System combination): This chapter details the feature level combination of HMM/GMM and IB systems.
- Chapter 7 (Conclusions): This chapter summarizes the conclusions of the thesis along with possible future directions.

2 Speaker Diarization: Background and resources

The goal of speaker diarization is to partition a given speech recording into speaker-homogeneous regions along with attributing these regions to the speakers in the audio stream [Tranter and Reynolds, 2006] as shown in Figure 2.1. Most of the state-of-the-art speaker diarization systems

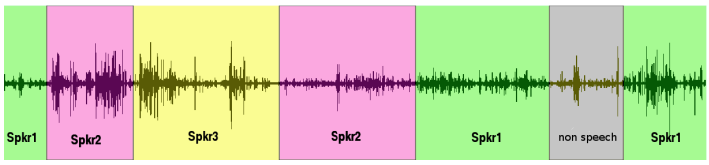


Figure 2.1: Speaker diarization segments the given audio stream based on speaker turns. Segments of same colour belong to a single speaker.

are based on the agglomerative clustering framework [Tranter and Reynolds, 2006, Wooters and Huijbregts, 2008, Vijayasenan, 2010]. The main steps involved in speaker diarization are processing of the speech signals captured by single or multiple microphones, feature extraction, speech/non-speech detection, segmentation and clustering.

In the preprocessing step, noise reduction methods such as Wiener filtering [Wiener, 1949, Adam et al., 2002a] are applied to the signal captured to reduce the effects of background noise. Also, when the signal is captured by multiple distant microphones (MDMs), beamforming methods could be applied to obtain an enhanced signal with high signal to noise ratio. These methods have shown to be effective especially in diarization of meeting room conversations and are explained in detail in Section 2.1.

The typical features used for speaker diarization are short-term spectrum based features which represent the vocal tract characteristics of the speaker. When the audio is captured by multiple distant microphones (MDM) like in a meeting room environment, the time delay of arrival (TDOA) of the speech signal at different microphones carries useful information about spatial location of the source speaker [Anguera et al., 2007]. Section 2.2 explains in detail the different types of features used for speaker diarization.

After feature extraction, speech activity detection is performed, and non-speech regions are excluded prior to the initialization of clustering to prevent corruption of speech regions. The non-speech class encompasses a broad range of signals such as background noise, silence, vocal sounds like cough or laughter and other sounds like music, door slamming, or any other ambient noises. These signals, if included in the clustering, add impurity to speaker clusters and increase the clustering error. Several speech activity detection algorithms have been proposed in the literature. They are presented in detail in Section 2.3.

The initialization of diarization is done by segmenting the speech regions identified by the speech activity detector. The segmentation can be done either by speaker change detection or by uniform segmentation. Several approaches for change detection have been proposed in literature which are based on different criteria like Bayesian information criterion (BIC), Kullback Leibler (KL) divergence, generalized likelihood ratio (GLR) etc.,. Studies on meeting diarization have shown that uniform segmentation based initialization followed by re-alignment after each cluster merge yields similar diarization output to that of initialization based on speaker change detection. Section 2.4 provides details of various segmentation methods proposed for speaker diarization systems.

The agglomerative clustering is initialized by representing segments obtained from uniform segmentation or speaker change detection as individual clusters. At each step clusters that are closest to each other are merged based on a distance measure. After each cluster merge the feature vectors are re-assigned to the cluster models to allow for corrections in any errors committed in the previous segmentation steps. The clustering continues until no suitable clusters are available for merging, which is decided based on a stopping criterion. Several distance measures and stopping criteria have been proposed in the literature. Section 2.5 outlines these approaches. A block diagram of a typical agglomerative speaker diarization system is shown in Figure 2.2.

2.1 Processing of multiple audio channels

Diarization of meeting room recordings typically uses audio captured by distant microphones placed at different locations in the meeting room. Due to this, the audio captured in a meeting room environment is far-field and is influenced by various factors such as background noise, room reverberation and different qualities of microphones used for the recording [Janin et al., 2003, Mccowan et al., 2005, Mostefa et al., 2007]. Several methods have been proposed in the literature to handle these issues. For example the audio signals can be pre-processed using Wiener filtering [Wiener, 1949] to attenuate the noise. Several works on speaker diarization use the toolkit developed in [Adam et al., 2002a] to perform this and improvements are observed over the systems that do not use noise reduction. The single channel speaker diarization can be adapted to multi-channel speaker diarization in several ways. In [Fredouille et al., 2004], speaker diarization is performed individually on each channel and the final output is obtained by merging the outputs of the individual channels. In [Jin et al., 2004], speaker

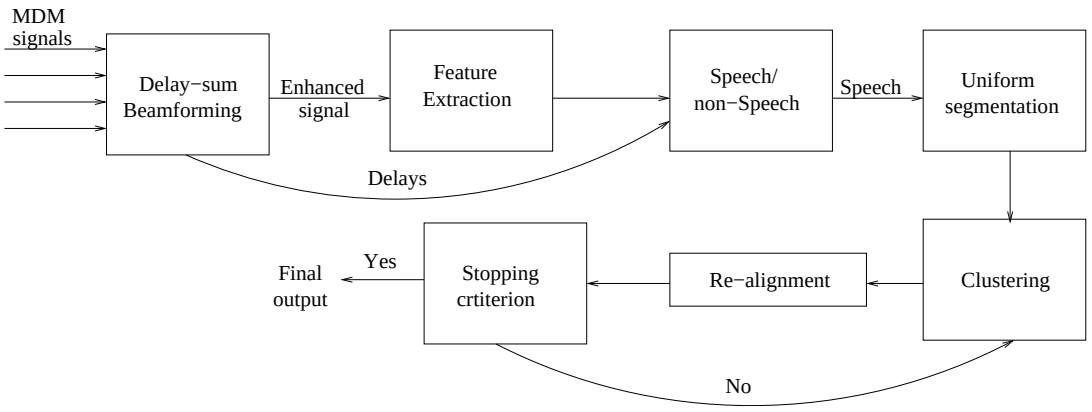


Figure 2.2: Block diagram showing agglomerative clustering framwork for speaker diarization: The captured speech signals at Multiple Distant Microphones (MDMs) are enhanced by the delay-sum beamforming followed by feature extraction and speech/non-speech detection. Non-speech regions are ignored and speech portions are uniformly segmented and agglomerative clustering is initialized with each segment as a cluster. Multiple iterations of cluster merging (based on a distance measure) and re-alignment are performed until the stopping criterion is satisfied.

segmentation is performed individually on each channel and the final diarization output is obtained by considering the channel whose segments have high signal to noise ratio (SNR). Several methods have tried to combine multiple audio signals to obtain a single signal which is used for diarization. In [Istrate et al., 2005] the signals captured by different microphones are combined by a SNR based weighted combination. But this method has the disadvantage that it does not consider the time delay of arrival at the multiple microphones which very often leads to poor performance.

Recent approaches to multi-channel diarization [Anguera et al., 2005, Pardo et al., 2007, Anguera et al., 2007] use acoustic beamforming as an important pre-processing step to speaker diarization. Popular open source implementation of delay-sum beamforming method known as BeamformIt [Anguera, X,] is often used for this purpose. Here, signals captured by multiple distant microphones are aligned to a reference channel by correcting for the time delay of arrival at different microphones before obtaining a weighted combination of these signals to get an enhanced signal. The weights for the combination are based on the individual channel SNRs or on cross-correlation based metric. The time delay of arrival at different microphones is obtained by using delay-sum beamforming which finds the location of peak in the cross-correlation function between the reference channel and other channels. This procedure helps speaker diarization in two ways, as it provides an enhanced signal and also as a by-product, the TDOA values at multiple distant microphones, which carry spatial information on the sound source are very helpful for speaker diarization. This method of beamforming is very useful for speaker diarization as it does not need any prior information on microphone locations, room architecture and ambient noise. Although several other sophisticated methods for beamforming like maximum likelihood beamforming [Seltzer et al., 2004], generalized side

lobe canceler (GSC) [Griffiths and Jim, 1982] and minimum variance distortion less response (MVDR) [Woelfel and McDonough, 2009] are available they either need prior information or are computationally very expensive.

2.2 Features for speaker diarization

This section gives an overview of different types of features that have been used for speaker diarization. These features can be broadly classified into three categories; short-term spectrum based features, spatial features (based on time delay of arrival) and long-term features (estimated from a longer context).

2.2.1 Short-term spectrum based features

Features extracted from the short-term spectrum are used in almost all speaker diarization systems. The rationale behind this is that, the short-term spectrum based features carry information about the vocal tract characteristics of individual speakers. These features are extracted from either the short-term Fourier transform of the windowed speech signal or from linear prediction (LP) analysis. These features are typically extracted for every 10 ms from a window size of around 30 ms. Many systems [Ajmera et al., 2004, Ajmera, 2004, Vijayasenan and Valente, 2012] use Mel frequency cepstral coefficients (MFCCs) as features for speaker diarization. The dimensionality of MFCCs is around 20, which is higher than what is typically used for automatic speech recognition (13). This helps in capturing the source characteristics of the speech signal. Delta, delta-delta features which are used in ASR are not used in speaker diarization as they mainly capture short-term dynamics that are mainly useful for distinguishing phoneme classes rather than speakers. Other short-term spectral features used for speaker diarization include line spectrum pair (LSP) features [Adam et al., 2002b], perceptual linear prediction coefficients (PLP) [Sinha et al., 2005], and linear prediction cepstral coefficients (LPCCs) [Ajmera and Wooters, 2003].

To improve the noise robustness of the features, several feature warping techniques have been proposed [Peleconas and Sridharan, 2001], which basically gaussianize the feature distribution prior to modelling. This has been shown to give significant improvements [Sinha et al., 2005, Zhu et al., 2006]. Also, RASTA-PLP [Hermansky and Morgan, 1994] features which are designed to deal with convolutive and additive noises have been used [van Leeuwen and Huijbregts, 2006, van Leeuwen and Konecny, 2008].

2.2.2 Time delay of arrival

When the speech signal is captured by multiple distant microphones (MDMs), the Time Delay of Arrival (TDOA) of the signal at MDMs carries useful information on speaker location. TDOA features are estimated as a by-product of delay-sum beamforming performed as a

preprocessing step. These features are very helpful for speaker diarization as they provide information on speaker location which is complementary to the short-term spectrum based features. The assumption here is that the position of the speaker remains more or less constant through out the recording. Given multiple signals captured by microphones placed at different locations in a room, the first step is to pick a reference channel. This can be done based on the SNR or on cross-correlation metric. Then, the TDOA features are estimated by locating the peak in the cross-correlation function between the reference channel and a microphone channel. The cross-correlation is performed in the spectral domain after applying a Fourier transform and magnitude normalization to improve robustness against noise and reverberation. The technique is called generalized cross correlation with phase transform (GCC-PHAT) [Knapp and Carter, 1976, Brandstein and Silverman, 1997]. The cross-correlation between two signals $x_i(n)$ and $x_j(n)$ using GCC-PHAT is computed as follows:

$$GCCPHAT(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{2.1}$$

where, $X_i(f)$ and $X_j(f)$ are the Fourier transform of the two signals and $[]^*$ denotes complex conjugate. The TDOA estimates from the cross-correlation are obtained as:

$$t(i,j) = \arg\max_d IFT(GCCPHAT(f)) \tag{2.2}$$

where, $t(i,j)$ is the TDOA estimate between the two channels i, j and $IFT()$ denotes inverse Fourier transform.

Once the TDOA features are estimated, they can be used for speaker diarization on their own or in combination with the short-term spectrum based features. Studies [Pardo et al., 2007, Pardo et al., 2006] have shown that the combination of MFCC features and TDOA features at the model level give significant improvements in speaker diarization output. The weights of the combination can be estimated on a separate set of recordings or can be estimated automatically using entropy based metrics [Anguera et al., 2007].

2.2.3 Long-term features

The short-term spectrum based MFCC features are used for both automatic speaker and speech recognition tasks. The main motivation for using long-term features for speaker diarization is that they can capture individual characteristics of a speaker's voice and speaking behavior which are not captured by short-term spectrum based features [Friedland et al., 2009]. In [Friedland et al., 2009] a systematic investigation of 70 long-term features which are mainly pitch related prosodic features is conducted for their usefulness for speaker discrimination. The combination of 10 top-rank long-term features with MFCCs has reduced the error in diarization by about 30% relative. In addition to this, these features have also been successfully used to initialize speaker clusters in an agglomerative clustering framework and have been shown to work better than previous initialization methods.

Modulation spectrogram features [Kingsbury et al., 1998] that depend on a long temporal context were also used to complement short-term spectrum based features [Vinyals and Friedland, 2008].

2.3 Speech/non-speech detection

The aim of a speech activity detection (SAD) system is to label speech and non-speech segments in a given audio stream. The errors made by a SAD system impact diarization in two significant ways. The missed speech segments and false-alarm speech detections contribute directly to the diarization error rate (DER) in the form of missed and false alarm errors. Therefore, a poor SAD system will adversely effect the diarization evaluation metric DER (explained in detail in Section 2.7.2). The other way in which the errors made by SAD system influence speaker diarization is, that these false-alarm errors introduce impurities into acoustic models of speaker clusters [Wooters et al., 2004]. The speech segments missed by a SAD system reduce the speech data available for speaker clusters. Both phenomena result in an increase of clustering error.

The problem of SAD has been well studied in speech processing literature in various contexts such as speech coding, enhancement, recognition etc.,. Many approaches have been proposed in the literature to solve this task [Lu et al., 2002]. Initial works on speaker diarization have attempted to perform diarization without SAD with the a logical reasoning that this would result in an extra cluster with non-speech regions. But it was observed that this adversely effected the speaker diarization output. Therefore, a dedicated SAD system is used from then on for this task. The main approaches used for SAD can be classified into three categories; energy based, model based and hybrid approaches which are detailed below.

2.3.1 Energy based detectors

Energy based detectors typically use thresholds on short-term spectral energy to decide whether a region contains speech/non-speech [Junqua et al., 1994, Lamel et al., 1981, Wooters et al., 2004]. This class of methods are generally employed in telephone speech and do not require any labelled training data. Since the channel, noise, and recording conditions vary drastically from one meeting room to another these methods usually do not generalize well to different meeting recordings [van Leeuwen and Huijbregts, 2006, Istrate et al., 2005].

2.3.2 Model based detectors

Model based detectors use pre-trained models on labelled data from speech and non-speech classes to identify the classes in unlabeled data [Zhu et al., 2008, Anguera et al., 2005, Fredouille and Senay, 2006]. The pre-trained models can also be adapted to the test data [Huijbregts and de Jong, 2011, Fredouille and Evans, 2008]. Usually, Gaussian mixture models are estimated

for each class and the detection is based on Viterbi decoding using these models. A minimum duration constraint for each class (speech,non-speech) is applied while decoding to prevent short segments. Apart from modelling speech and non-speech classes, several models can be trained for more specific classes such as speech + noise, speech + music, music, silence, etc.,. Works on broadcast news [Nguyen et al., 2002, luc Gauvain et al., 1998, Zhu et al., 2005] have proposed using gender and: channel dependent models to improve the SAD output. Although using more specific classes would result in improvements, its main disadvantages are the need for sufficient amount of labelled data to train the models, and generalizability to new environments. Discriminant classifiers based on linear discriminant analysis (LDA) [Rentzeperis et al., 2006], support vector machines (SVMs) [Temko et al., 2007] and multi layer perceptrons (MLPs) [Dines et al., 2006] have also been proposed in the literature. In the case of MLPs, the class posterior estimates obtained at the output layer are used as scaled likelihoods in Viterbi decoding to perform speech/non-speech detection.

2.3.3 Hybrid approaches

Hybrid approaches to SAD have been proposed to overcome the issue of generalizability of the pre-trained models to new conditions [Anguera et al., 2006a, Sun et al., 2009, Nwe et al., 2009, Wooters and Huijbregts, 2008, El-Khoury et al., 2009]. These approaches combine both threshold based energy detectors and model based detectors to perform speech/non-speech detection. In the first step, a threshold based energy detector is used to label unsegmented audio and in the second step, the segments of high confidence in the labelled data are used to train new models or adapt pre-trained models to current conditions. This method also alleviates the need for labelled training data.

2.4 Segmentation

The goal of the segmentation step in speaker diarization is to segment the given audio into speaker homogeneous segments which are used for clustering in the later stages. In other words, the goal is to find all the speaker change points in the audio under analysis. The most common approach to change detection is to perform hypothesis testing between two adjacent windows of speech. The hypotheses that are tested for change detection are: speech in the two windows is produced either by same speaker or different speakers. The windows are shifted at regular intervals to detect all the change points. The hypotheses are tested by estimating models over the speech data in the two windows and computing a criterion based similarity/distance metric between these models. A threshold on the metric is used to determine the change point. Several different criterion have been proposed in the literature and they are presented below.

2.4.1 Bayesian information criterion

The Bayesian Information Criterion (BIC) is a model selection criterion [Schwarz, 1978] used to determine which model best explains the given data. It has been used for acoustic change detection in [Chen and Gopalakrishan, 1998]. The BIC value of a model M representing data X is given by:

$$BIC(M) = \log \mathscr{L}(X|M) - \frac{\lambda}{2}\#(M)\log(N) \tag{2.3}$$

where, $\mathscr{L}(X|M)$ is the likelihood of data X given the model M, N is the number of data points in X, $\#(M)$ is the number of parameters in the model M, and λ is a trade-off between the model complexity and how well the model fits the given data.

For change detection between two segments X_i and X_j this criterion is computed for the two hypotheses $H1$ and $H2$:

- $H1$: The data is generated by a single speaker, in which case a single model best explains the data
- $H2$: The data is generated by two different speakers in which case two models (one for each segment) best explain the data.

The difference in the BIC values of these two hypotheses is denoted as $\Delta BIC(i,j)$ and is given by:

$$\Delta BIC(i,j) = \log \mathscr{L}(X_{ij}|M_{ij}) - [\log \mathscr{L}(X_i|M_i) + \log \mathscr{L}(X_j|M_j)] - \frac{\lambda}{2}\delta_{ij}\log(N) \tag{2.4}$$

where X_{ij} denotes the data resulting from combination of segments X_i and X_j, M_{ij} denotes the model estimated over the combined data, and δ_{ij} is the difference in the number of parameters in models representing $H1$ and $H2$. A change is detected when the value of $\Delta BIC(i,j)$ is less than zero. In practice, the two windows X_i and X_j are slid along the input audio stream at regular intervals to compute (2.4). The local minima in this sequence of ΔBIC values whose value is below zero are detected as change points. Whenever a minimum is observed, the windows are re-adjusted to start from the detected change point and when there is no change detected between two windows, the window size is increased to encompass both the windows and the computation is repeated. The trade-off parameter λ has to be tuned on a separate development data to get optimal change detection. To avoid this, a modification to the ΔBIC has been proposed [Ajmera et al., 2004] in which the number of parameters before and after the merge are made equal i.e., $\#(M_{ij}) = \#(M_i) + \#(M_j)$. This makes the term δ_{ij} in (2.4) equal to zero which gets rid of the last term. This modification is referred in literature as modified ΔBIC and has been shown to work very well for diarization [Ajmera, 2004, Wooters and Huijbregts, 2008].

The BIC based change detection has two main drawbacks [Tranter and Reynolds, 2006]. First,

it misses changes occurring in a short time span i.e., around two seconds and secondly, the sliding window based search for change point detection is computationally very expensive. To overcome the computational complexity, many systems [Delacourt and Wellekens, 2000] use a computationally less expensive distance measure like KL divergence to detect initial change points and these points are then refined in the second stage using BIC. Some works [Zibert et al., 2009] adapt the window length used to compute ΔBIC to reduce the number of computations.

2.4.2 Generalized likelihood ratio

The generalized likelihood ratio (GLR) [Jin et al., 2004, Gish et al., 1991, Han and Narayanan, 2008, Gangadharaiah et al., 2004] is similar to ΔBIC except that it does not take into account the model complexity factor. Given two segments X_i and X_j, the GLR computes the likelihood ratios of the two hypotheses similar to ΔBIC. In the first hypothesis, it is assumed that the data in the two segments is generated by a single model and in the second hypothesis, the data in each segment is assumed to be generated by a separate model. The GLR is computed as follows:

$$GLR(i,j) = \frac{\mathscr{L}(X_{ij}|M_{ij})}{\mathscr{L}(X_i|M_i) + \mathscr{L}(X_j|M_j)} \tag{2.5}$$

where $\mathscr{L}(X|M)$ denotes the likelihood of data X given a model M, X_{ij} and M_{ij} denote the data and model resulting from the combination of segments X_i and X_j. When the number of parameters in M_{ij} is made equal to $M_i + M_j$, the GLR boils down to the modified ΔBIC criterion.

2.4.3 KL divergence

Kullback Leibler (KL) divergence is used to measure the dissimilarity of two probability distributions. In the case of change detection, the KL divergence between the distributions of the data in two segments can be used to decide whether there is a change or not [Siegler et al., 1997, Zochová and Radová, 2005, Zhu et al., 2006]. The changes are detected by setting a threshold on the divergence value. The KL divergence is computed between the two segments X_i and X_j by modelling them with corresponding Gaussian distributions N_i and N_j. The divergence can be obtained in closed form as:

$$KL(X_i||X_j) = \frac{1}{2} tr[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1}) + (\Sigma_j^{-1} - \Sigma_i^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^{'}] \tag{2.6}$$

where, μ_k, Σ_k denote the mean and covariance of data X_k. The symmetric form of KL divergence (KL2) is also used for change detection. It is computed as:

$$KL2(X_i, X_j) = KL(X_i||X_j) + KL(X_j||X_i) \tag{2.7}$$

As no closed form expression for KL divergence exists for mixture models, Gaussian distributions are used to model individual segments. This limits the modelling capacity but it is computationally inexpensive to compute. For this reason, KL based measures are used in the first step of change detection which is later refined by more sophisticated measures such as BIC [Zochová and Radová, 2005, Delacourt and Wellekens, 2000].

2.5 Clustering and re-alignment

The clustering step takes the speaker homogeneous segments obtained from the segmentation step and clusters them until there is only one cluster for each speaker. The cluster similarity/distance measures used for clustering can be the same as those used for segmentation. The BIC measure [Chen and Gopalakrishan, 1998], KL divergence [Siegler et al., 1997, Zochová and Radová, 2005], KL2 distance [Zhu et al., 2006], and GLR [Jin et al., 2004, Gish et al., 1991] are used as distance measures for clustering.

When the segmentation and clustering steps are performed sequentially, the errors made in the segmentation step are passed on into the clustering step with no chance of correcting them. Due to this, most diarization approaches use iterative clustering and re-alignment steps to reduce the errors made during initial segmentation step. In this method, after merging two clusters, a few iterations of Viterbi re-alignment and model re-estimation steps are performed to reduce errors in the segmentation and to stabilize the cluster models. The Viterbi re-alignment is done with a minimum duration constraint to avoid spurious short speaker turns. The stopping criterion for clustering is usually based on a threshold over the distance measure used for clustering. When the distance between any two clusters is greater than a certain threshold, the clustering stops.

2.6 Speaker diarization systems

This section provides a detailed description of some of the state-of-the-art speaker diarization systems proposed in the literature. The main focus is on the parametric HMM/GMM based system which has been shown to achieve state of the art performance in NIST-RT [National Institute of Standards and Technology, 2003] evaluation campaigns and a non-parametric system based on an information bottleneck framework [Vijayasenan, 2010] which gives comparable performance to that of the parametric system but has a significantly lower running time [Vijayasenan and Valente, 2012].

2.6.1 HMM/GMM system

A HMM/GMM based speaker-diarization system represents each speaker by a state of an HMM and models the state emission probabilities using GMMs. The clustering is initialized by uniform segmentation of detected speech regions in a given recording, which generates

a set of segments $\{X_i\}$. The agglomerative clustering is initialized by treating each segment as a cluster and is represented by a state in the HMM. Let c_i denote the ith speaker cluster (HMM state), b_i denote the emission probability distribution corresponding to speaker cluster c_i and s_t denote a feature vector at time t. Then, the log-likelihood $\log b_i(s_t)$ of the feature s_t for cluster c_i using a GMM is obtained as:

$$\log b_i(s_t) = \log \sum_{(r)} w_i^{(r)} N(s_t, \mu_i^{(r)}, \Sigma_i^{(r)}), \tag{2.8}$$

where $N()$ is a Gaussian pdf and $w_i^{(r)}$, $\mu_i^{(r)}$ and $\Sigma_i^{(r)}$ are the weights, means and covariance matrices respectively of the rth Gaussian mixture component of cluster c_i. Figure 2.3 outlines the HMM/GMM based diarization framework. Clustering in an agglomerative framework starts by over-estimating the number of speaker clusters. At each iterative step, the clusters that are most similar to each other based on a distance measure are merged. The similarity between two clusters is measured using a modified delta Bayesian information criterion (modified ΔBIC) [Ajmera et al., 2004]. In modified ΔBIC, the complexity term in the standard BIC [Schwarz, 1978] cancels out because the total number of model parameters before and after merging is kept constant. The modified ΔBIC criterion $\Delta BIC(c_i, c_j)$ for two clusters c_i and c_j is given by:

$$\Delta BIC(c_i, c_j) = \sum_{s_t \in \{c_i \cup c_j\}} \log b_{ij}(s_t) - \sum_{s_t \in c_i} \log b_i(s_t) - \sum_{s_t \in c_j} \log b_j(s_t) \tag{2.9}$$

where b_{ij} is the probability distribution estimated over the combined data of cluster c_i and c_j. The clusters that produce the highest ΔBIC score are merged. After each merge step, a Viterbi decoding pass realigns the speech data to the new speaker clusters. A minimum duration constraint on each state prevents rapid speaker changes. The clustering stops when no two clusters have a ΔBIC score greater than zero.

When multiple feature streams are present, they can be combined by simply appending the synchronized feature vectors of the individual feature streams and running diarization over the appended feature vector stream. Alternatively, a separate set of GMMs for each feature stream is estimated, and a weighted combination of the individual stream log-likelihoods gives the combined log-likelihood. For the case of two feature streams p and q, let $b_i^{(p)}$, $b_i^{(q)}$ denote the probability distributions estimated from streams p, q respectively for cluster c_i. The combined log-likelihood for cluster c_i is given by:

$$\log b_i(s_t^{(p)}, s_t^{(q)}) = w^{(p)} \log b_i^{(p)}(s_t^{(p)}) + w^{(q)} \log b_i^{(q)}(s_t^{(q)}), \tag{2.10}$$

where $s_t^{(p)}$, $s_t^{(q)}$ are the feature vectors corresponding to feature streams p, q respectively, $w^{(p)}$, $w^{(q)}$ are the weights of the feature streams, such that $w^{(p)} + w^{(q)} = 1$. The weights $w^{(p)}$, $w^{(q)}$ are typically optimized on a held out development data set. This type of combination has been shown to be very effective in combining MFCC features and TDOA feature streams. The multi-stream HMM/GMM diarization framework is depicted in Figure 2.4. The baseline

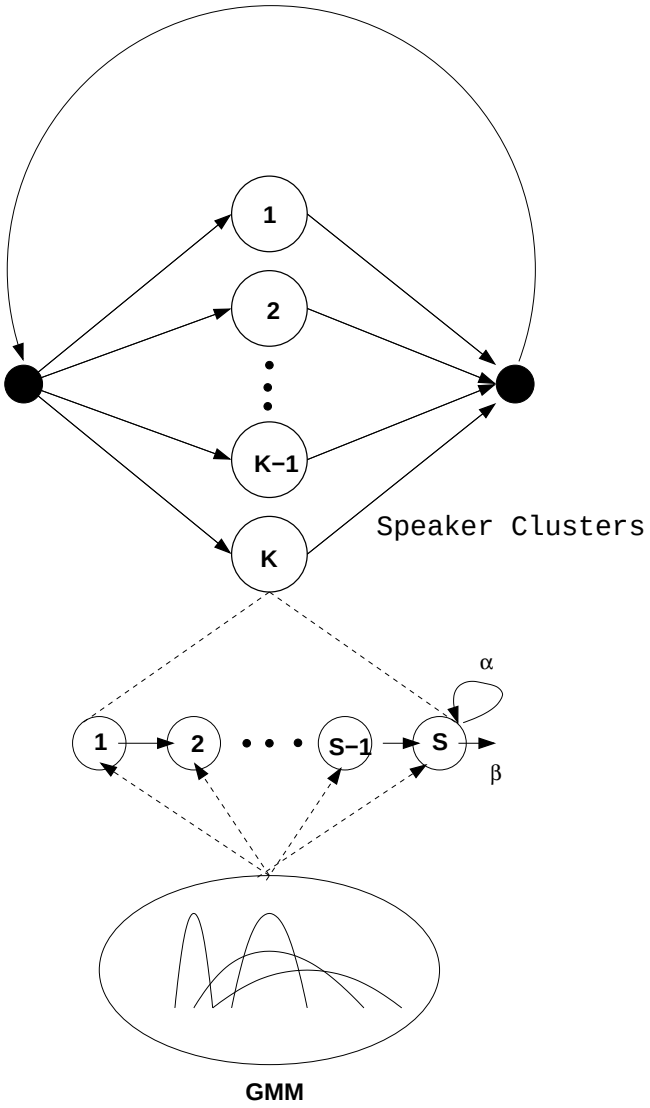


Figure 2.3: HMM/GMM speaker diarization system.

HMM/GMM diarization system used in the current study is modelled after the state-of-the-art system developed by ICSI [Wooters and Huijbregts, 2008].

2.6.2 Information bottleneck (IB) system

The information bottleneck (IB) diarization system [Vijayasenan et al., 2009, Vijayasenan and Valente, 2012, Vijayasenan, 2010] is a non-parametric system based on the IB clustering framework [Slonim et al., 1999]. The method has been shown to give similar performance to that of parametric systems based on the HMM/GMM framework [Wooters and Huijbregts, 2008] with the advantage of a significantly lower running time [Vijayasenan et al., 2009, Vijayasenan and Valente, 2012]. The IB method of clustering is a distributional clustering algorithm that clusters

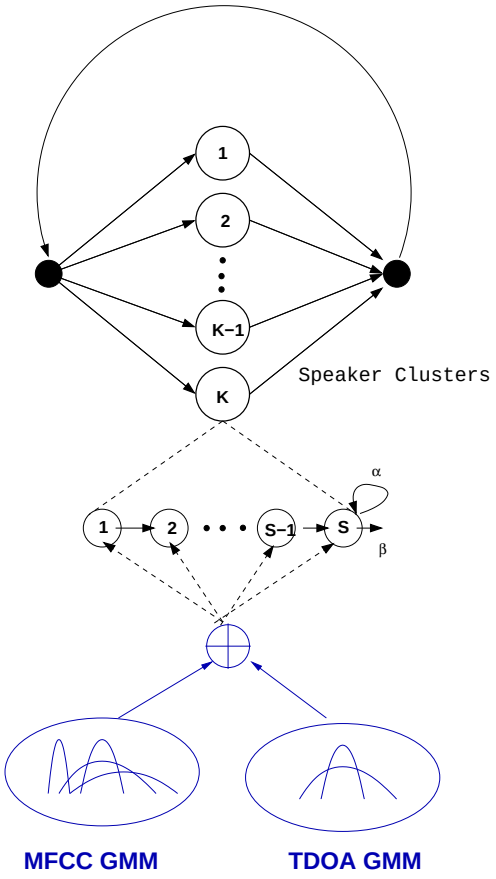


Figure 2.4: HMM/GMM speaker diarization system using multiple feature streams.

items with similar distributions over a set of variables known as relevance variables. It was initially applied to the task of document clustering, using the set of words in the documents as a relevance variable set [Slonim et al., 1999]. In this application, documents containing similar distributions over words were clustered together. In the case of speaker diarization, components of a background GMM estimated over speech regions of a given multi-party speech recording are used as a set of relevance variables. This is motivated from state-of-the-art methods in speaker identification where GMMs are used as universal background models (UBMs). Let $X = \{x_1, x_2, \ldots, x_M\}$ denote the set of input variables that need to be clustered and let $Y = \{y_1, y_2, \ldots, y_N\}$ denote the set of relevance variables that carry meaningful information about the desired clustering output $C = \{c_1, c_2, \ldots, c_P\}$. The IB method aims to find the optimal clustering by maximizing the function below:

$$\mathscr{F} = I(C,Y) - \frac{1}{\beta} I(C,X) \tag{2.11}$$

where β is a Lagrange multiplier, $I(C,Y)$ denotes mutual information between the set of relevance variables Y and the clustering output C, and similarly $I(C,X)$ denotes mutual information between the input variables set X and the clustering output C. By maximizing $\mathscr{F}$

in (2.11), the clustering algorithm aims at preserving as much information as possible about the relevance variables in the final clustering i.e., maximizing $I(C,Y)$ while being as compact as possible by minimizing mutual information $I(C,X)$ between the input variable set X and the clustering output C. The maximization is perfomed w.r.t the stochastic mapping $P(C|X)$. The IB function $\mathscr{F}$ can be maximized in several ways. The current system [Vijayasenan and Valente, 2012] uses a greedy agglomerative solution to the optimization. The clustering starts with a uniform (over) segmentation of speech regions, which are treated as set of input variables X. The set of relevance variables Y is formed by components of background GMM estimated over these speech regions. The posterior distribution of the relevance variables (components of background model) given the input segment x_i, $P(Y|x_i)$ is obtained by applying Bayes rule. The agglomerative clustering is initialized with each member of set X as an individual cluster and then at each clustering step of the IB method, the two clusters that have most similar distributions over the relevance variables are combined. The similarity is obtained in the form of loss in the IB function $\mathscr{F}$, resulting due to the merge of two clusters c_i, c_j as,

$$\nabla\mathscr{F}(c_i,c_j) = JS(P(Y|c_i),P(Y|c_j)) - \frac{1}{\beta}JS(P(X|c_i),P(X|c_j)) \tag{2.12}$$

where $JS()$ stands for the Jensen-Shannon divergence between two distributions, and is given by,

$$JS(P(Y|c_i),P(Y|c_j)) = \pi_i KL(P(Y|c_i),P(Y|c_{ij})) + \pi_j KL(P(Y|c_j),P(Y|c_{ij})) \tag{2.13}$$

where KL() stands for Kullback-Leibler divergence between two distributions, $\pi_i = \frac{P(c_i)}{P(c_i)+P(c_j)}$, $\pi_j = \frac{P(c_j)}{P(c_i+P(c_j))}$ and c_{ij} is the cluster formed after the merge of the clusters c_i and c_j. The relevance variable distribution of the cluster formed due to the merge is obtained by averaging the relevance variable distributions of the individual clusters in the merge. At each step, the two clusters that result in the lowest value of $\nabla\mathscr{F}$ are merged into one cluster. The stopping criterion is based on a threshold over the normalized mutual information $\frac{I(C,Y)}{I(X,Y)}$. Once the final clusters have been obtained, a re-alignment step is performed by estimating a GMM from the data assigned to each cluster and using these cluster models to perform Viterbi decoding with a minimum duration constraint. This step is intended to correct the errors in the segmentation introduced in clustering initialization by uniform segmentation of speech regions. Figure 2.5 depicts the agglomerative IB diarization framwork.

When multiple feature streams are given as input, the agglomerative IB diarization framework can be extended using aligned background GMMs for the feature streams. The background models for the feature streams have the same number of components. The parameters of components in the GMMs of feature streams are estimated using feature vectors from the same time indices in the respective feature streams. For example, if the r^{th} component parameters in the GMM of feature stream p are estimated from a set of features $\{s^p_{t'}\}$, the r^{th} component parameters in GMM of feature stream q are estimated from the set of features $\{s^q_{t'}\}$ that have the same time indices $\{t'\}$. The set of these aligned mixture components represents the relevance

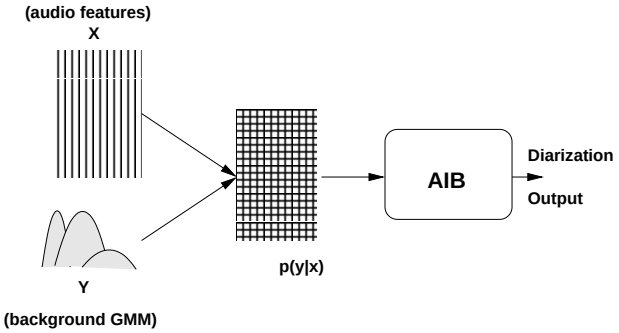


Figure 2.5: Agglomerative Information Bottleneck speaker diarization system.

variables. The estimation of the posterior distribution of the relevance variables $P(Y|x_i^p)$, $P(Y|x_i^q)$ is similar to the single feature stream case which is based on using Bayes' rule. The estimation of $P(Y|x_i^p, x_i^q)$ is obtained as a weighted average of the individual distributions given by:

$$P(Y|x_i^p, x_i^q) = w^p P(Y|x_i^p) + w^q P(Y|x_i^q) \tag{2.14}$$

where w^p and w^q represent respective feature stream weights such that $w^p + w^q = 1$. These weights are optimized on a separate development set of meetings. In the IB system, the combination of feature streams happens in the relevance variable space by taking a weighted average of posterior distributions of relevance variables given individual feature streams. This is depicted in Figure 2.6.

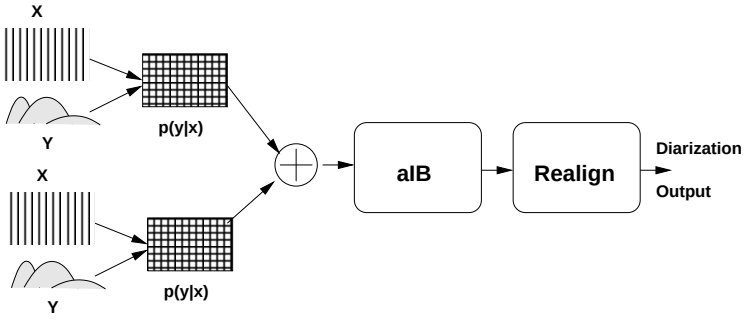


Figure 2.6: Agglomerative Information Bottleneck (aIB) speaker diarization system with multiple feature streams.

2.6.3 Other Approaches

The two approaches, HMM/GMM and IB based speaker diarization, are based on an agglomerative clustering framework. Several other approaches to speaker diarization have also been proposed in the literature. This section gives a brief overview of this type of methods.

Top down system

The top down/divisive clustering framework [Meignier et al., 2001, Fredouille and Evans, 2008, Fredouille et al., 2009] starts with a single model for the entire audio and progressively adds new speaker models until the optimal number of speakers is reached. This is done via selecting speech segments based on some criterion such as segment length. The segments that are obtained based on such criterion are used to represent new speaker clusters. After increasing the number of models, several iterations of Viterbi re-alignment and model re-estimation are performed to stabilize the speaker models. This process is repeated until no new speaker models need to be added. The top down approaches have not been as successful as the agglomerative clustering systems. But, recent works have shown that improvements in the top down approach can be obtained by cluster purification [Bozonnet et al., 2010b]. One advantage of using top down approaches is that they are computationally more efficient than the bottom up approaches.

Subspace methods for speaker diarization

Factor analysis based approaches [Kenny et al., 2008, Castaldo et al., 2007, Vogt and Sridharan, 2008] have been very successfully applied to tasks like speaker verification and identification and language recognition. Recently these methods have been adapted for speaker diarization [Kenny et al., 2010, Shum et al., 2011]. In these approaches, the goal is to separate the speaker and channel variabilities to determine the low dimensional speaker identity vector, referred to as the i-vector. This is performed using GMM supervectors, which are constructed by appending the means of a GMM components. The speaker and channel dependent supervector S is defined as:

$$S = m + Tv + \epsilon \tag{2.15}$$

where m is the speaker and channel independent supervector computed from a Universal Background Model (UBM) which is a GMM; T is a rectangular, low rank matrix of the total variability subspace, v is a low-dimensional vector, referred to as an i-vector, and ϵ is residual noise. Once the i-vectors of two speech segments X_i and X_j are obtained, their similarity is measured using the cosine distance between the respective i-vectors v_i and v_j as follows:

$$cos(v_i, v_j) = \frac{v_i^T v_j}{||v_i||.||v_j||} \tag{2.16}$$

The i-vectors are computed over short segments of around one second in duration. In [Shum et al., 2011], the i-vectors are extracted and the number of speakers is determined via spectral clustering. Then, k-means clustering based on the cosine distance is used to cluster the i-vectors (and their corresponding segments). After clustering, a number of post-processing steps are performed before obtaining the final segmentation.

Fast diarization

Fast diarization systems with low computational complexity are desired for some applications like meeting summarization and analysis in real time. This would enable realization of several applications (meeting browsing, speaker retrieval) on normal desktop machines. However, conventional diarization systems depend on parametric models to compute the BIC distance measure for diarization and require the re-estimation of parametric models for every possible cluster pair merge. This is computationally very expensive and requires considerable optimization to make it feasible for real time speaker diarization.

Apart from the IB system explained in Section 2.6.2, several other methods have been proposed in the literature to achieve fast speaker diarization. The speaker binary keys [Anguera and Bonastre, 2010] based framework was used for speaker diarization in [Anguera and Bonastre, 2011]. This method achieves a significant reduction in running time with comparable performance to state-of-the-art systems. Speaker binary keys are small binary vectors computed from the acoustic data using a Universal Background Model (UBM). Once they are computed all the processing takes place in the binary domain. Other works concerned with speeding up the speaker diarization systems include [Huang et al., 2007]. This system achieves faster than real-time processing through the use of several processing tricks applied to a standard agglomerative clustering approach. In [Friedland et al., 2010] the speed-up is achieved by parallelizing most of the processing in a GPU.

2.7 Resources for speaker diarization

Growing interest in automatic processing of natural multi-party conversations has resulted in a number of efforts to create resources for such tasks. Specifically, large scale efforts [Mccowan et al., 2005, Janin et al., 2003, Mostefa et al., 2007] were invested by multiple institutions in collection of natural multi-party conversations in a conference style meeting room environments. These conversations were collected in specially equipped meeting room environments with different types of microphones (both close talking and distant) to capture the audio. In some scenarios, videos of these conversations were also recorded to faciliate the analysis of non-verbal behavior of the participants. In this scenario, focussed efforts were applied to drive state-of-the-art in tasks like automatic speech recognition, speaker recognition and diarization, and detecting topic boundaries which result in generating rich transcription of the audio captured. To benchmark various algorithms proposed to solve these problems, evaluation campaigns with standardised data sets were conducted by NIST [National Institute of Standards and Technology, 2003]. The campaigns defined the task scenarios and evalution metrics used to compare different systems on these tasks.

For the experiments in the present work, we use meeting room recordings from three different corpora namely, AMI, NIST-RT and ICSI. The details of all the three data sets are summarized in Tab. 2.1.

2.7.1 Multi-party meeting corpora

AMI meeting corpus

The AMI corpus [Mccowan et al., 2005] consists of about 100 hours of meeting recordings recorded at multiple sites (Idiap, TNO, Edinbrough). The corpus contains both natural and scenario based meetings. The corpus is annotated at multiple levels with several low-level and higher-level information such as word transcripts, speaker segments, summaries, dialogue acts, topic changes etc.

NIST RT meeting corpus

The NIST RT meeting corpus [National Institute of Standards and Technology, 2003] consists of meetings belonging to NIST RT speaker diarization evaluation campaigns of 2005, 2006, 2007 and 2009. Each set consists of meeting recordings from multiple sites and varying number of speakers. The corpus also contains ground-truth speaker segmentation obtained by force-aligning the manual transcripts of individual head-microphone channels.

ICSI meeting corpus

The ICSI meeting corpus [Janin et al., 2003] contains 75 meetings comprising on the whole around 72 hours of spontaneous multi-party conversation recordings recorded at ICSI, Berkeley. All meetings were recorded by both close talking and distant microphones and contain word level orthographic transcription and speaker information along with other annotations like dialogue acts, hotspots and topic changes.

Table 2.1: *Details of meeting room conversations corpora.*

Name	Size	No. of speakers per meeting	Microphones used	Meeting rooms
AMI	100 hours	4	circular array	3
ICSI	72 hours	3-10(on avg. 6)	4 table top	1
NIST-RT	15 hours	4-11(on avg. 5)	ad-hoc	8

2.7.2 Evaluation Metric

The evaluations campaigns organized by NIST have defined a metric known as use Diarization Error Rate (DER) to measure the performance of speaker diarization systems. This metric makes use of a ground-truth reference segmentation to evaluate system generated diarization output. It is measured as the fraction of time that is not assigned correctly to a speaker or nonspeech. The computation of this metric is detailed below.

Since the sets of speaker labels provided by reference and hypothesis from the system are arbitrary, an optimal mapping between the two sets is first constructed such that the overlap

time between corresponding speakers is maximum. The DER is then calculated as:

$$DER = \frac{\Sigma_{allseg}\{dur(seg)[max(N_{ref}(seg), N_{hyp}(seg)) - N_{correct}(seg)]\}}{\Sigma_{allseg} dur(seg) N_{ref}(seg)} \tag{2.17}$$

where $N_{ref}(seg)$ and $N_{hyp}(seg)$ denote the number of speakers in reference and hypothesized (system generated) segmentation for speech segment seg with duration $dur(seg)$. $N_{correct}$ denotes the number of speakers that are correctly mapped between reference and hypothesized segmetations. Sements that are labelled as non-speech are considered to have zero speakers. When all speakers are correctly matched between the reference and hypothesized segmentation for a segment, the corresponding DER is zero. The three main components in DER are:

- **False-alarm error:** This is the percentage of time that a hypothesized speaker is labelled as non-speech in the reference. It can be expressed as:

$$E_{FA} = \frac{\Sigma_{seg:N_{hyp}>N_{ref}}\{dur(seg)[N_{hyp}(seg) - N_{ref}(seg)]\}}{\Sigma_{allseg} dur(seg) N_{ref}(seg)} \tag{2.18}$$

- **Missed speech error:** This is the percentage of time that a reference speaker segment is mapped to no hypothesized speakers. Missed speech thus includes undetected speakers. It is given by:

$$E_{MISS} = \frac{\Sigma_{seg:N_{hyp}<N_{ref}}\{dur(seg)[N_{hyp}(seg) - N_{ref}(seg)]\}}{\Sigma_{allseg} dur(seg) N_{ref}(seg)} \tag{2.19}$$

- **Speaker error:** This represents the percentage of time the hypothesized speaker label is mapped to a wrong speaker label in the reference segmentation. This is given by:

$$E_{SPKR} = \frac{\Sigma_{allseg}\{dur(seg)[min(N_{ref}(seg), N_{hyp}(seg)) - N_{correct}(seg)]\}}{\Sigma_{allseg} dur(seg) N_{ref}(seg)} \tag{2.20}$$

Given these three types of errors, the DER shown in (2.17) can be re-written as:

$$DER = E_{FA} + E_{MISS} + E_{SPKR} \tag{2.21}$$

The segment boundaries are evaluated upto an accuracy of a predefined collar that accounts for the inexact boundaries in the labelling. This value was fixed by NIST at 0.25s.

DER for a dataset with multiple meetings is computed from individual meeting DER values by performing a linear combination of individual DER values. The weight for each meeting is proportional to the total scored time ($\Sigma_{allseg} dur(seg) N_{ref}(seg)$) of the meeting.

3 Overlapping speech handling for Speaker diarization

3.1 Introduction

Overlapping speech occurs when there is more than one speaker speaking at any given instant of time in an audio recording. This is a very common phenomenon in spontaneous conversations like meeting room discussions, telephone conversations, television chat shows and other similar media [Shriberg, 2005, Adda-Decker et al., 2008]. The factors causing overlapping speech in a multi-party conversation are diverse [Cetin and Shriberg, 2006b, Cetin and Shriberg, 2006a, Schegloff, 2000, Kurtic et al., 2013]. It can occur when listeners use back-channels to show their involvement and also to convey their agreement with the foreground speaker. It also most commonly occurs when one or more participants try to interrupt the foreground speaker and take the conversation floor. In informal conversations among multiple participants, there are situations where multiple parallel conversations (schism) [Sacks et al., 1974] take place involving several participants in each sub-conversation. Also, it has been observed that overlaps are a common phenomenon during conversation floor exchanges among speakers. Apart from the above mentioned broad patterns of occurrences of overlaps, there are always some idiosyncrasies specific to a particular conversation or participant that can cause overlapping speech at any place during the conversation. Previous studies have shown that the error rates of automatic speech processing systems increase when processing speech from multiple simultaneous speakers [Shriberg et al., 2001, Cetin and Shriberg, 2006b]. Several diagnostical studies on speaker diarization systems have also shown that overlapping speech is one of the main sources of error in state of the art speaker diarization systems [Huijbregts and Wooters, 2007, Huijbregts et al., 2012, Knox et al., 2012].

3.2 Previous approaches for overlap detection

Several previous works have proposed methods to detect overlapping speech in meeting room conversations. Earlier works have concentrated on detecting overlapping speech in audio captured using head/lapel microphones worn by the participants in the conversation [Wrigley

et al., 2005, Pfau et al., 2001, Laskowski and Schultz, 2006]. These works have focussed on issues arising from cross-talk, breath noise and channel variations across different close talking microphones used to capture the audio. Pfau et al. [Pfau et al., 2001] have proposed a hidden Markov model (HMM) based approach to infer the sequence of hidden states speech and non-speech in each participant's channel. They have also proposed a method to detect overlapping speech segments by putting a threshold on the cross-correlation value between speech signals of multiple channels.

Wrigley et al. [Wrigley et al., 2005] proposed a more generalized approach to multi-channel speech activity detection, where a HMM with four states, single-speaker speech, cross-talk, overlapping speech and non-speech was used. They explored various acoustic features useful for this task and found kurtosis, 'fundamentalness' and cross-correlation related features to be most effective. Laskowski et al. [Laskowski and Schultz, 2006] have proposed a method to improve multi-channel speech activity detection in overlapping speech regions by modelling the turn taking behavior of participants in the conversation. These works were mainly concerned with improving speech activity detection on the meeting room data captured by close-talking microphones with an aim to facilitate reliable automatic speech recognition (ASR). Speaker diarization in its standard setup is evaluated on distant microphone speech where there is no one-to-one correspondence between the speakers and the channels used in the recording. So, recent works on overlap detection and speaker diarization have focussed on detecting the overlapping speech in recordings captured using distant microphones.

Otterson et al. [Otterson, 2008] trained overlapping speech models using synthesized overlapping speech obtained by adding multiple single-speaker speech utterances. But, experiments revealed that though the trained models were effective in detecting artificially synthesized overlaps, they did not generalize well to naturally occurring overlaps in meeting conversations. The same authors [Otterson and Ostendorf, 2007] proposed a two step method to handle overlapping speech in speaker diarization assuming oracle overlap detection. Boakye et al. [Boakye et al., 2008, Boakye et al., 2011, Boakye, 2008] explored various acoustic features for overlapping speech detection in distant microphone audio. They found features such as Mel-frequency cepstral coefficients (MFCC), energy, spectral flatness as being the most useful features for overlap detection. Experiments on meeting recordings from AMI meeting corpus [Mccowan et al., 2005] have shown that overlap detection is possible with reasonable accuracy which in turn reduced the diarization error rate (DER).

Huijbregts et al. [Huijbregts et al., 2009] proposed a method for overlapping speech detection in a two pass speaker diarization system by training overlapping speech models from speech surrounding speaker changes hypothesized by an automatic speaker diarization system. The trained overlap model was used in the second pass along with speaker models in Viterbi decoding to identify overlapping speech segments. Zelenak et al. [Zelenák et al., 2010, Zelenák et al., 2012, Zelenák, 2011] have proposed the use of time delay of arrival (TDOA) based features extracted from the cross-correlation of speech signals captured by multiple distant microphone channels to improve short-term spectral feature based overlap detection. The

cross-correlation based features were used along with short-term spectrum based features as two parallel feature streams in a multistream overlapping speech detection system. To reduce the dimensionality of the cross-correlation based feature vector, and also to make it independent of the number of distant microphone channels used for the recording, they have proposed a way to transform these features to a fixed dimension feature space by applying principal component analysis (PCA) or artificial neural networks (ANNs). Experiments have shown that the cross correlation based features improve the overlapping speech detection. Zelenak et al. [Zelenák and Hernando, 2011] have also shown improvements in overlapping speech detection by the use of prosodic features. More recently, convolutional non-negative sparse coding based approaches have been successfully applied to the problem of overlap detection [Vipperla et al., 2012, Geiger et al., 2012].

3.3 Motivation for the proposed features

All the methods described above rely directly on features computed from the acoustic signal captured by distant microphones. However, in the recordings of meeting room conversations by distant microphones, the speech signal is often corrupted by background noise resulting in a recording with low signal to noise ratio (SNR). In such scenarios, it is important to explore higher level information present in the structure of a conversation which carries useful cues in modelling the occurrence of overlapping speech. Such information if captured effectively, could potentially be more robust to noisy conditions in the recording. It can also be easily transferred across different corpora of multi-party conversations irrespective of the recording conditions and meeting room setup, which have a direct influence on acoustic features. Studies on conversational analysis have shown that overlaps are more likely to occur at specific locations in a conversation such as speaker turn changes [Shriberg et al., 2001, Cetin and Shriberg, 2006a] and have also shown that single-speaker speech, silence and overlapping speech patterns are related to each other [Laskowski et al., 2008, Laskowski et al., 2007, Laskowski, 2010]. This phenomenon is illustrated in Figure 3.1 which shows the speaking status of four speakers in a meeting snippet.

Motivated by these studies, the present work proposes the use of features that can be easily extracted automatically from conversations such as silence [Yella and Valente, 2012] and speaker change statistics [Yella and Bourlard, 2013] to capture higher level information in a conversation that is relevant for overlap detection. These features are extracted from a long context of about 3–4 seconds surrounding a given time instant and are used to estimate the probability of occurrence of overlap at that instant. We also explore methods to combine the complementary information present in the two features, silence and speaker change statistics, to improve upon the probability estimates obtained from either of the individual features. These probability estimates are incorporated into the acoustic feature based classifier as prior probabilities of the overlapping and single-speaker speech classes. We report experiments on the AMI [Mccowan et al., 2005], NIST-RT [National Institute of Standards and Technology, 2003] and ICSI [Janin et al., 2003] meeting corpora to validate the generalizability of the

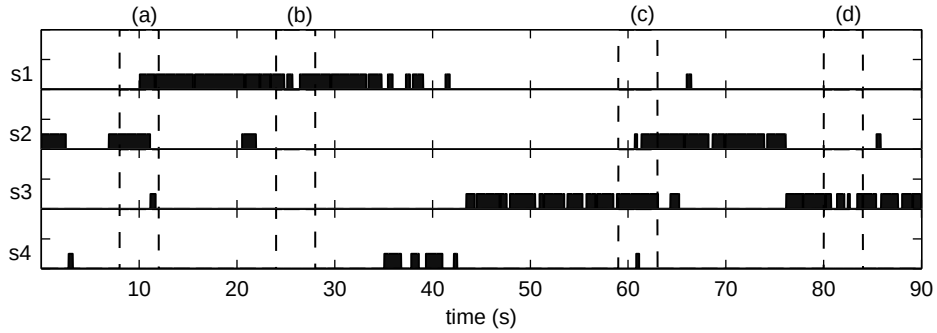


Figure 3.1: Speaker vocalizations from a snippet of multi-party conversation. The fixed length segments (a) and (c) are in regions of speaker change and contain overlap as well as less amount of silence whereas segments (b) and (d) contain single speaker speech, are reasonably away from any speaker changes, and also contain more silence.

Table 3.1: *Number of meetings used as part of train, development and test sets in each data set.*

Corpus	Train	Dev	Test
AMI	35	10	25
ICSI	35	-	15
NIST-RT	27	-	7

proposed method in overlap detection. In these experiments, we demonstrate that the model that is trained to estimate the probability of occurrence of overlap using the meetings from AMI corpus generalizes to other meeting corpora such as the NIST-RT and the ICSI. We also report speaker diarization experiments on the three data sets to evaluate the subsequent improvements in speaker diarization achieved due to improvements in overlap detection.

3.4 Data-sets used

For the experiments in the present work, we used meeting room recordings from three different corpora namely, AMI [Mccowan et al., 2005], NIST-RT [National Institute of Standards and Technology, 2003] and ICSI [Janin et al., 2003]. The train and test set splits are described in Table 3.1.

The train and test splits in AMI and ICSI corpus were obtained by random selection. As described earlier, AMI data set contains meetings recorded at three different locations. The train, dev and test sets of AMI corpus contain meetings from all these three sites. For the experiments on the ICSI data set, we used meetings from the groups Bmr, Bro, Bed. In the NIST-RT corpus, RT-05, RT-06, RT-07 data sets are used as part of train set, while RT-09 data set is used for testing. The meetings from the training set of each corpus are used to train the acoustic feature based overlap detector, which was later used for overlap detection on the respective test sets. The ground-truth segmentation obtained by force-aligning the manual

transcripts with close talking microphone audio is used to get the overlap and single-speaker speech boundaries.

3.5 Overlap speech diarization

This section presents the ways in which overlapping speech effects diarization output and reviews methods proposed in the literature to handle overlapping speech in diarization. It also provides details of acoustic feature based HMM/GMM overlap detector.

3.5.1 Overlap handling for speaker diarization

Overlapping speech causes errors in speaker diarization systems in two ways. First, it introduces impure segments containing speech from multiple speakers into the clustering process. Secondly, the overlap segments are scored n times, where n is the number of speakers in the overlap, which increases the missed speech error even if one of the speakers is not assigned correctly. To avoid these errors, Otterson et al. [Otterson and Ostendorf, 2007] have proposed a method to handle overlaps for speaker diarization that consists of two steps which are overlap exclusion and overlap labelling. In overlap exclusion, detected overlap segments are excluded from the clustering process so that they do not corrupt the speaker models. In overlap labelling, the segments detected as overlap are labelled by two speakers according to a heuristic such as assigning the overlap segment to the two nearest speakers in time. This heuristic is based on observations from studies on NIST RT meeting data which revealed that assigning speakers based on proximity to the overlap gives a significant reduction in DER [Otterson and Ostendorf, 2007]. The two speakers can also be assigned based on cluster likelihoods, in which the segment is assigned to two clusters for which the data in the overlap segment has highest likelihood [Boakye et al., 2008, Zelenák et al., 2010]. The overlap handling method in a typical speaker diarization system is summarized in the Figure 3.2.

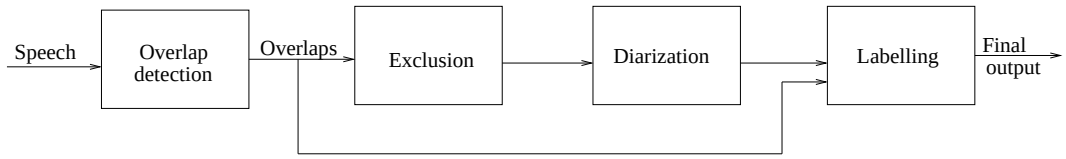


Figure 3.2: Block diagram of overlap handling for speaker diarization: Overlaps obtained from overlap detection are first excluded from the diarization process and later, appropriate speaker labels are assigned to the overlap segments based on methods described in Section 3.5.1 in the labelling stage, to generate the final diarization output.

3.5.2 Overlap detection using acoustic features

The baseline overlap detection system is based on acoustic features derived from short-term spectrum that have been shown to be effective in classifying single-speaker speech and

overlapping speech in the literature [Boakye et al., 2011]. In the present work, we use 12 Mel frequency cepstral coefficients (MFCCs) with log energy, spectral flatness, linear prediction (LP) residual energy computed from 12th order LP analysis, along with their deltas, resulting in a feature vector of dimension 30. All features are extracted from an acoustic frame of size 30 ms with a frame rate of 10 ms. Prior to overlap detection, automatic speech/non-speech detection is performed using the SHOUT algorithm [Huijbregts and de Jong, 2011] and non-speech regions are excluded from further processing. After this single-speaker, overlapping speech detection is performed on the detected speech regions in the recording. In baseline overlap detection system, the classes single-speaker and overlapping speech are represented by states of a hidden Markov model (HMM). The emission probability distributions of these states are modelled using Gaussian mixture models (GMMs). To control the tradeoff between the false overlap detections and total number of overlaps being detected, an overlap insertion penalty (OIP) is introduced which penalizes the overlap detections. High value for OIP hass a positive effect on the precision of the classifier while effecting the recall in a negative manner while the lower values increase the recall while sacrificing precision.

Let $V \in \{ov, sp\}$ denote the set of HMM states representing overlapping (ov) and single-speaker (sp) speech classes, and let F denote the sequence of acoustic feature vectors. Then, the most likely sequence of the states in a given audio recording can be obtained by Viterbi decoding as,

$$V^* = \arg\max_V P(V|F) = \arg\max_V P(F|V)P(V) \tag{3.1}$$

The prior probability $P(V)$ of the state sequence is usually approximated by a first order Markov model assuming that the current state is dependent only on its previous state and is represented by the state transition probability. In the baseline overlap detection system, these probabilities are fixed to a constant value based on the statistics observed in the training data. But studies on conversational analysis show that the probability of occurrence of an overlap is not constant across the recording, and there are places in a conversation where overlaps are more likely to occur [Shriberg et al., 2001, Cetin and Shriberg, 2006a]. In the current work, we propose a method to capture this information and use it in the acoustic feature based overlap detection system.

3.6 Conversational features for overlap detection

In this section, we present in detail the long-term features extracted from the structure of a conversation that are supposed to carry relevant information about occurrence of overlap. We explore features that can be easily extracted automatically such as silence and speaker change statistics.

3.6.1 Silence statistics for overlap detection

Several studies on multi-party conversational analysis have shown that single-speaker speech, silence and overlap patterns in a conversation are related to each other and carry useful information about the conversations and the participants in the conversations [Laskowski et al., 2008, Laskowski et al., 2007, Laskowski, 2010]. Motivated by these studies, we explore the relation between silence, single-speaker speech, overlapping speech durations in a segment to predict the occurrence of overlap in that segment. In particular we hypothesize that segments containing more silence are less likely to contain overlapping speech. To verify this hypothesis, we perform experiments using the AMI training set meetings.

Let D be a variable indicating the duration of silence in a segment and $n^S(D = d)$ be the number of segments that contain d seconds of silence where, the length of the segment is given by the variable S. Let $V \in \{ov, sp\}$ be a binary variable denoting the classes we are interested in detecting, which are overlapping speech (ov) and single-speaker speech (sp). Let $n^S(V = ov, D = d)$ denote the number of segments of length S seconds that contain d seconds of silence and an occurrence of overlap. Given these counts, it is possible to estimate the probability of overlap in a segment conditioned on the amount of silence in that segment as:

$$P^S(V = ov|D = d) = \frac{n^S(V = ov, D = d)}{n^S(D = d)} \tag{3.2}$$

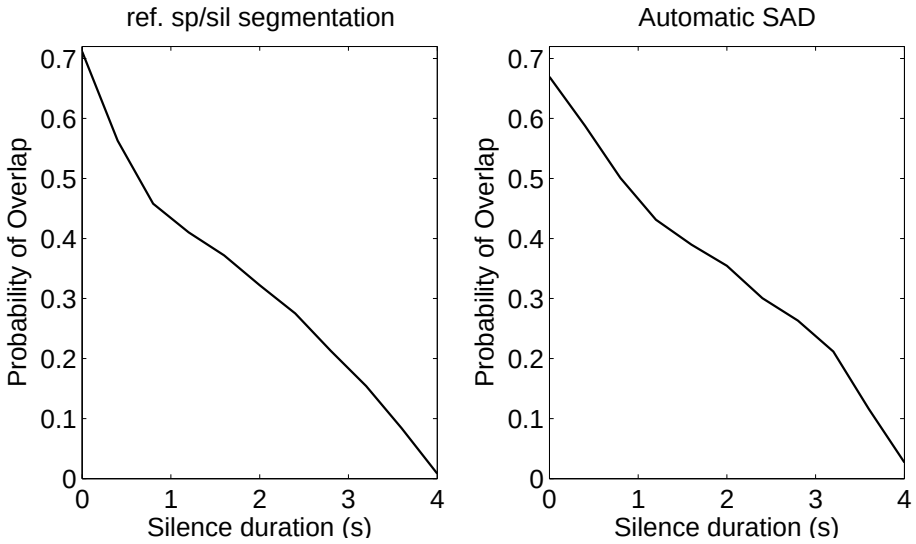


Figure 3.3: *Probability of overlap based on silence duration obtained using ground truth speech/sil segmentation and automatic speech activity detector output.*

Figure 3.3 shows $P^S(V = ov|D = d)$ for different values of d (silence duration) for a segment length of four seconds (i.e., $S = 4$). In the left plot, speech/silence segmentation is obtained from the ground-truth segmentation and in the right plot, it is obtained from the automatic speech activity detector (SAD) output [Huijbregts and de Jong, 2011]. It can be noticed from

Figure 3.3 that the probability of overlap in a segment is inversely proportional to the amount of silence in the segment which supports our hypothesis that segments containing more silence are less likely to contain overlapping speech. In particular, from the left subplot of Figure 3.3, it can be observed that when the amount of silence in a segment is zero, the probability of occurrence of overlap in that segment is around 0.7. In other words, this illustrates that it is possible to estimate the probability of occurrence of an overlap in a segment by the amount of silence present in that segment. This information is potentially useful as speech/silence detection is a simpler task compared to single-speaker/overlapping speech detection. The right subplot in Figure 3.3, which uses the SAD output to compute the amount of silence in a segment, also shows similar trends which means that ground-truth silence duration can be replaced by the estimates of silence from the SAD output. The probability of a single-speaker speech within a segment can be estimated as:

$$P^S(V = sp|SIL = x) = 1 - P^S(V = ov|D = d). \tag{3.3}$$

To compute these statistics for the whole recording, the segment is progressively shifted by one frame at each step. The probabilities $P_i^S(ov|d_i), P_i^S(sp|d_i)$ are estimated for $i \in \{1 \dots N\}$ where N is the total number of frames in the recording and d_i denotes the duration of silence in the segment centered around frame i. This process is depicted in Figure 3.4.

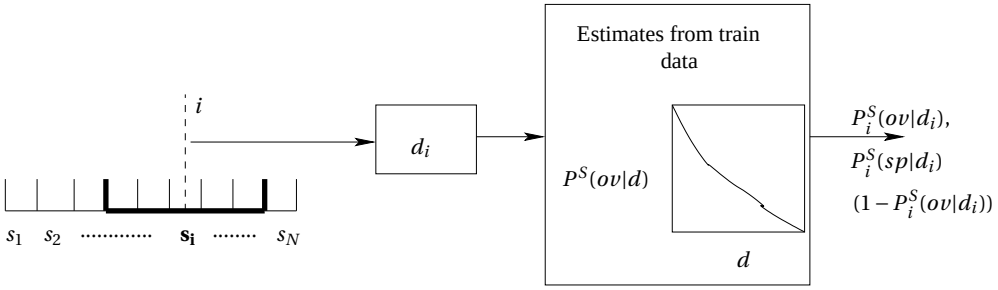


Figure 3.4: *Estimation of probabilities of single-speaker speech and overlapping speech states for a frame i based on duration of silence d_i present in the segment s_i centered around the frame i.*

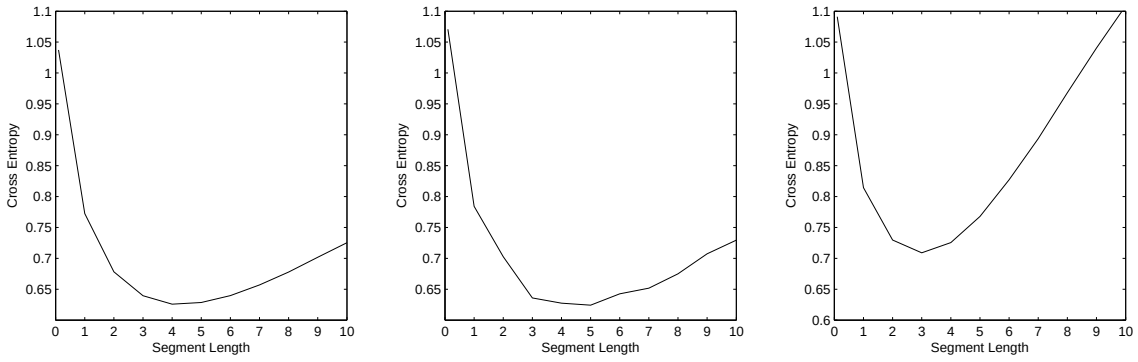
To verify how the estimated probabilities $P^S(ov_i|d_i)$ and $P^S(sp_i|d_i)$ generalize to sets of meetings that are different from those used during training, the cross-entropy between these estimated probabilities and the true distribution of the classes in a development set of meetings is computed. The probabilities for the true distributions are obtained for each frame $i \in \{1 \dots N\}$ as follows, $P^T(ov_i) = 1, P^T(sp_i) = 0$ if the frame i is overlapped and $P^T(sp_i) = 1, P^T(ov_i) = 0$ if the frame i belongs to single-speaker speech. The knowledge of whether a frame i belongs to the overlapped ($\{OV\}$) or single-speaker ($\{SP\}$) class is obtained from the ground-truth segmentation of these meetings. The cross-entropy between the true distribution and the

estimated distribution is computed as follows:

$$C(S) = -\frac{1}{L}\Big(\sum_{i\in\{OV\}} log(P_i^S(ov|d_i)) + \sum_{j\in\{SP\}} log(P_j^S(sp|d_j))\Big) \tag{3.4}$$

where L is the total number of frames used in the computation. To eliminate the bias in the estimate of $C(S)$ resulting from an uneven number of samples present in single-speaker speech and overlap classes, the cross-entropy measure is computed by considering an equal number of samples from each class.

Figure 3.5: Cross-entropy between estimates based on silence duration (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) NIST-RT 09 (c) ICSI.



(a) (b) (c)

It can be observed from Figure 3.5 (a) that the cross entropy decreases as we increase the segment length (S) used to estimate the probabilities of overlap until some point and then starts to increase again. This decrease in the cross-entropy suggests that the estimated probabilities are much closer to the true probabilities when they are estimated from a longer context than a frame. The lowest value of cross entropy is found around a segment length of 4 secs. This indicates that a segment length of 4 secs is optimal to compute the silence statistics. Similar plots of cross entropy are also shown for the RT 09 and the ICSI meetings respectively in subplots (b) and (c) of Figure 3.5. Note that the probability estimates for the NIST-RT 09 and the ICSI meetings are obtained using the model learned from AMI training set. These plots also show similar trends to those observed on the AMI development set, which gives an indication that the estimated statistics have similar effects on other data sets. It also shows that the model learned to estimate the probabilities of overlapping and single-speaker speech from the AMI training set can be generalizable to other meeting corpora such as NIST-RT and ICSI.

3.6.2 Speaker change statistics for overlap detection

Studies on conversational analysis have shown that overlaps occur more often at some specific parts of conversations [Shriberg et al., 2001]. Especially, it was shown that a significant proportion of the overlaps occurs during speaker turn changes [Shriberg et al., 2001]. Motivated by these studies, the current work analyzes the relationship between the occurrence of overlap in a segment and the number of speaker changes in the segment. Specifically, the study hypothesizes that the overlap probability in a segment is directly proportional to the number of speaker changes in the segment. In other words, segments containing more speaker changes are highly probable to have larger number of overlaps than those having fewer speaker changes. To verify this hypothesis, we perform experiments using the AMI training set.

In the first experiment, the distribution of the number of overlaps is analyzed for different number of speaker changes in a segment. Let C and O respectively denote the variables indicating the number of speaker changes and overlaps in a segment. In the present work, an occurrence of overlap is defined as a contiguous segment of overlapping speech surrounded by single-speaker speech or silence regions. The number of overlaps is obtained by counting such occurrences in the segmentations (obtained by force-aligning close talking microphone audio with manual transcripts) provided by the corpus authors. Let $n^S(C = c)$ denote the number of segments of length S seconds which contain c speaker changes and, let $n^S(O = o, C = c)$ denote the number of segments containing o overlaps and c speaker changes. Then, the probability $P^S(O = o|C = c)$ of having o overlaps in a segment of length S seconds conditioned on the fact that it contains c speaker changes can be estimated as:

$$P^S(O = o|C = c) = \frac{n^S(O = o, C = c)}{n^S(C = c)} \tag{3.5}$$

Figure 3.6 shows the distribution of $P^S(O|C = c)$ i.e., the distribution of the number of overlaps (O) in segments of length six seconds ($S = 6$) for different numbers of speaker changes (c). The speaker changes are obtained from the ground-truth speaker segmentation and automatic diarization output for left and right subplots respectively. It can be observed from Figure 3.6 that, as the number of speaker changes increases, the probability of occurrence of more overlaps also increases. Also, it can be observed that the distribution of $P^S(O|C = c)$ for different c seem to follow a Poisson distribution with a rate that is directly proportional to the number of speaker changes (c). The number of speaker changes in the diarization output is lower when compared to ground truth speaker segmentation due to constraints and errors introduced by the automatic system. Nevertheless, a similar phenomenon can also be observed for distributions estimated from the diarization output. Figure 3.6 supports our hypothesis that segments containing more speaker changes contain more overlaps. This information can be useful when incorporated into the baseline acoustic feature based overlap detector, since it does not contain evidence from the conversational patterns in the meetings.

Motivated by the empirical distributions in Figure 3.6, we model the probability of the number

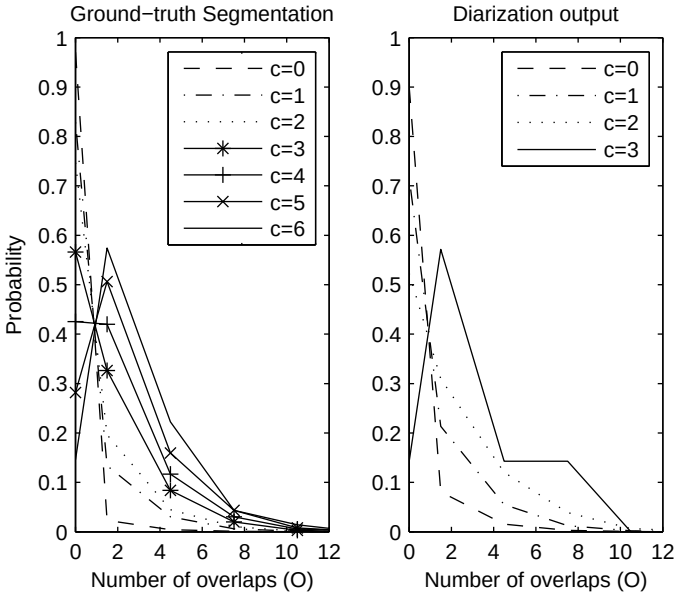


Figure 3.6: *Probability distributions of number of occurrences of overlaps (O) for different numbers of speaker changes (c) obtained using ground-truth segmentation and diarization output.*

of overlap occurrences in a given segment by a Poisson distribution whose rate λ_c^S depends on the number of speaker changes c in the segment S i.e.,

$$P^S(O=o|c) = \frac{(\lambda_c^S)^o e^{-\lambda_c^S}}{o!} \tag{3.6}$$

where the rate parameter λ_c^S is a maximum likelihood estimate from the training set of meetings. This estimate is simply the mean of the number of occurrences of overlaps in segments of length S seconds which contain c speaker changes. After estimating the set of rate parameters $\{\lambda_c^S\}$ for different values of c, the probability of overlap occurrence in a segment conditioned on the number of speaker changes in the segment can be obtained as,

$$P^S(V=ov|c) = 1 - P^S(O=0|c) \tag{3.7}$$

$$= 1 - e^{-\lambda_c^S} \tag{3.8}$$

and, the probability of single speaker speech can be obtained as,

$$P^S(V=sp|c) = 1 - P^S(V=ov|c) \tag{3.9}$$

$$= e^{-\lambda_c^S}. \tag{3.10}$$

The probabilities $P_i^S(V=ov|c_i)$ and $P_i^S(V=sp|c_i)$ are estimated for all the frames i in a given

recording as depicted in Figure 3.7. The cross-entropy between the estimated probabilities

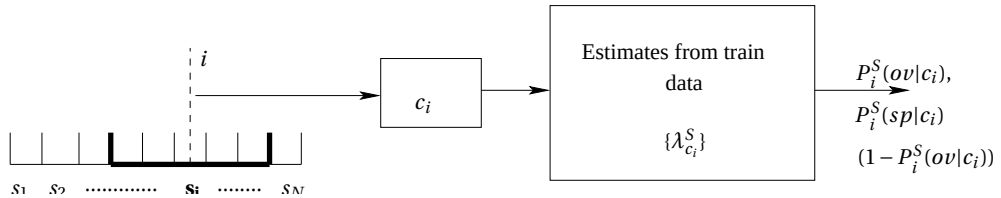
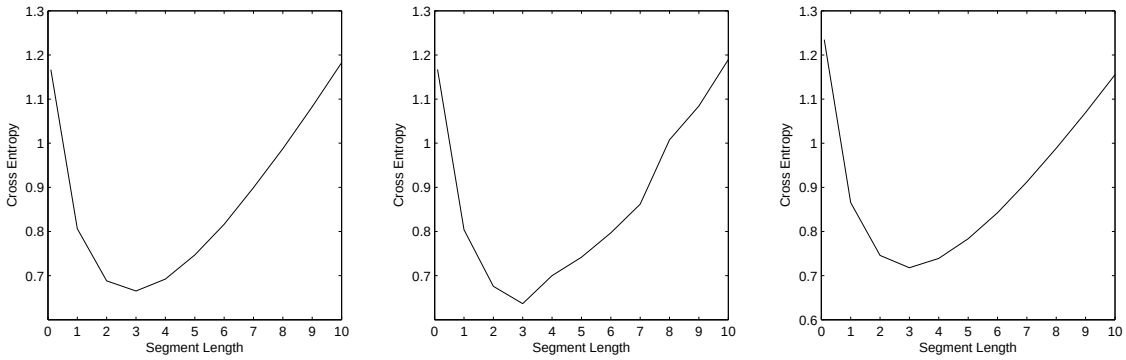


Figure 3.7: *Estimation of probabilities of single-speaker and overlapping speech classes for a frame i based on number of speaker changes c_i present in the segment s_i centered around the frame i.*

and the true distribution is computed by replacing the probability estimates in (3.4) by the estimates based on speaker changes $P^S(ov|c)$ and $P^S(sp|c)$. Figure 3.8 shows the cross entropy values for various segment lengths S computed on the AMI, NIST-RT 09 and ICSI data sets. It

Figure 3.8: Cross entropy measure between estimates based on speaker changes (obtained from model learnt from AMI-train set) and true distribution on various datasets (a) AMI development set (b) RT 09 (c) ICSI for various segment lengths.



(a) (b) (c)

can be observed from Figure 3.8(a) that a segment length of three seconds yields optimal estimates of the probabilities on the AMI development set. It can also be observed from subplots (b) and (c) in Figure 3.8 that the estimated statistics generalize well to unseen meetings from different corpora such as NIST-RT 09 and ICSI.

3.6.3 Combination of silence and speaker change statistics

The probability estimates of single-speaker and overlapping speech classes obtained using silence and speaker change statistics are based on different conversational phenomena and we hypothesize that combining the information captured by these two features might result in a better estimate of the class probabilities. Motivated by this hypothesis, we explore various combination strategies proposed in the literature to obtain an estimate that exploits the information captured by both features. In particular, we experimented with two types of

combination methods, early combination and late combination.

In the early combination strategy, a combined feature vector is formed for each frame by appending the individual features, corresponding to the frame. Let d_i and c_i denote the duration of silence and number of speaker changes in a segment centered around frame i. Let b_i denote the combined feature vector formed by appending the features d_i and c_i corresponding to the frame i i.e., $b_i = [d_i, c_i]$. A logistic regression classifier is trained using these feature vectors on a training set of meetings with a binary target variable {0,1} denoting single-speaker and overlapping speech classes respectively. The target labels for each frame are obtained from the ground-truth segmentation. The output of this classifier is in the range [0,1] and can be treated as a probability estimate of the overlapping speech class i.e., $P(V = ov|b_i)$. Given this estimate, the probability of single speaker speech $P(V = sp|b_i)$ is obtained as $1 - P(V = ov|b_i)$.

In the late combination method, we explore different ways to combine the probability estimates $P(V = ov|d_i)$ and $P(V = ov|c_i)$ obtained from the individual features to get the final overlap probability estimate $P(V = ov|d_i, c_i)$. We experimented with standard combination methods such as the sum and product rules with uniform weights for individual estimators as well as the inverse-entropy based weighting scheme [Misra et al., 2003]. The combination according to sum rule can be written as,

$$P(V|d_i, c_i) = w_d P(V|d_i) + w_c P(V|c_i) \tag{3.11}$$

The combination according to product rule can be written as:

$$P(V|d_i, c_i) = \frac{1}{P(V)} P(V|d_i)^{w_d} P(V|c_i)^{w_c}. \tag{3.12}$$

where $P(V)$ denotes the prior probability of a class where $V \in \{ov, sp\}$. When using uniform weights, both w_d and w_c are set to 0.5. In the inverse entropy based weighting scheme, the weights are set as explained below. Let H_d^i and H_c^i respectively denote the entropies of probability estimators based on silence and speaker change statistics for a frame i. The weights of individual probability estimators for a frame i are computed as, $w_d = \frac{1/H_d^i}{1/H_d^i + 1/H_c^i}$ and $w_c = \frac{1/H_c^i}{1/H_d^i + 1/H_c^i}$.

To evaluate the usefulness of the different combination strategies, we use the cross-entropy measure computed on the AMI development set. The cross-entropy is computed based on (3.4) using the probability estimates of the classes obtained by the above mentioned combination strategies. Cross-entropy based studies on development set of meetings revealed that the inverse-entropy based weighting gave similar results to the uniform weighting of the individual estimators. Table 3.2 shows the cross-entropy measures obtained for different combination methods mentioned above. It can be observed from Table 3.2 that estimates obtained by the product rule of combination have the lowest cross-entropy. Therefore, based on these findings, in the current work, we use the product rule based combination method to combine

Table 3.2: Cross entropy measure on AMI development set of meetings for various combination strategies.

method	cross entropy
prod-uniform	0.60
sum-uniform	0.63
log. reg.	0.62

information captured by different conversational features.

3.6.4 Combination of conversational and acoustic features

The probability estimates of single-speaker and overlapping speech classes obtained from silence and/or speaker change statistics are integrated into the acoustic feature based classifier as prior probabilities of these classes. As mentioned earlier, the prior probabilities of the classes in the acoustic feature based system usually are fixed to a constant value based on the proportion of samples in each class observed during the training phase. However, studies have shown that the probability of overlap occurrence is neither constant across different conversations nor within a conversation. Therefore, to address this issue in the acoustic feature based classifier, we introduce prior probabilities that are estimated based on the silence and the speaker change statistics in context of a frame. These probabilities encode information present in the long-term context of a frame and change depending on the context. Let F denote the sequence of acoustic features and let C denote the sequence of conversational features which can be either the individual features or their combination. Given these, the most probable state sequence V^* where the states in the sequence belong to the set $\{ov, sp\}$ can be estimated by Viterbi decoding as:

$$\begin{aligned} V^* &= \arg\max_V P(V|F,C) \\ &= \arg\max_V P(F|V,C)P(V|C) \\ &\doteq \arg\max_V P(F|V)P(V|C) \end{aligned} \tag{3.13}$$

assuming that, given the state/class the observed acoustic features F are independent of the conversational features C. The quantity $P(F|V)$ is modelled using GMM distributions of the corresponding states and is only dependent on the current frame. The term $P(V|C)$ estimates the probability of the states based on the conversational features such as silence duration and/or number of speaker changes in a segment surrounding the current frame. Also, it captures the information present in the long-term context of the frame which is not present in the acoustic features. Therefore, we hypothesize that this combination will improve the performance of the classifier.

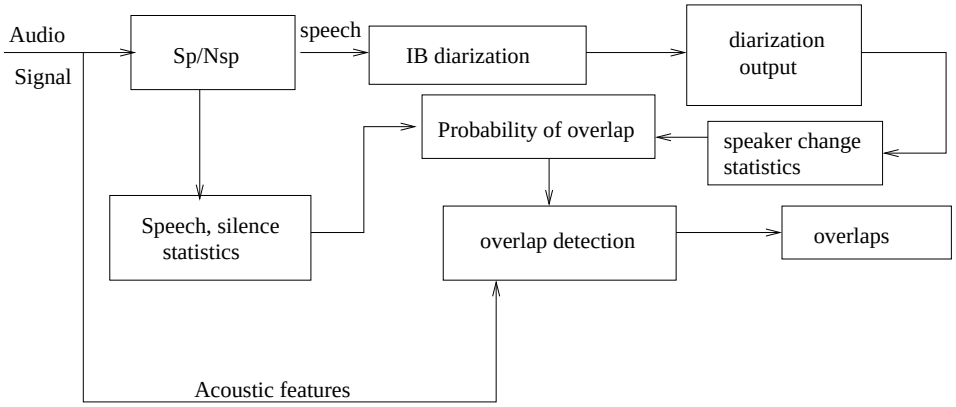


Figure 3.9: Block diagram of the proposed method of overlap detection: Speech, silence statistics and speaker change statistics are obtained as a by product of initial pass of IB speaker diarization. These statistics are used to estimate the probability of occurrence of overlap as explained in Sections 3.6.1, 3.6.2 and 3.6.3. These probabilities are incorporated into the acoustic feature based overlap detector (Section 3.5.2) as prior probabilities as detailed in Section 3.6.4

3.7 Experiments and Results

In this section, we present the experimental results of overlapping speech detection using standard acoustic features as explained in Section 3.5.2 and the proposed method of using probability estimates of the classes obtained from conversational features as prior probabilities of the states as explained in Section 3.6. We also present experiments evaluating the effect of the acoustic feature based overlap detection system and the proposed method for overlap detection on speaker diarization. Figure 3.9 summarizes the proposed method for overlap detection. The obtained overlaps are used for overlap handling in speaker diarization as explained in Section 3.5.1.

3.7.1 Experiments on overlap detection

We evaluate the performance of the acoustic feature based overlap detector and the proposed method of incorporating conversational information into acoustic feature based detector on three different meeting corpora namely, AMI, NIST-RT and ICSI. A HMM/GMM based overlap detector is trained for each corpus using training set of meetings for the respective corpus as described in Section 3.5.2. The parameters needed for estimating the probabilities of single-speaker and overlapping speech classes from the conversational features as described in Sections 3.6.1, 3.6.2 and 3.6.3 are obtained from the AMI training set of meetings. These parameters are then used to estimate the probabilities of the classes on the test set of meetings in all the corpora and are incorporated into the acoustic feature based HMM/GMM system as explained in Section 3.6.4.

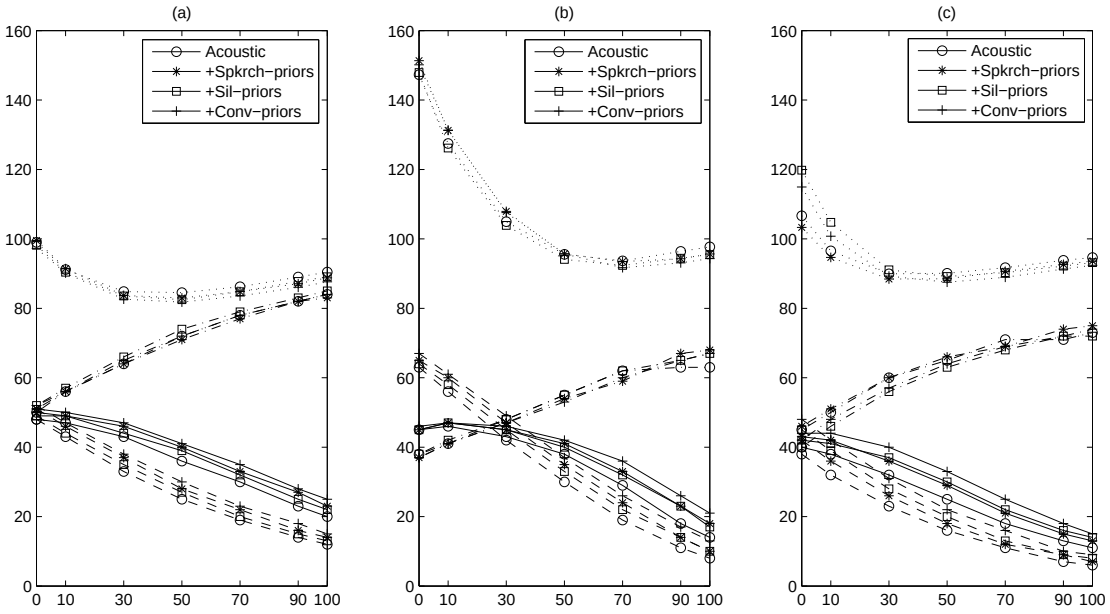
In the first experiment, we verify the hypothesis stated in the Section 3.6.4 that, incorporat-

ing the probabilities of overlapping and single-speaker speech classes estimated from the proposed conversational features such as silence and speaker change statistics as prior probabilities of the classes in the acoustic feature based classifier improves its performance. The systems are compared using the metrics such as recall, precision, F-measure and detection error. The recall of a system on overlap detection task is computed as the ratio between the duration of overlap that is correctly detected by the system and the total duration of actual overlap. The precision is computed as the ratio between duration of the overlap that is correctly detected by the system and total duration of overlapping speech detected by the system. The F-measure is the harmonic mean between recall and precision. It is a good indicator of classifier performance when the number of samples in classes is skewed as in the current study (proportion of speech in overlap regions is significantly less than that of in single speaker regions). The overlap detection error rate is computed as the ratio between the total duration of missed and false overlap detections and the total duration of overlap.

Figure 3.10 shows the performance of the acoustic feature based overlap detector and the proposed method on the test sets of three meeting corpora in terms of recall (dashed line), precision (-.- line), F-measure (solid line) and error (dotted line) of the respective classifiers on the task of overlap detection. In Figure 3.10, the acoustic feature based classifier (Section 3.5.2) is denoted by the label Acoustic, the system using probability estimates from silence statistics (Section 3.6.1) is denoted as +Sil-priors, the system using probability estimates from speaker change statistics (Section 3.6.2) is denoted as +Spkrch-priors and, the system using the probability estimates from the combination of the individual features by product rule (Section 3.6.3) is denoted as +Conv-priors. The figure plots various evaluation metrics for different values of overlap insertion penalty (OIP) which is introduced to have a trade-off between the true and the false overlap detections. In general, it can be observed that higher values of OIP tend to increase the precision of the classifier while sacrificing the recall.

The error rates achieved by the acoustic feature based overlap detector on the AMI-test set are similar to the error rates obtained by prior works [Boakye et al., 2008, Zelenák et al., 2012, Geiger et al., 2012] in literature on the data set. It can be observed from the Figure 3.10 (a) that incorporating the probabilities of the classes estimated from the individual conversational features (Sil-priors, Spkrch-priors) improves the performance of the acoustic feature based system as they consistently achieve higher f-measure and lower error over the acoustic feature based classifier for all the values of OIP. Also, the combination (Conv-priors) of the conversational features (Sil-priors, Spkrch-priors) leads to further improvements in overlap detection. In the experiments reported here, the combination is performed based on the product rule as described in Section 3.6.3, since the combination based on product rule obtained the lowest cross-entropy on AMI development set (see Tab. 3.2). The improvements in terms of f-measure and decrease in the error rate achieved by the proposed method is mainly due to increase in the recall of the classifier when compared to the acoustic feature based system. This indicates that the proposed method is able to identify instances of overlaps which are not detected by the acoustic feature based system. Similar trends can be observed in all the corpora though the absolute values of the evaluation metrics are different. In general, it can be observed that

Figure 3.10: Overlap detection evaluation on (a) AMI-test set (b) NIST-RT 09 set and (c) ICSI-test set. In each subplot (a), (b), (c), dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) in x-axis.

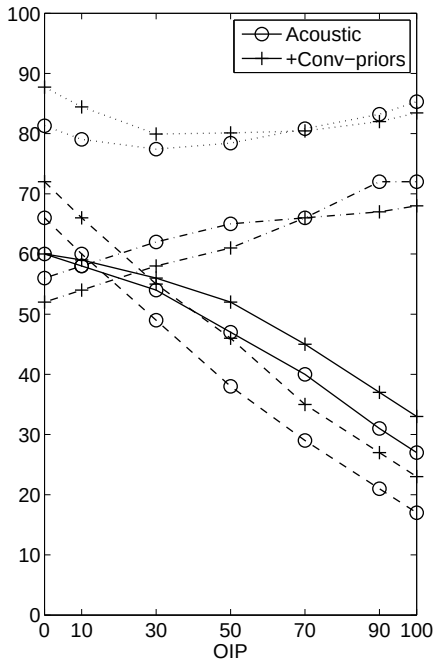


incorporating conversational features into the acoustic feature based classifier improves the performance of the classifier as shown by the consistent higher f-measure and the lower error rates achieved by the proposed method on all the three corpora when compared to that of the acoustic feature based system. This result is particularly encouraging as it demonstrates that, model trained to estimate probability of overlap based on conversational features using one corpus (AMI), generalizes well to meetings from the other corpora (NIST-RT and ICSI).

Laughter overlap detection

Laughter is a very common phenomenon in human interactions [Glenn, 2003]. Studies on spontaneous conversations have shown that overlaps and laughter occurrences are correlated with each other [Laskowski and Burger, 2007, Kennedy and Ellis, 2004]. Studies done on ICSI corpus have shown that 9% of speaking time contains laughter [Laskowski and Burger, 2007]. Based on these studies, we evaluate the performance of the proposed method for overlap detection in laughter segments of ICSI corpus. The start and end times of laughter segments and the corresponding speakers are obtained from the annotations done for analysis of laughter in [Laskowski and Burger, 2007]. Figure 3.11 presents results of overlap detection on laughter segments based on acoustic features alone(Acoustic) and combination of acoustic and conversational features (+Conv-priors). To obtain the conversational features silence and speaker change statistics which are used to estimate the prior probabilities of the classes, we superimpose the ground-truth laughter segments of a meeting recording over speaker

Figure 3.11: Overlap detection evaluation on laughter segments from ICSI-test set: dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection in percentage on y-axis for various overlap insertion penalties (OIP) on x-axis.

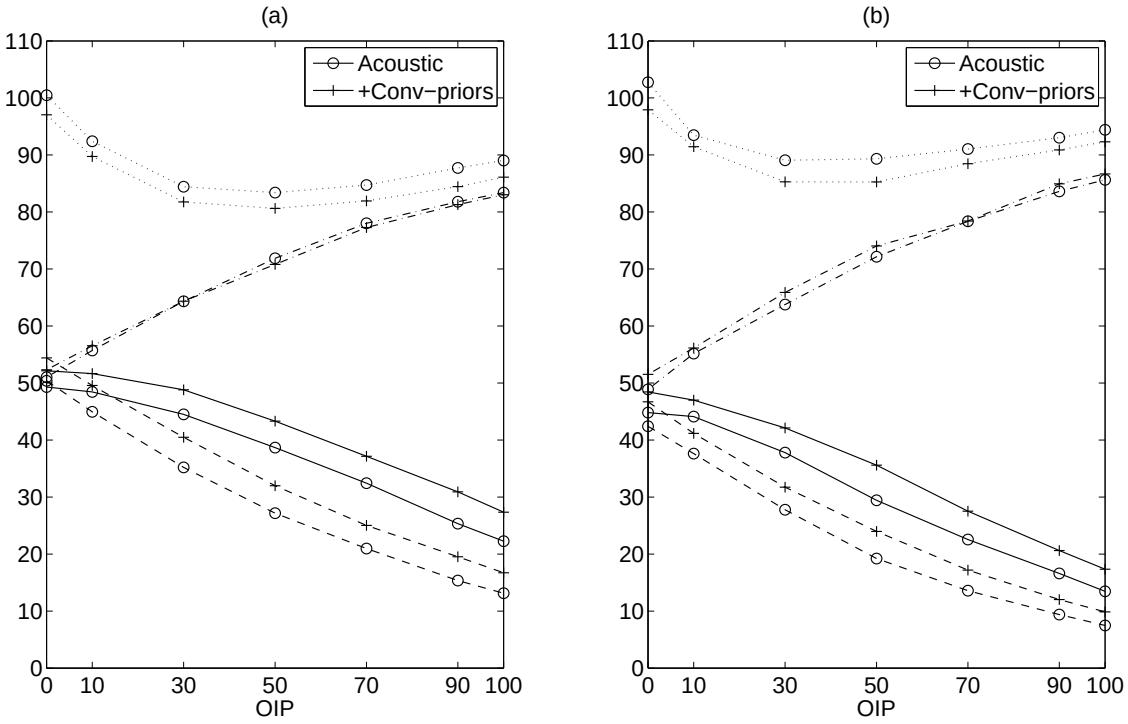


diarization output of the respective meeting. It can be observed from the Figure 3.11 that the combination of conversational features improves the performance of the acoustic feature based overlap detector as shown by the f-measures for various OIPs.

Robustness to speaker diarization errors

The conversational features proposed in the current work make use of the baseline IB diarization output to compute speaker change statistics. These statistics computed over a long-term context of a frame are used to estimate the probability of overlap at that frame. To evaluate the effect of errors made by the clustering algorithm of speaker diarization on overlap detection, we compare the overlap detection performance on meetings with high and low diarization error. For this purpose, we divided the AMI-test set into two subsets, based on the speaker error of the diarization output. All the meetings with a speaker error less than 15% were put in the low error set and rest of the meetings were put in high error set. Figure 3.12 plots the evaluation metrics for overlap detection for acoustic feature based system (Acoustic) and the combination of acoustic and conversational feature based system (+Conv-priors) on high and low error sets. From Figure 3.12, it can be observed that the performance of the acoustic feature based detector is improved on both the sets by the combination of conversational features. The low error set (Figure 3.12(b)) has slightly high precision when compared to the

Figure 3.12: Effect of diarization errors: (a) High error set (b) Low error set. Dotted line shows the error, solid line shows f-measure, dashed line shows recall, and '-.-' shows precision of the classifiers in overlap detection on in percentage y-axis for various overlap insertion penalties (OIP) on x-axis.



high error set (Figure 3.12(a)).

3.7.2 Speaker diarization with overlap handling

In this section, we evaluate the effect of overlap detection on speaker diarization on three different corpora AMI-test set, NIST RT-09 and ICSI-test sets. For experiments on the ICSI-test set, we included laughter segments in speech regions given as input to the diarization system to make the scenario as natural as possible. We used ground-truth speech/non-speech segmentation for experiments on ICSI corpus to avoid missing laughter segments because of automatic speech activity detection, which might classifying laughter segments as non-speech. The detected overlaps are used in speaker diarization by performing overlap exclusion and labelling techniques explained in Section 3.5.1. Using these methods, we compare the two overlap detection systems, one based on just the acoustic features and the other incorporating the information from conversational features as prior probabilities of the classes. Since the combination of the conversational features showed the best performance in overlap detection (Figure 3.10), we use the overlap detection hypothesis generated by the system using the combination of the conversational features based on the product rule (Conv-priors). To decide the optimal value of OIP to use in the overlap detection system, we perform tuning experiments

on ami AMI development corpus and pick an OIP that gives lowest DER. As proposed in earlier works [Zelenák et al., 2010, Boakye et al., 2011], we perform overlap exclusion and overlap labelling steps using different overlap detection hypothesis dependent on the value of OIP used. This is based on the rationale that, high precision overlap detection hypothesis is desirable for overlap labelling step to avoid increasing false alarm errors. Also, high recall hypothesis is desirable for overlap exclusion, as it helps in avoiding as much overlapping speech as possible from corrupting the speaker (cluster) models.

Overlap exclusion

To tune the overlap insertion penalty (OIP) for overlap exclusion, we ran experiments on a development set of meetings by performing overlap exclusion using the overlap hypothesis generated by various values of OIP. These experiments have revealed a similar trend to that observed in previous studies [Zelenák et al., 2012], where the DER reduction was not a smooth function of OIP. Therefore, for overlap exclusion we use the detection hypothesis obtained with OIP set to zero as done in previous studies [Zelenák et al., 2012, Geiger et al., 2012]. Tab. 3.3 reports DER and its components speech/non-speech error (SpNsp) and speaker error (Spkr) obtained on a test set of meetings from the AMI and NIST-RT 09 corpus using the baseline speaker diarization system in three scenarios; without any overlap exclusion (Baseline), overlap exclusion using the hypothesis generated by acoustic feature based system (Acoustic) and overlap exclusion using the hypothesis generated by the proposed method (Conv-priors). First of all, it can be observed from the Tab. 3.3 that performing overlap exclusion reduces the speaker error (Spkr) of the diarization as expected, since it avoids corruption of resulting speaker models. Also, it can be observed from Tab. 3.3 that on AMI-test set, the proposed method reduces the DER by around 17% relative to the baseline speaker diarization system that does not do any overlap exclusion. The acoustic feature based overlap detector reduces the DER by around 13% relative to the baseline. The table also reports the f-measures of both the overlap detection systems used to perform overlap exclusion. This reveals that higher reduction in DER achieved by the proposed method is due to the ability of the proposed method to detect more overlap at the given OIP (0).

Tab. 3.3 also presents results of similar experiments on the NIST-RT 09 and ICSI-test data sets. On the RT-09 data set the acoustic feature based overlap detector reduces the DER by 3.5% relative and adding conversational features further reduces the DER by 6.4% relative. On the ICSI-test set, the acoustic feature based overlap detector reduces the DER by 2.4% relative and the DER reduces by 4.2% relative when conversational features are added. The drop in the relative reductions in DER when compared to the AMI data set (Tab. 3.3) is mainly due to the performance drop in the acoustic feature based classifier as indicated by the f-measures of the classifiers on the corpora in Tab. 3.3. Nevertheless, the proposed method (Conv-priors) achieves a lower DER than the system using overlaps from acoustic feature based system (Acoustic) on RT 09 and ICSI-test data sets also.

Table 3.3: *Overlap exclusion on the AMI-test, NIST-RT 09 and ICSI-test data sets: SpNsp represents error in speech/non-speech detection, Spkr denotes speaker error (clustering error), DER is the total diarization error rate and f-measure of the respective overlap detection systems at the operating point (OIP=0) used for exclusion.*

Corpus	System	SpNsp	Spkr	DER	f-measure
AMI	Baseline	13.5	16.9	30.4	-
	Acoustic	13.5	12.8	26.3	0.48
	+Conv-priors	13.5	11.6	25.1	0.51
RT 09	Baseline	12.7	21.2	33.9	-
	Acoustic	12.7	20.0	32.7	0.44
	+Conv-priors	12.7	19.0	31.7	0.46
ICSI	Baseline	17.7	15.6	33.3	-
	Acoustic	17.7	14.8	32.5	0.44
	+Conv-priors	17.7	14.2	31.9	0.49

Overlap labelling

To determine the optimal value of OIP for the labelling task, we ran tuning experiments on the AMI development set using overlap hypothesis obtained for different values for OIP. Figure 3.13 plots the relative change in DER due to overlap labelling as a function of OIP used to generate the overlap hypothesis. In the present work, we use nearest neighbor based labelling as it gave similar results to the cluster likelihood based labelling. It can be observed from Figure 3.13 that the OIP value of 90 gives the highest relative decrement in DER on the AMI development set. Based on this observation, we use the overlap hypothesis obtained by setting the value of OIP to 90 while performing overlap labelling. Tab. 3.4 presents the overlap labelling results on the test set of meetings in the AMI and RT corpora. It can be observed from Tab. 3.4 that on the AMI corpus, the proposed method (Conv-priors) decreases the DER by 2.9% relative to the baseline diarization system that does not use any overlap information (Baseline). It also achieves lower DER than the system using overlaps detected by acoustic feature based system (Acoustic). The decrease in DER achieved by the proposed method is due to the decrease in speech/non-speech error (SpNsp). The speech/non-speech error (SpNsp) is reduced from 13.5% in the baseline diarization system to 11.9% in the proposed method, which is around 12% relative reduction. This reduction is due to the detection of overlapping speech and labelling it as speech. But the improvement in the final DER is not in the same range due to the errors introduced during labelling, which increase the speaker error when the identified overlap segments are not assigned to correct speakers. This highlights the need for a novel speaker labelling method for the detected overlaps. Similar trends can be observed in Tab. 3.4 on the RT 09 data set also. The lower DERs achieved by the proposed method on the two data sets can be attributed to better detection of overlaps as indicated by the higher values of f-measure and lower error rate obtained by it on the data sets at the given OIP of 90 when compared to the acoustic feature based system (see Figure 3.10). On the ICSI corpus also, overlaps obtained by the combination of acoustic and conversational features achieve the

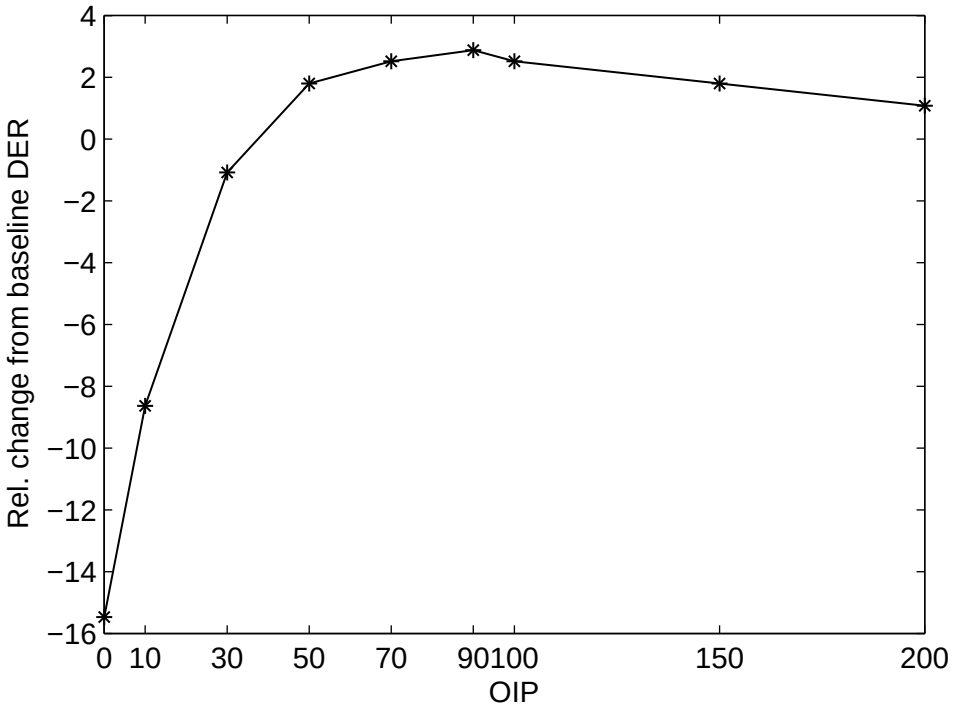


Figure 3.13: Overlap labelling: Tuning for optimal overlap insertion penalty (OIP) on AMI development set.

highest reduction of DER by labelling. The relative reduction on the ICSI-test set is better than the RT-09 even though the overlap detection performance is similar. This is due to the fact that the ICSI-test set contains overlaps from laughter segments which tend to have more speakers than normal speech overlaps. Therefore, the labelling method is less prone to errors when labelling overlaps in laughter segments.

Overlap exclusion followed by labelling

In this section we summarize the results of the diarization experiments when overlap exclusion and labelling are performed alone and together, where exclusion is followed by labelling as described in Figure 3.2. Tab. 3.5 summarizes the DERs obtained on all the three data sets AMI-test and NIST-RT 09 and ICSI-test while performing overlap handling techniques Exclusion, Labelling, and Both (exclusion followed by labelling) using overlaps detected by the acoustic feature based system (Acoustic) and the proposed method of combining acoustic and conversational features(+Conv-priors). It can be observed from Tab. 3.5 that the proposed method achieves highest reduction in DER on all the corpora. It increases the relative error reduction due to overlap detection to around 20% on meetings from AMI-test from around 15% achieved by the acoustic feature based overlap detection system. On RT 09 meetings, it increases the relative DER reduction to around 7% from 3.5% achieved by the system based on acoustic features and on the ICSI-test set it increases the relative error reduction to around 7%

Table 3.4: *Overlap labelling on AMI-test and NIST-RT 09 and ICSI-test data sets: missed (Miss) and false-alarm (FA) errors in speech/non-speech detection and total speech/non-speech error (SpNsp), speaker error (Spkr), diarization error rate (DER) and f-measures of the respective overlap detection systems at the operating point (OIP=90) chosen to do labelling.*

Corpus	System	Miss	FA	SpNsp	Spkr	DER	f-measure
AMI	Baseline	12.5	1.0	13.5	16.9	30.4	-
	Acoustic	10.9	1.4	12.3	17.4	29.7	0.23
	+Conv-priors	10.2	1.7	11.9	17.6	29.5	0.28
RT 09	Baseline	11.6	1.1	12.7	21.2	33.9	-
	Acoustic	11.0	1.5	12.5	21.4	33.9	0.18
	+Conv-priors	10.6	1.6	12.2	21.5	33.7	0.26
ICSI	Baseline	17.7	0	17.7	15.6	33.3	-
	Acoustic	16.6	0.1	16.7	15.9	32.6	0.18
	+Conv-priors	15.8	0.4	16.2	16.1	32.3	0.26

Table 3.5: *DERs (with relative improvements over baseline diarization within parenthesis) obtained while performing overlap handling by using overlaps detected by proposed method (+Conv-priors) and acoustic feature based overlap detector (Acoustic).*

Corpus	System	Exclusion	Labelling	Both
AMI	Acoustic	26.3 (+13.4%)	29.7 (+2.3%)	25.8 (+15.1%)
	+Conv-priors	25.0 (+17.7%)	29.5 (+2.9%)	24.2 (+20.3%)
RT 09	Acoustic	32.7 (+3.5%)	33.9 (-)	32.7 (+3.5%)
	+Conv-priors	31.7 (+6.4%)	33.7 (+0.6%)	31.5 (+6.7%)
ICSI	Acoustic	32.5 (+2.4%)	32.6 (+2.1%)	31.8(+4.5%)
	+Conv-priors	31.9 (+4.2%)	32.3 (+3%)	30.9 (+7.2%)

from 4% achieved by the acoustic feature based system.

3.8 Summary

Motivated by the studies done on conversational analysis, in the current work, a method to improve the acoustic feature based overlap detector using long-term conversational features was proposed. Experiments done in the present work revealed that features extracted automatically from conversations such as silence and speaker change statistics carry relevant information about overlap occurrence in a conversation. The features were computed over a window of around 3-4 secs to capture the information present in the long-term context of a frame. These features were then used to estimate the probability of overlap and single-speaker speech classes in the window. Cross entropy measure based studies on development data revealed that the probability estimates of the classes are closer to the true distribution as the length of the context used to compute the features is increased and reaches an optimum around 4 secs. The probability estimates of the overlapping and single-speaker classes obtained from the long-term context of each frame in the conversation were incorporated into

the acoustic feature based classifier as prior probabilities of the classes.

Experimental results on overlap detection using three different meeting corpora (AMI, NIST-RT, ICSI) revealed that the proposed method improves the performance of the acoustic feature based classifier. These experiments also revealed that the model learnt to estimate the class probabilities using data from the AMI corpus is generalizable to other meeting corpora such as NIST-RT and ICSI. Experiments were also done to evaluate the effect of overlap detection on speaker diarization using the standard methods of exclusion and labelling. These experiments revealed that the proposed method decreases the DER by 20% relative to the baseline speaker diarization system on the AMI-test data set. Using overlap detection from the acoustic feature based system reduced the DER by 15% relative. Speaker diarization experiments on NIST-RT 09 and ICSI-test data sets have also revealed that the proposed method achieves higher reduction in DER when compared to the system using only acoustic features for overlap detection. These experiments also highlight the need for an effective overlap labelling mechanism to assign speakers to the detected overlap segments as the reductions obtained in speech/non-speech error are compensated to an extent by the increase in speaker error due to the errors done during labelling.

4 Auxiliary information for speaker diarization

4.1 Introduction

Speaker diarization on far-field audio is adversely effected by various factors such as background noise and channel characteristics. Several diagnostical studies on speaker diarization have shown that after overlapping speech one of the main causes for degradation of diarization output is errors in speech/non-speech detection [Huijbregts and Wooters, 2007, Huijbregts et al., 2012]. Non-speech segments in meeting rooms contain different types of signals such as door slamming, paper shuffling or other types of ambient noises or vocal sounds like laughter or cough which have very different characteristics than a usual speech signal. When such signals are combined with speech segments they degrade the purity of speaker clusters and increase the speaker error. In [Anguera et al., 2006b], a method was proposed to improve the diarization output by detecting and eliminating non-speech frames that are included in clustering due to errors in speech/non-speech detection. In that work, it was observed that non-speech frames included in clustering have a negative influence on the cluster merging decisions. In such scenarios, extra information sources which are helpful in speaker discrimination have to be explored. The current chapter proposes two such auxiliary information sources which are non-speech regions and phoneme transcripts of a given recording to help improve speaker diarization in information bottleneck framework.

To incorporate information from non-speech regions in a meaningful way, we propose the information bottleneck with side information (IBSI) framework for speaker diarization. We explain the clustering based on IBSI framework and use the agglomerative solution of the IBSI clustering framework to perform speaker diarization. We also show that having information about what is being spoken (phoneme transcripts) can be useful in identifying who is speaking. The information from phoneme transcripts is used to estimate a phoneme background model (PBM) which is then used for speaker diarization in information bottleneck framework. These two approaches are explained in detail and evaluated on meetings from different corpora.

4.2 Non-speech as side information for speaker diarization

The use of non-speech regions as side information (irrelevant variable) for clustering is motivated by two reasons. In a typical diarization system, automatic speech/non-speech detection is performed before initializing the agglomerative clustering. Errors in speech/non-speech detection are a common artefact, which introduces some non-speech frames into clustering. Since a typical agglomerative speaker diarization system is initialized by uniform segmentation, segments of different speakers containing similar non-speech (background noise) might get merged due to their similarity in the non-speech regions rather than speech regions. The second reason is that, segments with low signal to noise ratio (SNR), belonging to different speakers, but corrupted by similar background noise, might get merged into a cluster due to the similar noise characteristics rather than their speech characteristics. In the present work, we hypothesize that using information from the non-speech segments in clustering can overcome these issues. To achieve this, we use information from non-speech regions as irrelevant variable for clustering in information bottleneck with side information (IBSI) framework. The IBSI framework [Chechik and Tishby, 2003] aims to cluster the input segments such that the resulting clusters maximize the mutual information with respect to relevant variables to clustering while minimizing the mutual information with respect to irrelevant variables. This method penalizes the cluster similarity in the relevant variable space with the similiarity in the irrelevant variable space, Since the method only uses the non-speech segments from a given recording, it does not require any pre-labelled non-speech data. It also has the advantage of using the data that best represents the given recording.

4.2.1 Agglomerative Information Bottleneck with Side Information (IBSI)

The goal of any clustering algorithm is to find meaningful structure in the given data and construct clusters based on this structure. But often there are multiple conflicting structures in the data; for example documents can be clustered based on topic or writing style; speech segments can be clustered based on speaker identity or background/environmental conditions etc. All the above clustering solutions are valid alternatives and the desired solution depends on the problem formulation. The IBSI framework was proposed to identify relevant patterns among several conflicting ones that might exist in the data [Chechik and Tishby, 2003]. The method has been successfully applied in document clustering, processing neural spike train activity, and face recognition [Chechik and Tishby, 2003, Chechik, 2003]. The method incorporates information about irrelevant components of the data to better extract the relevant pattern information. Given a set of input variables X that need to be clustered, a set of relevant variables Y^+ whose characteristics should be preserved in the final clustering, a set of irrelevant variables Y^-, and the joint distributions $P(X,Y^+)$ and $P(X,Y^-)$, the IBSI framework aims at clustering the input variable set X into clusters C such that the resulting clusters maximize mutual information w.r.t the relevant variable set Y^+ and minimize mutual information w.r.t the irrelevant variable set Y^-. This can be represented as a maximization of

the objective function below:

$$F_{IBSI} = I(Y^+, C) - \gamma I(Y^-, C) - \frac{1}{\beta} I(X, C) \tag{4.1}$$

where, γ and β are Lagrange multipliers.

F_{IBSI} can be optimized using various approaches [Chechik, 2003] such as deterministic annealing, greedy agglomerative hard clustering and sequential K-means based clustering. To be compatible with the already existing diarization framework [Vijayasenan et al., 2009], in the current work, we used the agglomerative hard clustering solution to the optimization problem. In this method, loss in the objective function due to a merge of two clusters c_i and c_j can be obtained as:

$$\nabla F_{IBSI}(c_i, c_j) = [p(c_i) + p(c_j)] d_{ij}^{IBSI} \tag{4.2}$$

The distance d_{ij}^{IBSI} between two clusters c_i and c_j can be obtained as:

$$JS[p(Y^+|c_i), p(Y^+|c_j)] - \gamma JS[p(Y^-|c_i), p(Y^-|c_j)] - \frac{1}{\beta} JS[p(X|c_i), p(X|c_j)] \tag{4.3}$$

where $JS()$ denotes the Jensen-Shannon divergence between two distributions. At each step of agglomerative clustering, the algorithm merges the two clusters that result in the lowest value of ∇F_{IBSI}. By comparing the two distance measure of the original IB clustering and d_{ij}^{IBSI}, it can be observed that d_{ij}^{IBSI} incorporates an extra penalty term $\gamma JS[p(Y^-|c_i), p(Y^-|c_j)]$ which measures the similarity between two clusters in irrelevant variable domain Y^-. Due to this, the current distance measure penalizes the merge of clusters with similar distribution over irrelevant variables. The whole method is summarized in Figure 4.1. The model selection criterion which gives the number of final clusters is based on a threshold on the normalized mutual information given by $\frac{I(C,Y^+)}{I(X,Y^+)}$.

4.2.2 Agglomerative IBSI for speaker diarization

To apply this method to speaker diarization, the set of relevant variables Y^+ is defined as the components of the background GMM trained on the speech regions of a given recording similar to aIB framework. The set of irrelevant variables Y^- is defined as the components of the background GMM trained on non-speech regions of a given recording. Similar to the agglomerative IB (aIB), the clustering starts with uniformly segmented speech regions represented by X, and the posterior distributions of relevant and irrelevant variables $p(Y^+|X)$,

$p(Y^{-}|X)$, are obtained using Bayes' rule. Clusters that have the lowest distance measure (4.2) are merged at each step. The final number of clusters is obtained using the model selection criterion.

4.3 Phoneme background model for speaker diarization

The knowledge of what is being spoken has been shown to be a very useful information in speaker identification/verification tasks [Sturim et al., 2002, Stolcke et al., 2007]. This gives a chance to model individual variations in pronunciation of an acoustic class (phoneme/-word) [Weber et al., 2000, Eatock and Mason, 1994]. Due to this, text constrained speaker identification/verification tasks usually have higher accuracies than their text independent counterparts. In the current work, we perform experiments to investigate whether knowledge of what is being spoken helps to improve speaker diarization.

4.3.1 Diarization experiments using oracle annotations

First of all, we perform an oracle speaker diarization experiment, with ground-truth phone transcription and speaker segmentation. In this experiment, we compare the IB diarization systems using different background models to perform speaker diarization. We compare plain background GMM (Plain-UBM) estimated from the speech regions of a given meeting recording, the background model estimated with the knowledge of speaker segmentation (Spkr-UBM) and the model estimated with the knowledge of both speaker and phoneme being spoken (Spkr-phone-UBM). In the Spkr-phone-UBM, each phoneme spoken by a speaker is represented by a Gaussian component in the background GMM. This is done by accumulating all the utterances of a phoneme by a speaker according to the ground-truth transcripts and approximating them with a Gaussian. In total there are 45 phonemes including silence class. To estimate the Spkr-UBM, speech segments belonging to a speaker are used to estimate n components per speaker in the background GMM, where n is varied from 5 to 45. The ground-truth phone transcripts are obtained by force-aligning the manual transcripts to individual head microphone channels, which also produces speaker start and end times as a by product. Once the background model is obtained based on one of the approaches explained above (Plain-UBM/Spkr-UBM/Spkr-phone-UBM), IB speaker diarization is performed using the components of the background model as a relevance variable set. We used 100 meetings from the AMI corpus [Mccowan et al., 2005] in this experiment. Figure 4.2 plots the speaker error in IB diarization when using the three background models (Plain-UBM/Spkr-UBM/Spkr-phone-UBM). The lowest error for Spkr-UBM is at $n = 20$ which means 20 is the optimal number of components per speaker in the background GMM. It can be observed that Spkr-phone-UBM achieves the lowest error among all the background models which shows that having the knowledge of what is being spoken helps speaker diarization.

Input:

Joint Distribution $p(x, y^+)$, $p(x, y^-)$

Trade-off parameters γ, β

Output:

C_m: m-partition of X, $1 \leq m \leq |X|$

Initialization:

$C \equiv X$

For $i = 1 \ldots N$

$c_i = \{x_i\}$

$p(c_i) = p(x_i)$

$p(y^+|c_i) = p(y^+|x_i) \forall y^+ \in Y^+$

$p(y^-|c_i) = p(y^-|x_i) \forall y^- \in Y^-$

$p(c_i|x_j) = 1$ if $j = i, 0$ otherwise

For $i, j = 1 \ldots N, i < j$

Find $\nabla F_{IBSI}(c_i, c_j)$

Main Loop:

While $|C| > 1$

$\{i, j\} = \arg\min_{i',j'} \nabla F_{IBSI}(c_{i'}, c_{j'})$

Merge $\{c_i, c_j\} \Rightarrow c_r$ in C

$p(c_r) = p(c_i) + p(c_j)$

$p(y^+|c_r) = \frac{[p(y^+|c_i)p(c_i) + p(y^+|c_j)p(c_j)]}{p(c_r)}$

$p(y^-|c_r) = \frac{[p(y^-|c_i)p(c_i) + p(y^-|c_j)p(c_j)]}{p(c_r)}$

$p(c_r|x) = 1, \forall x \in c_i, c_j$

Calculate $\nabla F_{IBSI}(c_r, c)$, $\forall c \in C$

Figure 4.1: Agglomerative IBSI algorithm

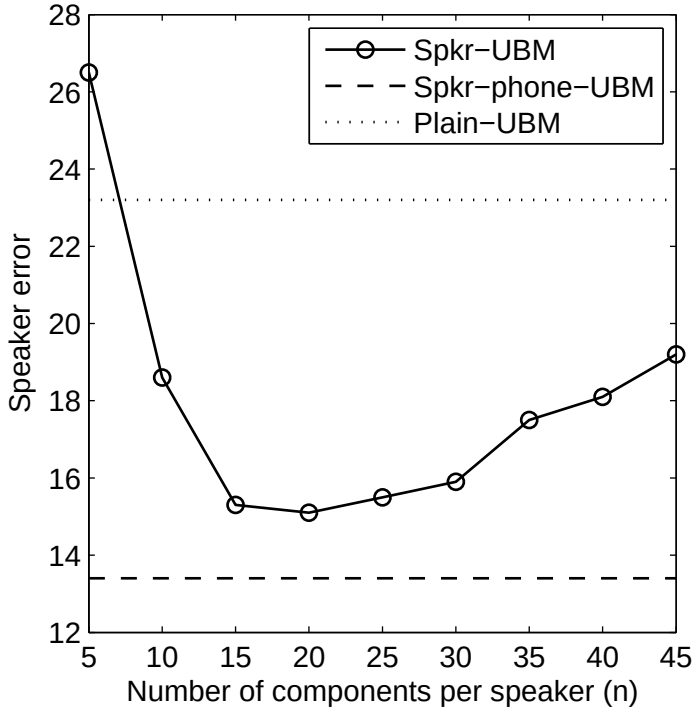


Figure 4.2: *Comparison of oracle UBMs: Spkr-UBM is estimated with the knowledge of speaker labels, Spkr-phone-UBM is estimated with the knowledge of both the speaker and phoneme being spoken and Plain-UBM is the regular UBM.*

4.3.2 Estimation of a phoneme background model (PBM)

Motivated by the above oracle experiment, we propose a method to estimate a background model that captures different modes of articulation of a phoneme. The hypothesis is that different modes of articulation of a phoneme arise due to differences in the individual speakers pronouncing it. Given a phone transcript of a meeting recording (obtained either from ground-truth or an ASR system), we employ a simple clustering algorithm such as Ward's method to estimate the clusters in a phoneme and use the obtained clusters from all the phonemes to build a background GMM which is referred to as the phoneme background model (PBM). Let $S = \{p_1, p_2, \ldots, p_i, \ldots, p_N\}$ denote the set of phonemes in the given transcript, where N is the number of unique phonemes in the transcription. Let $P_i = \{p_i^1, p_i^2, \ldots, p_i^{n_i}\}$ denote the set of all the occurrences of a phoneme p_i in the transcript, where n_i indicates the number of occurrences. Each occurrence p_i^j of phoneme p_i is approximated by a mean vector x_i^j computed from the feature vectors corresponding to that phoneme occurrence. This results in a set of points $X_i = \{x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^{n_i}\}$ where, x_i^j represents the j^{th} occurrence of the phoneme p_i in the transcript. Agglomerative clustering is performed using Ward's method to cluster the occurrences of a given phoneme. The agglomerative clustering is initialized by a set of single-ton clusters represented by X_i. Ward's method is a greedy clustering method, where, at each step, it merges two clusters that results in minimum increase of total cluster

variance. The distance measure $\Delta(c_k, c_l)$ between two clusters c_k and c_l, is given by:

$$\Delta(c_k, c_l) = \frac{n_{c_k} n_{c_l}}{n_{c_k} + n_{c_l}} \|m_k - m_l\|^2 \tag{4.4}$$

where, $\|\|$ denotes Euclidean distance, n_{c_k} and n_{c_l} represent the number of samples in cluster c_k and c_l respectively and m_k, m_l represent the mean vectors (centroids) of clusters c_k and c_l respectively. The clustering continues until the desired number of clusters M is obtained. The clustering is performed for all phonemes p_i in the set S and the final set of clusters $C = \{c_i^j\}$ $\forall\ i \in \{1,\ldots,N\}$ and $\forall\, j \in \{1,\ldots,M\}$ is obtained. Each of the clusters in the final cluster set C is represented as a Gaussian component in the PBM. The mean of component is equal to the centroid of the respective cluster and variance is equal to the cluster variance. The weight of a component is obtained as the proportion of samples assigned to the respective cluster. The number of clusters for each phoneme class M is selected based on cross-validation on development set. Figure 4.3 summarizes the procedure of estimating the PBM with the help of a block diagram.

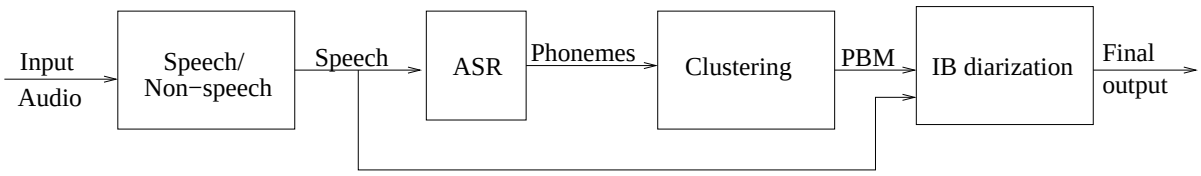


Figure 4.3: *Block diagram of the proposed diarization method using PBM.*

4.4 Experiments and Results

4.4.1 Diarization experiments using phoneme background model

Speaker diarization experiments are conducted on meetings from the AMI [Mccowan et al., 2005] and NIST-RT [National Institute of Standards and Technology, 2003] meeting corpora. Out of 170 meetings present in the AMI corpus, 100 meetings are used in the current experiments. The number of speakers in each meeting in the AMI dataset varies between 3 to 5, but most of the meetings have 4 speakers. For experiments on the NIST-RT corpus, we have used meetings from RT-05,06,07,09 datasets. The number of speakers in each meeting of the NIST-RT corpus varies between 4 to 11. Both the corpora contain meetings recorded in multiple meeting room environments. The audio captured by the distant microphone array is enhanced by beamforming using *BeamformIt* [Anguera, X,] toolkit. 19 Mel-frequency cepstral coefficients (MFCC) are extracted for each frame of length 30 ms with a frame shift of 10 ms from this enhanced signal. These features are used both for speaker diarization and for clustering of data within each phoneme class to obtain a PBM. Prior to performing diarization, speech/non-speech detection is performed using the SHOUT toolkit [Huijbregts and de Jong, 2011] and non-speech segments are ignored. The speaker diarization systems are evaluated using the metric diarization error rate (DER) used in NIST evaluation campaigns.

The ASR system used in the current study is a conventional HMM/GMM system [Motlicek et al., 2013]. The acoustic models are trained on 150 hours of labelled speech from the AMI and ICSI [Janin et al., 2003] corpora. 13 dimensional MFCC along with their first and second order derivatives resulting in a 39 dimensional feature vectors are extracted every 10 ms from a speech frame of length 30 ms. The features are extracted from individual head microphone (IHM) channels worn by the speakers in the meetings. These features are used for training the models and for decoding. The 1-best word recognition output is subsequently transformed into a sequence of phonemes (i.e., top-down approach to deriving a phoneme sequence) which is later used for PBM estimation.

The optimal number of clusters M for each phoneme class in the PBM is decided based on diarization experiments on development data. Since the number of clusters is inherently dependent on the number of speakers in a given recording and since the number of speakers varies significantly between AMI and NIST-RT meetings, we performed separate development experiments for each corpus. For the development experiments on the AMI corpus, we used 20 meetings that are not included in the 100 test meetings. For NIST-RT meetings, we used meetings from RT-05,06 as the development set and used RT-07,09 as our test set. Development experiments on AMI data set revealed that $M = 6$ is optimal for meetings from the AMI corpus. For the NIST-RT meetings, the optimal value of M on the development data was 15. Table 4.1 presents the speaker errors for the baseline IB system (Bas-IB), and the system using PBMs estimated from phone transcripts obtained from ground-truth (GT-PBM) and ASR (ASR-PBM) systems on the AMI and the NIST-RT development sets.

Table 4.1: *Diarization experiments on development sets: Speaker error for different systems on development set of meetings from AMI (20 meetings) and NIST-RT (RT-05,06) corpora.*

Corpus	Bas-IB	GT-PBM	ASR-PBM
AMI	24.3	17.4	20
NIST-RT	16.8	13.4	14.8

To evaluate the performance of the proposed method on the test set of meetings, we compare the performance of the baseline IB speaker diarization system using a normal background model without any information about the phoneme transcripts with the IB diarization system using the PBMs estimated using phoneme transcripts obtained from ground-truth reference transcripts and ASR system. Table 4.2 presents the evaluation results on the AMI data set. Speech/non-speech error (SpNsp) for all the systems is constant as the speech activity detector output given by the SHOUT toolkit [Huijbregts and de Jong, 2011] is systematically used. It can be observed from Tab. 4.2 that the baseline IB system achieves a DER of 38.2. Also, the system using PBM estimated from ground-truth phone transcripts (GT-PBM) gives the lowest DER among all the systems. It reduces the DER from 38.2 to 33.4. The PBM estimated from ASR transcripts also performs better than the baseline system. It reduces the DER from 38.2 to 34.8. This shows that the proposed method of using PBM for IB speaker diarization improves the performance of the baseline IB system. To check that phone transcripts generated by the ASR

Table 4.2: *Diarization experiments on 100 test meetings from AMI corpus: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems.*

System	SpNsp	Spkr	DER
Bas-IB	15.0	23.2	38.2
GT-PBM	15.0	18.4	33.4
ASR-PBM	15.0	19.8	34.8
Rand-PBM	15.0	23.0	38.0

system are providing reliable information for PBM estimation, we performed an experiment in which, the PBM is estimated by randomizing the phone transcript, and the phoneme class for a segment was chosen randomly out of the 45 phoneme classes. Since this randomization results in the loss of phoneme information, it is expected that the performance of this system will be similar to the baseline system which uses a background model estimated by ignoring the phoneme information. Tab. 4.2 also reports the performance of the system using the PBM estimated from randomized phone transcripts (Rand-PBM) which is similar to the baseline IB system(Bas-IB).

Tab. 4.3 compares the performance of the various systems on NIST-RT test set of meetings from RT-07,09 sets. It can be observed from Tab. 4.3 that the proposed method (ASR-PBM)

Table 4.3: *Diarization experiments on NIST-RT test set: Speech/non-speech error (SpNsp), Speaker error (Spkr) and total diarization error rate (DER) for different diarization systems on NIST-RT 07, 09 data sets.*

Corpus	System	SpNsp	Spkr	DER
RT 07	Bas-IB	3.7	10.8	14.5
	GT-PBM	3.7	8.3	12
	ASR-PBM	3.7	9.9	13.6
RT 09	Bas-IB	12.7	21.2	33.9
	GT-PBM	12.7	16.6	29.3
	ASR-PBM	12.7	18.2	30.9

reduces the DER of the baseline IB system from 14.5 to 13.6 on the RT-07 data set and from 33.9 to 30.9 on the RT-09 data set. Using ground-truth phone transcripts (GT-PBM) further reduces the error to 12.0 and 29.3 on the RT-07 and RT-09 data sets respectively.

4.4.2 Diarization experiments in agglomerative IBSI framework

This section presents the results of speaker diarization experiments in the IBSI framework using non-speech as side information for clustering. Experiments are conducted on meetings from the NIST RT 06, 07 and 09 evaluation data sets. The audio captured by multiple distant microphone channels is beamformed to get an enhanced signal using *BeamformIt*

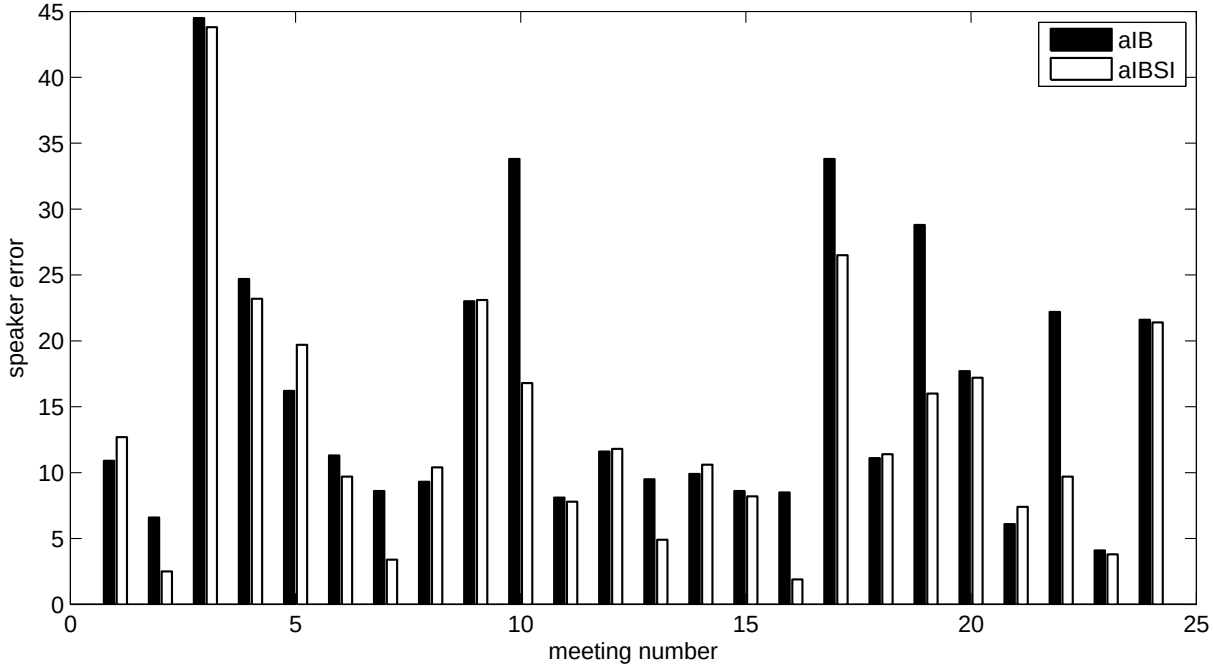


Figure 4.4: *Meeting wise speaker error values for baseline aIB diarization system and aIBSI system.*

toolkit [Anguera, X, , Anguera et al., 2007]. 19 dimensional Mel frequency cepstral coefficients (MFCCs) are extracted per each frame with a frame rate of 10 ms and frame length of 30ms. These features are used as input features for the speaker diarization system. Speech/non-speech detection is performed using the SHOUT system [Huijbregts and de Jong, 2011].

It was observed that non-speech segments detected by the SHOUT system contained instances of laughter in the meeting conversations. Since the aim of using data from non-speech regions in the current method is to capture the background environment in the meeting, these laughter instances have to be eliminated from the non-speech regions before using them in clustering. Since the laughter segments detected as non-speech by the SHOUT system are usually very loud as they involve several people laughing together, they can be easily separated from silence/background noise in the recording. In the current study, we used a simple short-term spectral energy based system to detect the laughter segments in non-speech segments detected by the SHOUT system. The detected laughter segments are excluded from non-speech regions and the remaining data is used to train the non-speech background model.

The optimal value of γ in (4.1) for the aIBSI framework is obtained by picking the value of the parameter that minimized speaker error on the RT 05 evaluation set of meetings which is used as a development set. The speaker error obtained for various values of γ on this set of meetings is plotted in Figure 4.5. These development experiments as reported in Figure 4.5 show that $\gamma = 0.1$ is optimal on the RT 05. Therefore, the parameter value is fixed to 0.1 when testing on the RT 06, 07 and 09 meetings. The value of β is fixed to 10 according to prior

work [Vijayasenan et al., 2009].

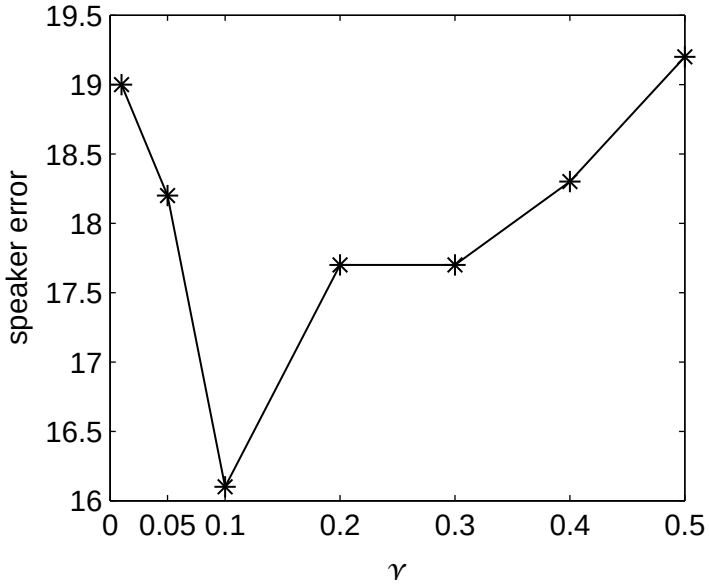


Figure 4.5: *Speaker error for various values of γ on development set of meetings from RT05 eval set.*

The performance of the two systems, the baseline aIB system and the proposed aIBSI system is measured in terms of Diarization Error Rate (DER) which is a standard metric used to evaluate speaker diarization systems in NIST RT evaluation campaigns [National Institute of Standards and Technology, 2003] given a reference ground-truth segmentation. DER is the sum of speech/non-speech error and the speaker error.

The performance of the automatic speech/non-speech detector in terms of miss (Miss) and false-alarm (FA) rates on the different test meeting sets is summarized in Table. 4.4. Since

Table 4.4: *Speech/non-speech errors for RT 06, 07 and 09 sets.*

Data set	Miss	FA	TOTAL
RT-06	6.5	0.1	6.6
RT-07	3.7	0	3.7
RT-09	11.6	1.1	12.7
ALL	7.3	0.4	7.7

automatic speech/non-speech output is kept constant for all the meetings for both the baseline aIB framework and the proposed aIBSI framework, we compare the meeting wise speaker error or the error in the clustering for the two systems in Figure 4.4. It can be observed from Figure 4.4 that the proposed method either decreases or makes insignificant changes to DER for most of the meetings.

In Table 4.5, we summarize the results at the data set level for the baseline aIB and the

proposed aIBSI diarization systems. It can be observed from this table that the proposed

Table 4.5: *Speaker error for aIB, aIBSI diarization systems (with relative improvements over aIB baseline in parenthesis) on RT 06, 07 and 09 sets. We also report the performance of HMM/GMM system for comparison.*

Data set	aIB	aIBSI
RT-06	16.8	14.9 (+11.3)
RT-07	10.8	9.8 (+9.2)
RT-09	21.2	15.3 (+27.8%)
ALL	16.3	13.3 (+18.4%)

method decreases speaker error on all the meeting sets when compared to the baseline aIB system. The overall speaker error on all three data sets is reduced from 16.6 to 13.3 by around 18.4% relative when compared to the baseline aIB system.

4.5 Summary

This chapter proposed two methods to incorporate auxiliary information sources which are non-speech regions and phoneme transcripts to improve IB diarization system. The information about non-speech regions was incorporated into IBSI clustering framework in the form of an irrelevant variable set represented by a set of components of background GMM estimated over non-speech regions in a given recording. Experimental results on meetings from the RT 06, 07, 09 meeting sets have shown that, the proposed method decreases the speaker error or the clustering error on all three data sets when compared to the baseline aIB diarization system. The combined speaker error on all three data sets was reduced from 16.3% to 13.3% which is a reduction of around 18% relative to the baseline aIB system. Also, meeting level comparison between the two systems showed that the proposed method decreases the speaker error on most of the meetings.

The information from phoneme transcripts was used to estimate a phoneme background model which was used instead of the regular background model used in IB diarization. Experiments have shown that incorporating information about "what is being spoken" (represented by phoneme transcripts) improves identification of "who is speaking when" (speaker diarization) in an information bottleneck based speaker diarization system. The estimation of a PBM was motivated by the oracle experiment, which showed that the background model estimated from the knowledge of speaker and phoneme being spoken is more useful for diarization than the background model estimated using only speaker information and a model estimated by ignoring both speaker and phoneme information. The PBM estimation was based on the hypothesis that clusters within a phoneme class roughly correspond to different speakers that have spoken it. The PBM was estimated by clustering the data within each phoneme class in a phoneme transcript of an audio file and representing each cluster with a Gaussian component in the PBM. The usefulness of such a PBM was evaluated by using it as a background model

in IB speaker diarization. Experiments conducted on meetings from the AMI and NIST-RT corpora showed that the PBMs estimated from ASR transcripts reduce the DER from 38.2 to 34.8 on the AMI corpus, from 14.5 to 13.6 on the RT-07 and 33.9 to 30.9 on the RT-09 data sets.

5 Artificial neural network features for speaker diarization

5.1 Introduction

The state-of-the-art speaker diarization systems typically use short-term spectral features, such as Mel-frequency cepstral coefficients (MFCCs) which represent the vocal tract characteristics of a speaker, as features for diarization. The MFCCs are not optimized for speaker discrimination as they also encompass various other information sources like phoneme being spoken, background noise and channel characteristics etc. In the current chapter, we propose new features based on artificial neural networks (ANNs) which are trained on tasks related to speaker diarization such as speaker comparison, speaker classification and auto encoding.

To extract speaker discrimininant features, previous works have adapted factor-analysis based techniques, which are popular in the speaker-verification domain, to the speaker diarization task [Shum et al., 2011]. These methods cluster i-vectors extracted from speech segments using a cosine similarity measure to provide speaker diarization output. Experiments on summed telephone channels have shown that i-vector based methods improve the performance of speaker diarization when compared to the traditional MFCC features. Another approach based on feature transforms uses linear discriminant analysis (LDA) after initial passes of diarization to obtain discriminative features [Lapidot and Bonastre, 2012]. However, none of these methods developed for two-party telephone conversations have so far been applied to multi-party, conference-style meetings.

In this work, we explore various artificial neural network (ANN) architectures to extract features for diarization. We train three different neural networks; which are then used to generate features for speaker diarization. The first network is trained to decide whether two given speech segments belong to the same or different speakers. The second network is trained to classify a given speech segment into a pre-determined set of speakers. The third network is an auto-encoder which is trained to reconstruct the input at the output layer with as low reconstruction error as possible. We hypothesize that the hidden layers of networks trained in this fashion should transform spectral features into a space more conducive to speaker discrimination. We propose to use the hidden layer activations from the bottleneck layer of

such networks as a new feature for speaker diarization. We conduct experiments to evaluate the usefulness of these bottleneck features for the task of speaker diarization on various meeting-room data sets.

5.2 ANN features for speaker diarization

Artificial neural networks are extensively used in supervised tasks such as automatic speech recognition and speaker identification/verification tasks. In these problems neural networks are trained to predict the posterior probabilities of the desired classes (speakers/phonemes). The posterior probabilities obtained from a neural network can be directly used to infer the class. Another way of using neural networks for these tasks is to use a network trained to identify the classes as a discriminative feature extractor. Here, the activations of the hidden layer prior to the final layer are used as input features to another classifier (typically HMM/GMM). This section explains different ANN architectures which are trained on tasks related to speaker diarization. Once trained, the networks are used to generate features for speaker diarization.

5.2.1 ANN for speaker comparison

Speaker diarization is an unsupervised task and there is no a-priori information about the speakers. Therefore, in this work, we propose a neural network that is trained to compare two given input speech segments and predict if they are spoken by the same or different speakers. Such a network, when fully trained, would initially transform the input features into a space which makes the task of prediction (same/different speaker) easier. Therefore, we extract features from this network to use as a new stream in a diarization system. Figure 5.1 shows the architecture of the four-layer network we use in this work. We split the input layer of the network into two halves, left and right, to represent acoustic features belonging to the two speech segments being compared. The first hidden layer (bottleneck) is also split into two halves similar to the input layer, so each half receives input from the respective input segment i.e., the right half of the hidden layer only gets input from the right half of the input layer and the left half from the left half of the input layer. We tie the weight matrices (denoted by W in Figure 5.1) connecting the right and left halves of input and hidden layers so that the network learns a single common transform for all speakers. The second hidden layer connects each half of the first hidden layer to the output layer. The output layer has two units denoting the class labels—-same or different speakers—deciding the identity/non-identity of the speakers providing the two input speech segments (segment1, segment2 in Figure 5.1). All the hidden layers have sigmoid activation functions and the output layer has a softmax function to estimate the posterior probabilities of the classes (same/different). We train the network using a cross-entropy objective function.

After training the network, we use the first hidden layer activations, before applying the sigmoid function, as features for speaker diarization. The resulting features obtained are a linear transform of the input features as they are extracted before applying the non-linearity.

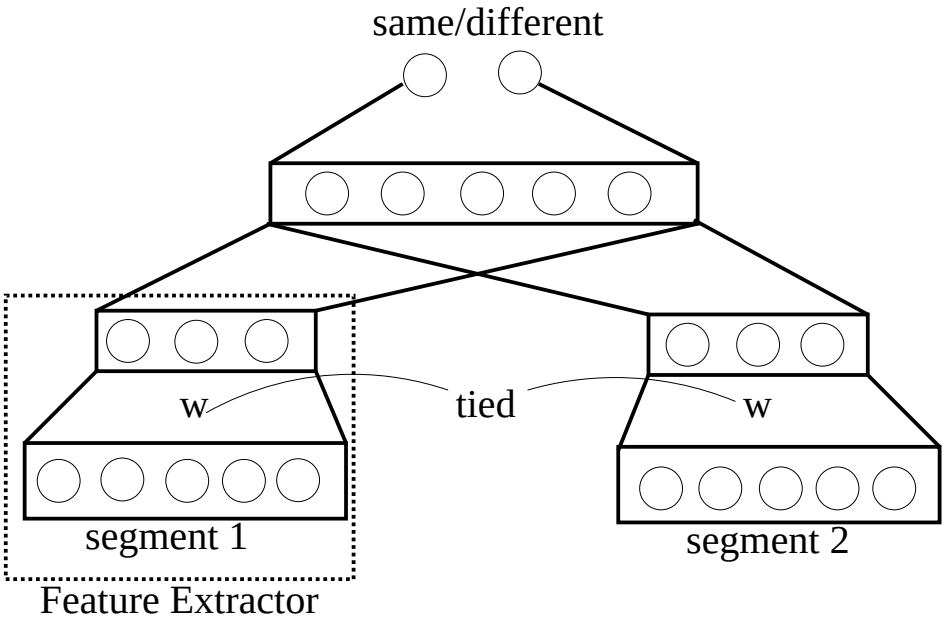


Figure 5.1: *An ANN architecture to classify two given speech segments as belonging to same or different speakers. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.*

To generate features from the network, we give a speech segment as input to one half of the input layer and extract activations at the corresponding half of the bottleneck layer. It should be noted that it does not matter to which half a speech segment is given as input to generate features since, the weight matrices connecting left and right halves of input layer to the corresponding halves in bottleneck layer are tied.

5.2.2 ANN for speaker classification

Konig et al. [Konig et al., 1998] used a multi-layer perceptron (MLP) with five layers, trained to classify speakers, as a feature extractor for speaker recognition. The MLP was discriminatively trained to maximize speaker-recognition performance. They used the outputs from the second hidden layer (units of which had linear activation function) as features in a standard GMM-based speaker-recognition system.

In the current work, we trained a similar network with speakers as output classes. The architecture of the network is depicted in Figure 5.2 The network is trained by providing a frame along with its context as input and the corresponding speaker as the output class label. The output layer uses a softmax function to estimate the posterior probability of the speaker. The second hidden layer (bottleneck) uses a linear activation function and the units in the rest of the hidden layers use a sigmoid non-linearity. After, training the network, the hidden layer activations obtained from the bottleneck layer (2nd hidden layer) of the network are used as features in speaker diarization. This network generates a non-linear transform of the

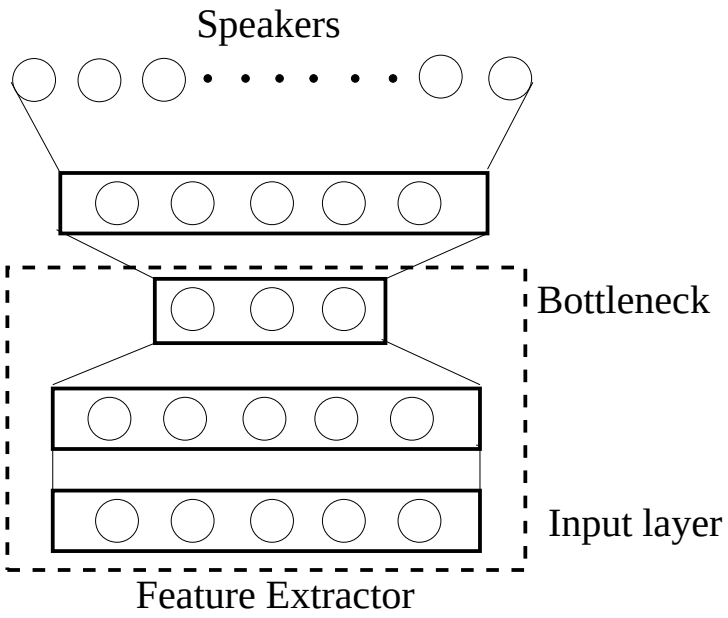


Figure 5.2: *An ANN architecture to classify a given input speech segments into one of speakers in the output layer. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.*

input features as the the input features are fed through the first hidden layer with a sigmoid activation function before extracting the activations at the second hidden layer.

5.2.3 ANN as an auto-encoder

Auto-encoders are used in the literature to generate feature representations and non-linear dimensionality reduction [Bengio, 2009]. An auto-encoder encodes the input in a representation which in turn is used to reconstruct the input. Therefore, while training the network the target at the output is the input it self. When the network has a single hidden layer with linear activations and is trained with mean square error criterion, it was shown that the activations at the n dimensional hidden layer project the input data into the space of first n principal components of the data [Bourlard and Kamp, 1988]. When the hidden layers have non-linear activations, the representation diverges from the principal component space and captures different aspects of the input feature distribution [Japkowicz et al., 2000]. In the current work, we use an auto-encoder with three hidden layers. The input to the network is the current frame along with its context. The network is trained to reconstruct the current frame at the output with as less re-construction error as possible. The network is trained with a mean square error criterion. The architecture of the network is depicted in Figure 5.3. All the hidden layers of the network use a sigmoid non-linearity. Once the network is trained, the features are generated by giving an input frame along with a context as input to the network and obtaining the activations of the bottleneck layer (second hidden layer) before applying the sigmoid non-linearity. The network performs a non-linear transform of the input features as they are

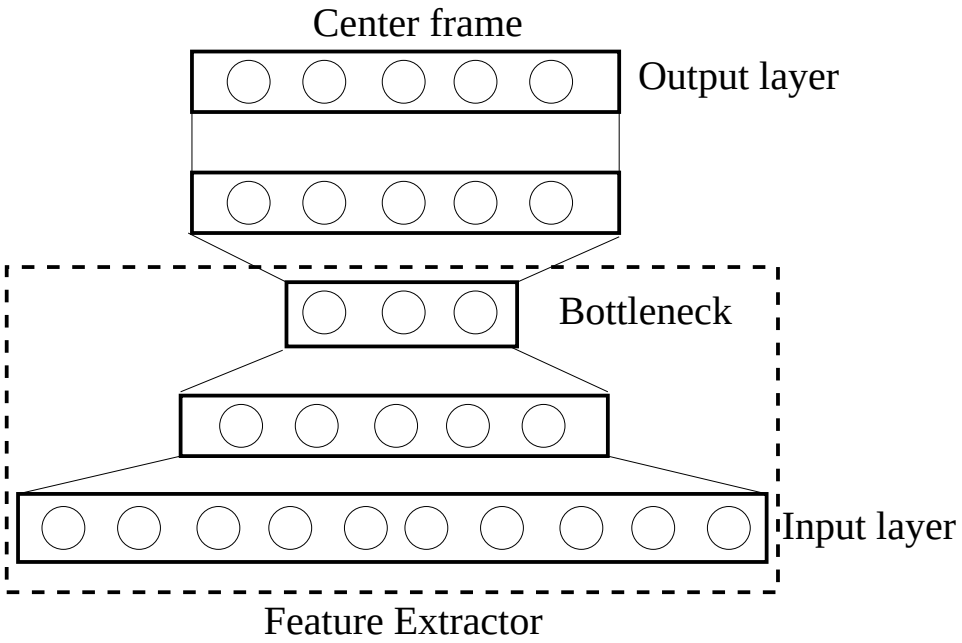


Figure 5.3: *An ANN as an auto-encoder: The network reconstructs the center frame of the input (center frame + context) at the output layer. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.*

Table 5.1: *Meeting corpus statistics as used in experiments, including numbers of distinct speakers, meeting room sites, and number of meetings used as part of train, development and test sets.*

Corpus	Speakers	Sites	Meetings		
			Train	Dev	Test
AMI	150	3	148	-	12
ICSI	50	1	-	20	55
NIST-RT	100	6	-	-	24

fed through sigmoid activation function in the first hidden layer.

5.3 Experiments and Results

This section describes the data sets, methodology, and experiments done on speaker diarization using ANN features.

5.3.1 Datasets used in experiments

The experiments make use of meeting room recordings from various corpora: AMI [Mccowan et al., 2005], ICSI [Janin et al., 2003], and 2006/2007/2009 NIST-RT [National Institute of Standards and Technology, 2003]. Table 5.1 summarizes the characteristics of these data sets.

The AMI data set is split into train and test sets of 148 and 12 meetings, respectively. The AMI-test and AMI-train sets are disjoint in speakers. We use only speech data from the AMI train set to train the neural network classifiers described in Section 5.2. Twenty ICSI meetings are set aside for the purpose of development and tuning, and the remaining 55 ICSI meetings form an additional test set. The ICSI-dev and ICSI-test sets have speaker overlap as the ICSI data set contains speakers that are common in most of the meetings. All NIST-RT evaluation sets (2006/2007/2009) are also used for testing.

5.3.2 Training of the ANNs

All the ANNs are trained using the data from the AMI-train set. To avoid skewing the training toward particular speakers we sampled 50 utterances from each of 138 speakers in the meetings of the AMI-train set. Each utterance has a duration of around 10 seconds. The cross validation (CV) set contained 10 utterances from each speaker in the training set. The networks are trained using 19-dimensional MFCC features extracted every 10 ms from a frame length of 30 ms. We force-aligned the manual speech transcripts to the close-talking microphone recordings to obtain frame-level speaker labels needed for training the ANNs. For training purposes we ignored speech segments containing overlapping speech. The objective function for the ANN was cross entropy for the speaker comparison and speaker classification networks and mean square error for the auto-encoder network. All the networks are trained using error back propagation and stochastic gradient descent for 25 epochs.

5.3.3 Speaker diarization evaluation

In this section, we report the speaker diarization experiments performed using the bottleneck features obtained from various ANN architectures. The bottleneck features are compared against the baseline 19 dimensional MFCC features extracted from one of the single distant microphone channels used to record the meeting room conversation. The diarization output is evaluated using a metric called diarization error rate (DER), which is a standard metric used in NIST-RT evaluation campaigns [National Institute of Standards and Technology, 2003]. We used speech/non-speech labels obtained from the ground-truth speaker segmentation to get the speech regions for diarization.

To evaluate the usefulness of the features obtained from the different ANN architectures for speaker diarization, we use them as features for diarization systems based on the HMM/GMM framework and the Information Bottleneck (IB) framework. To identify the optimal context length at the input to generate features for diarization from different ANN architectures, we performed tuning experiments on meetings from ICSI-dev set.

ANNs are trained with features of different context lengths varying between 10–50 frames (100–500 ms) at the input layer. The dimension of the bottleneck layer was fixed to 20. The bottleneck features obtained from these networks are used for speaker diarization. Figure 5.4

shows the speaker errors of the HMM/GMM system while using bottleneck features obtained from different ANN architectures with different input context lengths. It can be observed from

Figure 5.4: Input context length tuning using ICSI-dev set for various bottleneck features in HMM/GMM based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder.

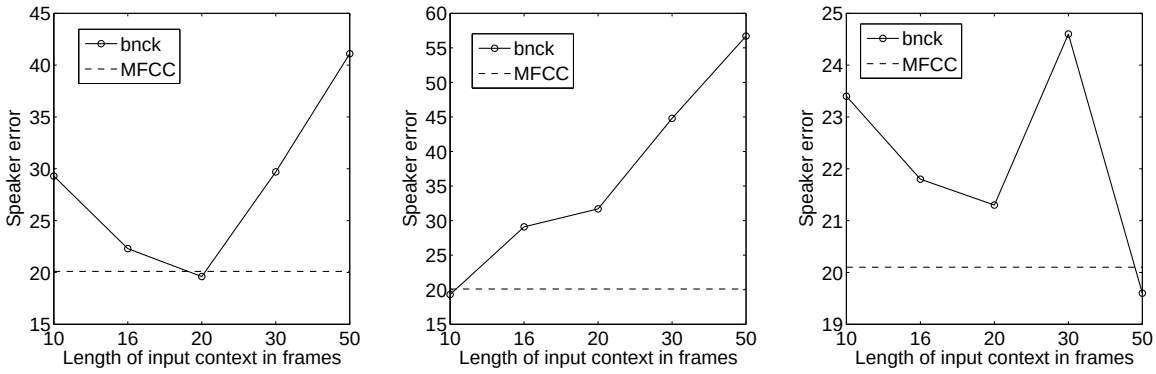
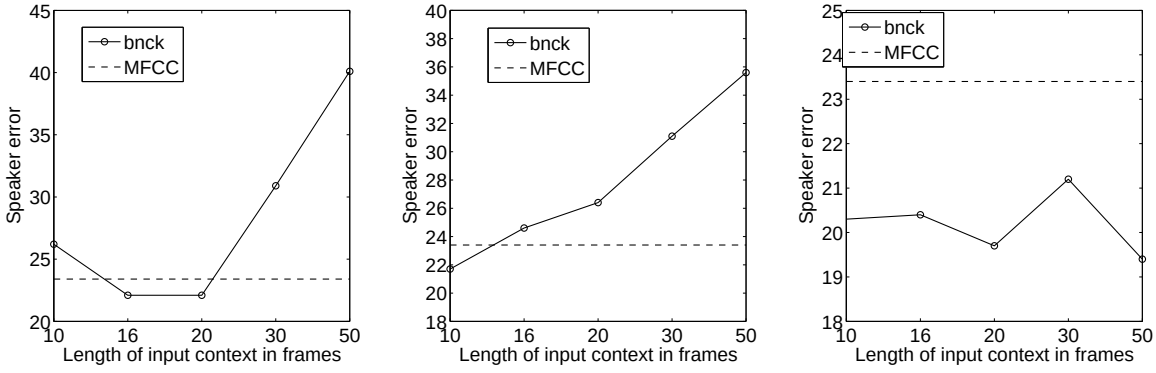


Figure 5.4 that the optimal input context length for speaker comparison ANN is 20 frames, for speaker classification network is 10 frames and for auto-encoder it is 50 frames. Similar plots for IB diarization system are shown in Figure 5.5.

Figure 5.5: Input context length tuning using ICSI-dev set for various bottleneck features in information bottleneck based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder.



It can be observed from Figure 5.4 and Figure 5.5 that the optimal input context lengths to generate ANN features are the same for both the HMM/GMM system and the IB system. The context lengths optimized in the above fashion using the development set are used to generate features for the meetings in test set. The results obtained for the different test sets are summarized in Tab. 5.2 for the HMM/GMM and the IB systems. It can be observed from Tab. 5.2 that, while the bottleneck features from the speaker comparison network (spkr-com)

Table 5.2: *Speaker errors on test data sets for various bottleneck features in HMM/GMM and IB speaker diarization system.*

System	Test-set	spkr-com	autoen	spkr-class	MFCC
HMM/GMM	AMI-test	22.9	25.9	29.3	24.8
	ICSI-test	23.1	20.9	19.8	19.8
	RT-06	26.6	11.0	22.1	14.5
	RT-07	16.8	9.7	16.7	12.0
	RT-09	20.9	16.4	25.6	16.3
IB	AMI-test	20.5	21.1	29.7	22.5
	ICSI-test	21.3	19.9	20.7	20.9
	RT-06	24.8	19.5	18.5	17.5
	RT-07	18.0	12.9	15.4	13.1
	RT-09	28.3	26.8	23.1	23.0

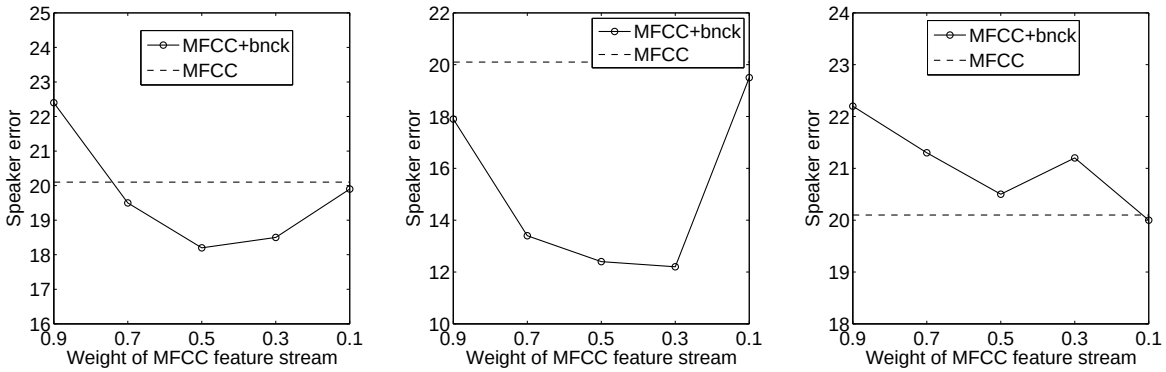
give lower speaker error than MFCC features on AMI-test set for both HMM/GMM and IB diarization systems, they increase the error on all the remaining test sets. Bottleneck features from the auto-encoder produce lower error on the RT-06/07 data sets for HMM/GMM systems and for the AMI-test and the ICSI-test sets for IB system. The bottleneck features obtained from speaker classification network increase error on all test sets for both systems except on the ICSI-test set for the HMM/GMM system. In summary, it can be observed that none of the neural network features perform as well as baseline MFCC features on all test sets.

Combination of MFCCs with ANN features.

Though the results in Tab. 5.2 show that the bottleneck features in general perform worse than MFCC features on most of the test sets, one can hypothesize that these features capture some complementary information to that of MFCC features, which can be helpful for speaker diarization. To test this hypothesis, we combine the MFCC features with bottleneck features obtained from different ANN architectures. The combination in the HMM/GMM system is performed at the model level, where separate GMM models are estimated for each feature stream for every cluster (state) and the final log-likelihood for a cluster is obtained as a weighted combination of log-likelihoods of individual feature streams.

In the IB system, the combination occurs in the posterior space. Here, the posteriors of the relevance variable set obtained for each feature stream are combined by weighted combination to obtain the final posteriors of the relevance variable set. The weights in both systems sum to unity and are fixed based on tuning experiments done on ICSI-dev set. The bottleneck features in both cases are obtained from the best architecture in the single stream case whose diarization performance is reported in Tab 5.2. Figure 5.6 shows the speaker error for different combination weights for MFCCs and various bottleneck features in HMM/GMM system. It can be observed from the Figure 5.6 that the optimal weights for the combination of MFCCs and bottleneck features from the speaker comparison network are 0.5 and 0.5 for MFCCs

Figure 5.6: Feature stream weight tuning using ICSI-dev set for combination of MFCC features with bottleneck features obtained from different ANNs in HMM/GMM based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder.

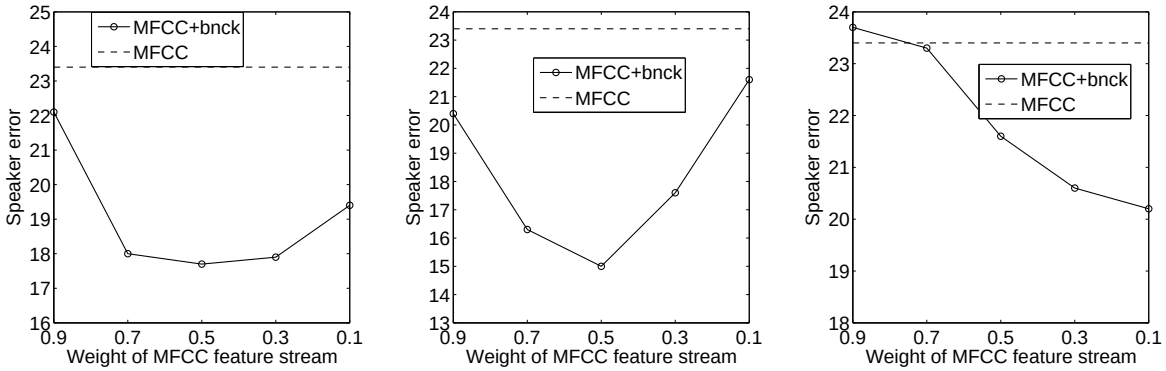


(a) (b) (c)

and bottleneck features respectively. The optimal weights for the combination of MFCCs and bottleneck features from the auto-encoder network are 0.1 and 0.9 for the MFCCs and bottleneck features respectively. The optimal weights for the combination of the MFCCs and bottleneck features from speaker classification network are 0.3 and 0.7 for the MFCCs and bottleneck features respectively.

Similar plots for weight combination tuning experiments on the ICSI-dev set are shown for the IB system in Figure 5.7. The Figure 5.7 shows that the optimal combination weights for

Figure 5.7: Feature stream weight tuning using ICSI-dev set for combination of MFCC features with bottleneck features obtained from different ANNs in information bottleneck based speaker diarization system: (a) Speaker comparison (b) Speaker classifier (c) Auto-encoder.



(a) (b) (c)

MFCCs and bottleneck features from the speaker comparison and auto-encoder networks are same as the optimal weights obtained for HMM/GMM system. In IB system, the optimal weights for combination of MFCCs and bottleneck features from the speaker classification network are 0.5 and 0.5 for MFCCs and bottleneck features respectively.

The optimal combination weights obtained for the MFCCs and the different bottleneck features are used for performing multi-stream diarization on the meetings from test sets. Tab. 5.3 shows the speaker errors of the HMM/GMM and the IB systems which combine MFCCs and different bottleneck features on various test sets. It can be observed from Table 5.3 that

Table 5.3: *Speaker errors on test data sets after combining different bottleneck features with MFCC features in HMM/GMM and IB speaker diarization systems.*

System	Test-set	M+spkr-com	M+autoen	M+spkr-class	MFCC
HMM/GMM	AMI-test	19.7	24.9	19.2	24.8
	ICSI-test	18.5	21.1	13.1	19.8
	RT-06	18.4	14.9	10.9	14.5
	RT-07	12.7	10.8	10.6	12.0
	RT-09	20.7	17.0	14.0	16.3
IB	AMI-test	19.9	21.3	18.8	22.5
	ICSI-test	17.0	19.2	15.3	20.9
	RT-06	17.9	18.7	17.1	17.5
	RT-07	13.5	12.5	10.0	13.1
	RT-09	24.5	24.2	21.0	23.0

the combination of MFCCs and bottleneck features from the speaker comparison network (M+spkr-com) reduces the error on the AMI-test and the ICSI-test sets while performance is worse than MFCC features on the RT-06/07/09 data sets in both the diarization systems. The combination of the MFCCs and the bottleneck features from the auto-encoder (M+autoen) does not yield significant changes from the single stream system using only auto-encoder based bottleneck features (refer to Table 5.2). This shows that the auto-encoder based bottleneck features are not capturing any information complementary to that of the MFCC feature stream. The highest gains are achieved by combining the MFCCs with bottleneck features from the speaker classification network. It can be observed from Tab. 5.3 that this combination consistently reduces the speaker error significantly in both the diarization systems on all test sets.

5.4 Summary

We have developed a speaker diarization framework that uses ANNs as trainable acoustic feature extractors. Three different ANN architectures are explored for this purpose. The first ANN, is trained to compare pairs of speech segments and predict if they are spoken by the same or different speakers, while forcing the raw MFCC features to undergo a shared transform via a bottleneck layer. The learned transform (linear) can then be applied to unseen data to generate features that are combined with the baseline MFCCs and given as input to standard agglomerative clustering diarization systems based on the HMM/GMM and the IB frameworks. We find that the resulting system reduces speaker error substantially when trained on data that is reasonably matched to the test data (AMI or ICSI test data when trained on AMI speakers not seen in testing).

The second ANN was trained to classify an input speech segment into one of the pre-determined speaker classes. The network consists of three hidden layers and the activations from the second hidden layer (which has a linear activation function) are used as features for speaker diarization. These features when combined with MFCCs produced significant reduction in speaker error for all the test data sets in both HMM/GMM and IB diarization systems.

The third ANN was trained to perform auto encoding. The network was trained to replicate the input at the output with as less re-construction error as possible. The network consisted of three hidden layers and the activations of the second hidden layer before applying the sigmoid are used as features for speaker diarization. These features did not show any significant reduction in the error when combined with MFCC features, which shows that they did not capture any complementary information to MFCCs.

In summary, the experiments in this chapter suggest that neural network features extracted from networks discriminatively trained on related classification tasks to speaker diarization capture complementary information to MFCCs and are useful for speaker diarization.

6 Combination of different diarization systems

6.1 Introduction

Speaker diarization evaluation campaigns conducted by NIST [National Institute of Standards and Technology, 2003] have revealed that there is a significant variance in performance of different systems based on different clustering frameworks [Knox et al., 2012, Mirghafori and Wooters, 2006]. These studies have observed that there is no single best system on all the meeting recordings used in the test sets. One way to overcome this issue is by combining different diarization systems so that the combination can take advantage of the complementary nature of the systems being combined. In the current chapter we propose a novel combination strategy for the parametric HMM/GMM based diarization system and non-parametric Information Bottleneck (IB) system. The combination takes place at the feature level where the output of the IB diarization system is used to generate features for the HMM/GMM system. The rationale behind this is that the new feature set will complement the HMM/GMM system at each step with the information provided from the output of the IB system. The proposed combination method does not require any changes to the original systems.

6.2 Previous approaches to system combination

A number of studies on broadcast data have discussed the combination of speaker diarization outputs from different systems to improve results. The simplest combination consists of voting schemes [Tranter, 2005] between the outputs of multiple systems. Also, a system can be initialized with the output of another one, such as in case of bottom-up and top-down diarization proposed in [Moraru et al., 2003]. Finally integrated approaches [Moraru et al., 2004], i.e., systems that integrate two different diarization methods into a single one, have been considered for broadcast data. Recently, they have also been revisited in the context of meeting recordings [Bozonnet et al., 2010a]. While combinations are able to outperform the individual diarization systems, each combination technique has advantages and pitfalls; in particular the voting scheme performs only late combination, i.e. at the output level. The initialization

approaches only benefit from a different starting point and the integrated approaches require modifications to all parts of the systems.

6.3 Combination of HMM/GMM and IB systems

The HMM/GMM and IB systems differ in a number of implementation issues which are summarized in Tab. 6.1. The HMM/GMM system uses GMMs as speaker models for clustering, where as the IB system uses components of background model as a set of relevance variables (Y) for clustering. The distance measure used for merging two clusters in HMM/GMM framework is based on modified ΔBIC where as in IB system it is based on Jensen-Shannon (JS) divergence. The output of HMM/GMM system assigns an input speech segment X to a cluster C. The output of IB diarization system, in addition to assigning the input segments to a cluster, each cluster is also represented by a relevance variable distribution $P(Y|C)$. Given these differences in the two systems, one could expect complementarity between them. Hence, we describe a combination method for HMM/GMM and IB diarization systems which captures this complementarity.

6.3.1 Motivation for the proposed method

The current combination method is largely inspired from the TANDEM framework used in Automatic Speech Recognition (ASR) [Hermansky et al., 2000]. TANDEM aims at using the probabilistic output of a Multi Layer Perceptron (MLP) that estimates phoneme posterior probabilities, as features to a conventional HMM/GMM system. Given an input speech frame X and a set of phonetic targets Y, the MLP estimates the posterior probabilities $p(Y|X)$. After that, the $p(Y|X)$ are first gaussianized using a logarithm and then de-correlated with a PCA transform followed by a dimensionality reduction. The resulting features are referred as the TANDEM features. After concatenation with MFCC, they are used to train a standard HMM/GMM system. TANDEM features are able to reduce the Word Error Rate (WER) by $10-15\%$ relative (see [Morgan et al., 2005] for a review of tasks and improvements) thus complementing well the standard spectral features. However, contrary to ASR, speaker diarization is an unsupervised task thus there is no direct equivalent to the phoneme posterior probabilities $p(Y|X)$. This work proposes to generate TANDEM-like features using the probabilistic output of the Information Bottleneck system [Vijayasenan et al., 2009].

Table 6.1: Main differences between the HMM/GMM and the IB diarization systems.

	HMM/GMM	IB
Modeling	a separate GMM for each speaker c	relevance variables Y from a background GMM
Distance	Modified ΔBIC	JS divergence
Output	mapping $X \rightarrow C$	mapping $X \rightarrow C$ and $p(Y\|C)$

6.3.2 Information Bottleneck features

This section describes how the output of the IB system can be used as features in HMM/GMM diarization. Consider MFCC feature vectors $S = \{s_1, \ldots, s_T\}$ where s_t denotes the feature vector at time t; those are then segmented in $X = \{x_j\}$ chunks each containing D consecutive speech frames (feature vectors). The feature vectors S can be re-designated as $S = \{s_t^j\}$, where the superscript j denotes to which segment the feature vector belongs to. The output of the IB diarization is a hard partition of speech segments $x_j \in X$ into C clusters, i.e., $p(c_i|x_j) \in \{0,1\}$, meaning that each segment x_j is assigned to only one cluster. For each cluster, the associated relevance variable distribution $p(Y|c_i)$ is available.

Thus each feature vector s_t^j belonging to segment x_j (given by the initial segmentation) can be associated to a cluster z_t obtained from the diarization output, i.e.,

$$z_t = \{c_i | s_t^j \in x_j, p(c_i|x_j) = 1\}, \;\; t = 1, ..., T. \tag{6.1}$$

Let F denote a matrix that contains the relevance variable distributions $p(Y|z_t)$ associated with each z_t, i.e.,

$$F = [p(Y|z_1), ..., p(Y|z_T)], \;\; t = 1, ..., T. \tag{6.2}$$

F is a $|Y| \times T$ matrix where T is the number of speech frames and $|Y|$ is the cardinality of the relevance variable space.

F contains both information on the clustering output (if two feature vectors s_t and $s_{t'}$ belong to the same cluster), and characterizes each cluster with the distribution $p(Y|z_t)$ (different clusters will have different $p(Y|z_t)$). Thus TANDEM processing [Hermansky et al., 2000] can be applied, probabilities $p(Y|z_t)$ are gaussianized by a logarithm on their individual elements and then de-correlated using Principal Component Analysis (PCA). The PCA is also used to reduce the initial dimensionality, equal to the relevance variable space cardinality ($|Y|$). The resulting matrix, designated as F_{IB} and referred to as Information Bottleneck (IB) features can be used as input to a conventional diarization system where the GMM speaker models can be learnt from these features. The integration with MFCC can happen in two possible ways:

1 by concatenating IB features with MFCC features (as done in ASR) thus forming a single input vector to the HMM/GMM system. This approach will be referred as IB_aug (the IB feature stream is augmented with MFCC features).
2 by multi-stream modeling, i.e., estimating a separate GMM model for each feature stream and combining their log-likelihoods [Pardo et al., 2007]. This approach is used for instance in diarizing with features having very different statistics (like MFCC and Time Delay of Arrival features) and will be referred to as IB_multistr. In this case, the

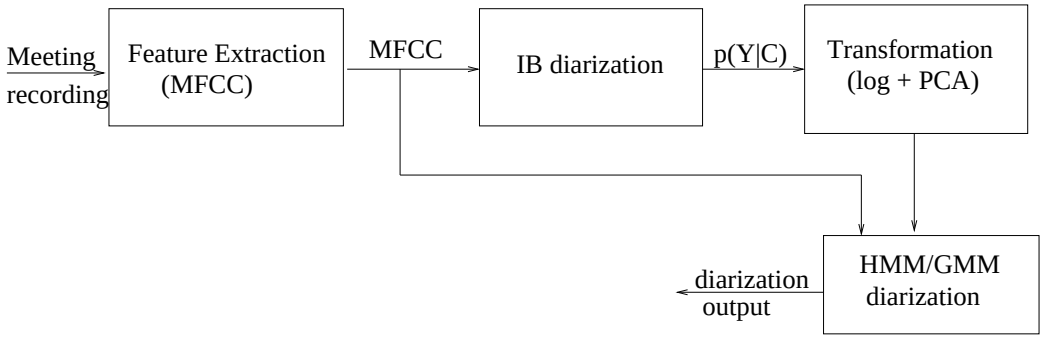


Figure 6.1: *Block diagram of the proposed method.*

clustering is based on the combined log-likelihood:

$$w_{mfcc} \log b_c^{mfcc} + w_{F_{IB}} \log b_c^{F_{IB}} \tag{6.3}$$

where b_c^{mfcc} and $b_c^{F_{IB}}$ are GMMs trained on MFCC and F_{IB} features and $(w_{mfcc}, w_{F_{IB}})$ are the combination weights.

The overall method can be summarized in three main steps given below and a block diagram of the proposed approach is shown in Figure 6.1:

1 Perform IB diarization and estimate $p(C|X)$ and $p(Y|C)$.
2 Map $p(Y|C)$ to input frames S and apply TANDEM processing to obtain IB features (F_{IB})
3 Use F_{IB} as complementary features to MFCC in a conventional HMM/GMM system.

6.4 Experiments and Results

The experiments are conducted on 24 meetings recordings from the NIST-RT data set [National Institute of Standards and Technology, 2003]. We used meetings from NIST RT06, RT07, RT09 evaluation sets for tesing and RT05 set for development experiments. The audio from multiple distant microphone channels of each meeting is beamformed using the *BeamformIt* toolkit [Anguera, X,]. The beamformed output of each meeting is used for speech, non-speech detection and feature extraction. Acoustic features consist of 19 MFCCs. The speech/non-speech detection is based on the SHOUT toolkit [Huijbregts and de Jong, 2011] and evaluated in terms of missed speech rate (Miss) and false alarm rate (FA) summing into the speech/non-speech error rate (SpNsp) (see Tab. 6.2). The performance is evaluated in terms of Diarization Error Rate (DER) which is the sum of speech/non-speech error and speaker error. For the purpose of comparison, only speaker error is reported here as same speech/non-speech is used for all the systems.

The number of principal components to be kept after PCA and the weights $(w_{mfcc}, w_{F_{IB}})$ are selected as the ones that minimize the speaker error on a separate development data set. The optimal number of principal components is found to be equal to two, covering more than 80% of the PCA variability. The feature weights $(w_{mfcc}, w_{F_{IB}})$ are found to be equal to $(0.9, 0.1)$.

Table 6.2: *Speech/non-speech error rate in terms of missed speech (Miss), false alarm speech (FA) and the total error (SpNsp) in the evaluation data sets.*

Data-set	Miss	FA	SpNsp
RT-06	6.5	0.1	6.6
RT-07	3.7	0.0	3.7
RT-09	11.6	1.1	12.7
ALL	7.3	0.4	7.7

These values are then used for evaluation on the RT06, RT07, RT09 meetings. Tab. 6.3 reports speaker error for the baseline system as well as the IB_aug and IB_multistr approaches. The meeting-wise performance is reported in Figure 6.2 and data-set wise performance is reported in Table 6.4.

Table 6.3: *Total speaker error with relative improvement over baseline in parenthesis on the evaluation data sets (RT06, RT07, RT09 combined) for various diarization systems.*

system	Baseline	IB_aug	IB_multistr
spkr err	12.0 (-)	13.5 (-12.5%)	9.7 (+19%)

Table 6.4: *Data-set wise speaker error on RT06, RT07, RT09 data sets for various diarization systems.*

corpus	system	Spkr
RT-06	HMM/GMM	13.6
	IB_aug	14.7
	IB_multistr	10.3
RT-07	HMM/GMM	6.4
	IB_aug	7.5
	IB_multistr	6.0
RT-09	HMM/GMM	14.3
	IB_aug	18.1
	IB_multistr	12.7

The baseline HMM/GMM system achieves a speaker error equal to 12%. The first approach IB_aug, which concatenates MFCC and F_{IB} features, degrades the performance producing an error equal to 13.5%. On the other hand, the second approach IB_multistr, which estimates separate GMM models for MFCC and F_{IB} features, reduces speaker error to 9.7%, i.e., an improvement of approximatively 19% relative to the baseline. The degradation in performance produced by concatenation can be explained by the very different statistical properties of the MFCC and F_{IB} features. In fact, F_{IB} features have smaller dimensionality compared to the MFCC and are a compact representation of IB diarization output, thus they do not share the same distribution as MFCC. Therefore, whenever the modeling is done using separate GMMs, speaker error decreases from 13.5% (IB_aug) to 9.7% (IB_multistr). This is similar to what was

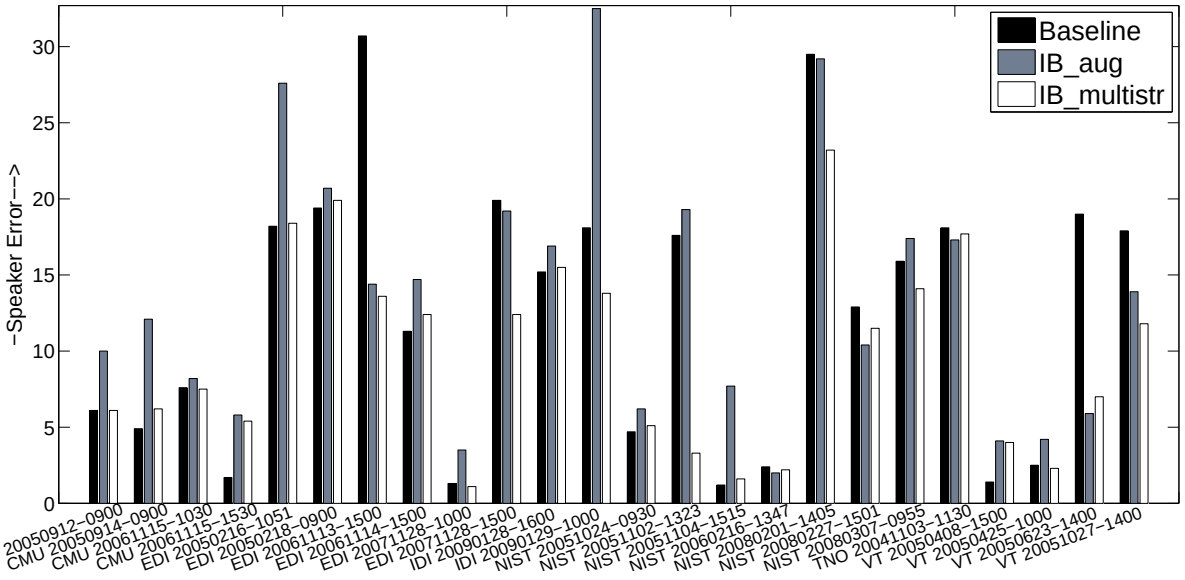


Figure 6.2: *Meeting wise speaker error values for baseline HMM/GMM diarization system and IB_multistr.*

observed in case of TDOA features, as they also become affective only through multistream modeling [Pardo et al., 2007].

It can be noticed from Figure 6.2 that the IB_multistr shows significant improvement upon the baseline system in meetings with high error (over 15%). It is observed that the IB features have an effect on purity of clusters, i.e., assignment of segments uttered by different speakers to the same clusters is reduced, thus producing much purer clusters compared to MFCC only (baseline). Reversely IB_aug often degrades the performance.

We also investigate the effect of the F_{IB} features at different stages of the clustering. Figure 6.3 plots the speaker error for the baseline and IB_multistr after each merge, for the meeting EDI_20061113-1500. It can be noticed that both systems have similar error rates in initial iterations but after a few iterations, the F_{IB} features avoid wrong cluster merges, which increase error rate and produce a smooth and decreasing error curve. Similar trends are verified for other meetings where IB_multistr achieves improvements over baseline.

We also combine the diarization systems using non-speech and phoneme transcripts proposed in this thesis with HMM/GMM system. Here, the features for HMM/GMM system are generated in a similar way as explained in Section 6.3.2 but using the output of agglomerative information bottleneck with side information (IBSI) clustering framework. In this system we use the phoneme background model (PBM) as relevant variable set for clustering and the components of GMM estimated from non-speech regions as irrelevant variable set. The diarization output of this system is used to generate features for HMM/GMM system. The results obtained by this combination on various data sets are presented in Table 6.5. It also

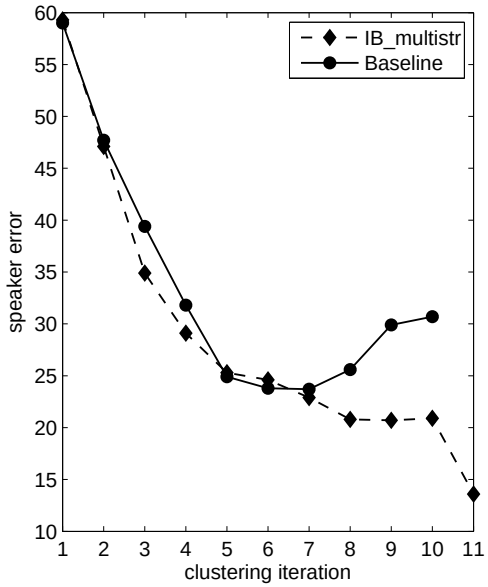


Figure 6.3: *Speaker error after each merge for baseline and IB_multistr systems for meeting EDI_20061113-1500*

shows the results for combination of IB and HMM/GMM systems (IB_multistr) for reference.

Table 6.5: *Combination of IBSI and HMM/GMM systems: The IBSI system uses phoneme background model (PBM) as relevance variable set and components of GMM estimated from non-speech regions as irrelevant variable set. Speaker error is reported for various test sets NIST-RT 06/07/09.*

corpus	system	Spkr
RT-06	HMM/GMM	13.6
	IB_multistr	10.3
	IBSI_multistr	9.7
RT-07	HMM/GMM	6.4
	IB_multistr	6.0
	IBSI_multistr	5.8
RT-09	HMM/GMM	14.3
	IB_multistr	12.7
	IBSI_multistr	11.4

It can be observed from Table 6.5 that the combination of the IBSI and HMM/GMM systems (IBSI_multistr) reduces the speaker error further when compared to combination of IB and HMM/GMM systems (IB_multistr). Also, the reduction in errors is consistent across all the test sets. This can be attributed to the improved speaker diarization output in IBSI framework when compared to IB system.

6.5 Summary

This chapter proposes and investigates a novel approach to combine diarization systems through the use of *features*. The Information Bottleneck system is used to generate a set of features that contain information relevant to the clustering and characterizes each speaker in terms of probabilities; these features are later used to complement MFCC in a conventional HMM/GMM system. The approach is largely inspired from the TANDEM framework used in ASR and has the advantage of being fully integrated (features are used at all steps of agglomerative clustering), while it does not require any change to individual diarization components.

The combination with MFCC features is investigated using simple concatenation and using multi-stream modeling. Results on 24 meetings from the NIST RT06/RT07/RT09 evaluation campaigns reveal that the Information Bottleneck features reduce the speaker error from 12% to 9.7%, i.e., a 19% relative improvement when they are used together with MFCC in multi-stream fashion. The improvements are consistent across all the test sets. The approach is particularly effective in meetings where the baseline system has a speaker error higher than 15%. On the other hand, simple concatenation increases speaker error to 13.5% as F_{IB} and MFCC have very different statistical distributions to be modeled using same GMM. In summary the IB system provides complementary information to the HMM/GMM whenever the integration is based onby multi-stream modeling.

The combination of IBSI and HMM/GMM systems has yielded further reduction in the error across all the test sets. Here, the combination was performed using the features generated from the IBSI diarization output. This system used components of phoneme background model (PBM) as the relevant variable set and the components of background GMM estimated from non-speech regions as irrelevant variable set. This shows that improvements observed in the diarization output in IBSI framework also improve the combination of IBSI and HMM/GMM system.

7 Conclusion and Future Directions

7.1 Conclusion

In this thesis, we have tackled issues in the state-of-the-art speaker diarization systems such as overlapping speech, effects of background noise, errors in speech/non-speech detection, and performance variations across different systems.

To handle overlaps, we proposed to use long-term features extracted from the structure of a conversation to improve the overlap detection system based on acoustic features. These features such as silence and speaker change statistics are computed over a long-term window (3-4 seconds) and are used to predict the probability of overlapping speech in the window. These probabilities are incorporated into the overlap detection system as prior probabilities of overlapping, single-speaker speech classes. Experiments on meetings from different meeting corpora revealed that the proposed method improves the overlap detection system significantly, which also results in the reduction of diarization error.

To suppress the effects of background noise and errors made in speech/non-speech detection, we proposed the use of auxiliary information from non-speech regions in a given recording in the Information Bottleneck with Side Information (IBSI) based clustering framework. The objective function of the IBSI clustering tries to find a clustering representation that maximizes the mutual information w.r.t the relevant variable set i.e., speech regions, while minimizing the mutual information w.r.t the irrelevant variable set (non-speech segments). Experimental results on meeting corpora showed that IBSI based speaker diarization system reduces the diarization error significantly when compared to the baseline Information Bottleneck (IB) system.

To capture speaker discriminative information present in a speech signal, we proposed to use one more auxiliary information source in the form of phoneme transcript of a given recording. The use of phoneme transcript as auxiliary information for speaker diarization is motivated from previous works in speaker verification/identification, where the knowledge of what is being spoken was shown to be useful to identify who is speaking. Oracle experiments using

ground-truth transcripts showed that phoneme transcripts are indeed useful for speaker diarization. Motivated from these findings, we proposed a Gaussian mixture model based Phoneme Background Model (PBM) as a replacement to the normal background model in IB diarization system. The phoneme background model (PBM) is estimated from clusters within different phonemes in a given transcript. Experiments on meeting corpora have shown that the use of PBM as a background model in the IB diarization system reduces the error.

In an another attempt to extract speaker discriminative information, we trained Artificial Neural Networks (ANNs) to extract features relevant for speaker diarization. We investigated three ANN architectures for this task. The first network was trained to compare two given speech segments and predict if they are spoken by the same speaker or not. The second network was trained to classify a given input speech segment into on of the pre-determined speaker classes. The third network was trained as an auto-encoder which aims to reconstruct the input at the output with as less reconstruction error as possible. The bottleneck layer activations from these three networks are used for speaker diarization. Experimental results showed that bottleneck features from all the ANNs perform poorly when used on their own in both HMM/GMM and IB diarization systems. Bottleneck features from the speaker comparison network (first ANN) reduce error when testing on data reasonably matched to training data when combined with MFCCs. Bottleneck features from speaker classification network consistently reduce error on all test data sets when combined with MFCCs. On the other hand, bottleneck features obtained from the auto-encoder (third ANN) do not change the error significantly when combined with MFCC features, which shows that it did not capture any information complementary to the MFCCs.

In the final part of the thesis we proposed a method to combine two speaker diarization systems; HMM/GMM and IB that are based on different clustering frameworks to exploit the complementary nature of these systems. The combination takes place at the feature level and does not require any changes to the original systems. This combination method is inspired from the TANDEM method used in automatic speech recognition (ASR) systems. Similar to what is done in TANDEM processing, the posterior distributions over a set of relevance variables obtained for different clusters as a by-product of the IB diarization are used as features in HMM/GMM diarization system. These features referred to as IB features are combined with MFCCs in a multi-stream mode in a HMM/GMM system. This method of combination of the two systems has yielded a significant reduction in the error rate over both individual systems on meetings from various datasets.

7.2 Future directions

The work done in this thesis motivates several future directions of research. They are outlined below.

In current work, we have used overtly available features such as silence and speaker change statistics to improve overlap speech detection. Several studies [Cetin and Shriberg, 2006b,

Cetin and Shriberg, 2006a, Shriberg et al., 2001] on meeting room conversations have shown that overlaps are strongly related to various other factors such as type of dialogue acts [Stolcke et al., 2000] and hot spots [Wrede et al., 2005]. Dialogue acts (DAs) convey the nature of speech act such as question, statement, back-channel etc.,. Hot spots [Wrede et al., 2005] in a multi-party conversation are defined as regions where participants are highly involved in the conversation. Studies have shown that overlaps are highly correlated with some of the DAs such as back-channels and interruptions. Though DA tagging and hot spot detection are not trivial tasks, future works on overlap detection would benefit from leveraging the relation between the DAs, hot spots, and overlap occurrence.

The evaluation results on overlapping speech diarization have shown that reduction in diarization error by overlap labelling which attributes speaker labels to detected overlaps requires a high precision overlap detector. This is due to the reason that the current overlap labelling methods are based on heuristics like nearest speakers in time and cluster likelihoods. There is a great scope of improvement in overlapping speech diarization if a overlap labelling method that can overcome this problem is proposed.

The Information Bottleneck with Side Information (IBSI) based clustering framework was proposed in the current work to suppress the effects of background noise and errors in speech/non-speech detection on speaker diarization. Here, the side information is provided in the form of an irrelevant variable set, which consists of components of background GMM estimated over non-speech regions of a given recording. Future works can explore the applicability of this framework in different scenarios to suppress influences of specific channel or meeting room environment conditions. Also, pre-trained non-speech background models trained on various types of ambient noises encountered in a typical meeting room environment could be used instead of using background model estimated from non-speech regions of a given recording.

The phoneme background model was proposed in the current work to capture speaker dependent phoneme articulation. Phonemes are very often influenced by their context. Therefore, future works can explore the use of context dependent phonemes or triphones to estimate such a background model to alleviate the effects of context.

Features extracted from three different ANNs are proposed for the purpose of speaker diarization. The features extracted from ANN trained to compare two speech segments are only a linear transform of the input features as they are extracted from the first hidden layer before applying the non-linearity. Exploring deeper architectures with more hidden layers and extracting features from these deep layers can be more useful for speaker diarization. Several other neural network architectures such as convolutional networks [Lecun et al., 1998] and recurrent networks [Graves et al., 2009] can also be used to extract features.

Bibliography

[Adam et al., 2002a] Adam, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., and Sivadas, S. (2002a). Qualcomm-ICSI-OGI features for asr. In *Proc. ICSLP*, pages 4–7.

[Adam et al., 2002b] Adam, A. G., Kajarekar, S. S., and Hermansky, H. (2002b). A new speaker change detection method for two-speaker segmentation. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages 3908–3911.

[Adda-Decker et al., 2008] Adda-Decker, M., Claude, B., Gilles, A., Patrick, P., Philippe, B. d. M., and Benoit, H. (2008). Annotation and analysis of overlapping speech in political interviews. In *Proceeding of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[Ajmera, 2004] Ajmera, J. (2004). *Robust audio segmentation*. PhD thesis, Ecole polytechnique fédérale de Lausanne.

[Ajmera et al., 2004] Ajmera, J., McCowan, I., and Bourlard, H. (2004). Robust speaker change detection. *Signal Processing Letters, IEEE*, 11(8):649–651.

[Ajmera and Wooters, 2003] Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition Understanding Workshop*, pages 411–416.

[Anguera, 2006] Anguera, X. (2006). *Robust speaker diarization for meetings*. PhD thesis, Universitat Politecnica de Catalunya.

[Anguera et al., 2006a] Anguera, X., Aguilo, M., Wooters, C., Nadeu, C., and Hernando, J. (2006a). Hybrid speech/non-speech detector applied to speaker diarization of meetings. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–6.

[Anguera and Bonastre, 2010] Anguera, X. and Bonastre, J.-F. (2010). A novel speaker binary key derived from anchor models. In *Interspeech*, Makuhari, Japan.

[Anguera and Bonastre, 2011] Anguera, X. and Bonastre, J.-F. (2011). Fast speaker diarization based on binary keys. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4428–4431.

[Anguera et al., 2012] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):356–370.

[Anguera et al., 2006b] Anguera, X., Woofers, C., and Hernando, J. (2006b). Purity algorithms for speaker diarization of meetings data. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I.

[Anguera et al., 2007] Anguera, X., Wooters, C., and Hernando, J. (2007). Acoustic beamforming for speaker diarization of meetings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2011–2022.

[Anguera et al., 2005] Anguera, X., Wooters, C., Peskin, B., and Anguilo, M. (2005). Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *In Proc. of NIST MLMI Meeting Recognition Workshop*, pages 26–38.

[Anguera, X,] Anguera, X. Beamformit: Fast and robust acoustic beamformer. http://www.xavieranguera.com/beamformit/.

[Bengio, 2009] Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127.

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

[Boakye et al., 2008] Boakye, K., Vinyals, O., and Friedland, G. (2008). Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech. In *Interspeech*, pages 32–35, Brisbane, Australia.

[Boakye et al., 2011] Boakye, K., Vinyals, O., and Friedland, G. (2011). Improved overlapped speech handling for speaker diarization. In *Interspeech*, pages 941–943, Florence, Italy.

[Bourlard and Kamp, 1988] Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4-5):291–294.

[Bozonnet et al., 2010a] Bozonnet, S., Evans, N., Anguera, X., Vinyals, O., Friedland, G., and Fredouille, C. (2010a). System output combination for improved speaker diarization. In *Interspeech*, pages 2642–2645, Makuhari, Japan.

[Bozonnet et al., 2010b] Bozonnet, S., Evans, N., and Fredouille, C. (2010b). The LIA-EURECOM RT-09 speaker diarization system: enhancements in speaker modelling and cluster purification. In *ICASSP*, pages 4958–4961, Dallas, USA.

[Brandstein and Silverman, 1997] Brandstein, M. and Silverman, H. (1997). A robust method for speech signal time-delay estimation in reverberant rooms. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, pages 375 –378 vol.1.

[Castaldo et al., 2007] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. (2007). Compensation of nuisance factors for speaker and language recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):1969–1978.

[Cetin and Shriberg, 2006a] Cetin, O. and Shriberg, E. (2006a). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition. In *ICSLP*, pages 293–296, Pittsburgh, USA.

[Cetin and Shriberg, 2006b] Cetin, O. and Shriberg, E. (2006b). Overlap in meetings: Asr effects and analysis by dialog factors, speakers, and collection site. In *3rd Joint Workshop on Multimodal and Related Machine Learning Algorithms*, Washington DC, USA.

[Chechik, 2003] Chechik, G. (2003). *Information theoretic approach to the study of auditory coding*. PhD thesis, Hebrew University.

[Chechik and Tishby, 2003] Chechik, G. and Tishby, N. (2003). Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems*, pages 857–864. MIT press.

[Chen and Gopalakrishan, 1998] Chen, S. and Gopalakrishan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Broadcast News Transcription and Understanding Workshop*, pages 127–132.

[Delacourt and Wellekens, 2000] Delacourt, P. and Wellekens, C. J. (2000). DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Commun.*, 32(1-2):111–126.

[Dines et al., 2006] Dines, J., Vepa, J., and Hain, T. (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Ninth IEEE International conference on Spoken Language Processing*.

[Eatock and Mason, 1994] Eatock, J. P. and Mason, J. (1994). A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, pages 133–136 vol.1, Adelaide, Australia.

[El-Khoury et al., 2009] El-Khoury, E., Senac, C., and Pinquier, J. (2009). Improved speaker diarization system for meetings. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4097–4100.

[Fredouille et al., 2009] Fredouille, C., Bozonnet, S., and Evans, N. W. D. (2009). The LIA-EURECOM RT'09 speaker diarization system. In *RT'09, NIST Rich Transcription Workshop*, Melbourne, FL.

[Fredouille and Evans, 2008] Fredouille, C. and Evans, N. (2008). The LIA RT'07 speaker diarization system. In Stiefelhagen, R., Bowers, R., and Fiscus, J., editors, *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 520–532. Springer Berlin Heidelberg.

[Fredouille et al., 2004] Fredouille, C., Moraru, D., Meignier, S., Besacier, L., and francois Bonastre, J. (2004). The NIST 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation. In *The NIST 2004 spring rich transcription evaluation workshop*, Montreal, QC, Canada.

[Fredouille and Senay, 2006] Fredouille, C. and Senay, G. (2006). Technical improvements of the e-hmm based speaker diarization system for meeting records. In Renals, S., Bengio, S., and Fiscus, J., editors, *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 359–370. Springer Berlin Heidelberg.

[Friedland et al., 2010] Friedland, G., Chong, J., and Janin, A. (2010). Parallelizing speaker-attributed speech recognition for meeting browsing. In *Multimedia (ISM), 2010 IEEE International Symposium on*, pages 121–128.

[Friedland et al., 2009] Friedland, G., Vinyals, O., Huang, Y., and Muller, C. (2009). Prosodic and other long-term features for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):985–993.

[Gangadharaiah et al., 2004] Gangadharaiah, R., Narayanaswamy, B., and Balakrishnan, N. (2004). A novel method for two-speaker segmentation. In *Interspeech*, Jeju island, Korea.

[Geiger et al., 2012] Geiger, J., Vipperla, R., Bozonnet, S., Evans, N., Schuller, B., and Rigoll, G. (2012). Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization. In *Interspeech*, Portland, USA.

[Gish et al., 1991] Gish, H., Siu, M.-H., and Rohlicek, R. (1991). Segregation of speakers for speech recognition and speaker identification. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 873–876 vol.2.

[Glenn, 2003] Glenn, P. (2003). *Laughter in Interaction*. Cambridge University Press.

[Graves et al., 2009] Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):855–868.

[Griffiths and Jim, 1982] Griffiths, L. J. and Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *Antennas and Propagation, IEEE Transactions on*, 30(1):27–34.

[Han and Narayanan, 2008] Han, K. J. and Narayanan, S. S. (2008). Agglomerative hierarchical speaker clustering using incremental gaussian mixture cluster modeling. In *Interspeech*, Lisbon, Portugal.

[Hermansky et al., 2000] Hermansky, H., Ellis, D., and Sharma, S. (2000). Tandem connectionist feature extraction for conventional hmm systems. In *ICASSP*, volume 3, pages 1635–1638.

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589.

[Huang et al., 2007] Huang, Y., Vinyals, O., Friedland, G., Muller, C., Mirghafori, N., and Wooters, C. (2007). A fast-match approach for robust, faster than real-time speaker diarization. In *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on*, pages 693–698.

[Huijbregts and de Jong, 2011] Huijbregts, M. and de Jong, F. (2011). Robust speech/non-speech classification in heterogeneous multimedia content. *Speech Communication*, 53(2):143–153.

[Huijbregts et al., 2009] Huijbregts, M., Leuwen, D. A. v., and Jong, F. M. G. d. (2009). Speech overlap detection in a two-pass speaker diarization system. In *Interspeech*, pages 1063–1066, Brighton, United Kingdom.

[Huijbregts et al., 2012] Huijbregts, M., van Leeuwen, D., and Wooters, C. (2012). Speaker diarization error analysis using oracle components. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):393–403.

[Huijbregts and Wooters, 2007] Huijbregts, M. and Wooters, C. (2007). The blame game: Performance analysis of speaker diarization system components. In *Interspeech*, pages 1857–1860, Antwerp, Belgium.

[Istrate et al., 2005] Istrate, D., Fredouille, C., Meignier, S., Besacier, L., and Bonastre, J. F. (2005). RT05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings. In *NIST 2005 Spring Rich Transcrition Evaluation Workshop*.

[Janin et al., 2003] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The icsi meeting corpus. In *ICASSP*, pages 364–367, Hong Kong.

[Japkowicz et al., 2000] Japkowicz, N., Jose Hanson, S., and Gluck, M. A. (2000). Nonlinear autoassociation is not equivalent to pca. *Neural Comput.*, 12(3):531–545.

[Jayagopi and Gatica-Perez, 2010] Jayagopi, D. and Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *Multimedia, IEEE Transactions on*, 12(8):790–802.

[Jayagopi et al., 2009a] Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2009a). Modeling dominance in group conversations using nonverbal activity cues. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(3):501–513.

[Jayagopi et al., 2009b] Jayagopi, D., Raducanu, B., and Gatica-Perez, D. (2009b). Characterizing conversational group dynamics using nonverbal behaviour. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 370–373.

[Jin et al., 2004] Jin, Q., Laskowski, K., Schultz, T., and Waibel, A. (2004). Speaker segmentation and clustering in meetings. In *8th International Conference on Spoken Language Processing, Jeju Island, Korea*.

[Junqua et al., 1994] Junqua, J.-C., Mak, B., and Reaves, B. (1994). A robust algorithm for word boundary detection in the presence of noise. *Speech and Audio Processing, IEEE Transactions on*, 2(3):406–412.

[Kennedy and Ellis, 2004] Kennedy, L. and Ellis, D. P. W. (2004). Laughter detection in meetings. In *Proc. NIST Meeting Recognition Workshop*, pages 118–121, Montreal.

[Kenny et al., 2008] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):980–988.

[Kenny et al., 2010] Kenny, P., Reynolds, D., and Castaldo, F. (2010). Diarization of telephone conversations using factor analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 4(6):1059–1070.

[Kingsbury et al., 1998] Kingsbury, B., Morgan, N., and Greenberg, S. (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1–3):117–132.

[Knapp and Carter, 1976] Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327.

[Knox et al., 2012] Knox, M. T., Mirghafori, N., and Friedland, G. (2012). Where did I go wrong?: Identifying troublesome segments for speaker diarization systems. In *Interspeech*, Portland, USA.

[Konig et al., 1998] Konig, Y., Heck, L., Weintraub, M., Sonmez, K., and E, R. E. (1998). Nonlinear discriminant feature extraction for robust text-independent speaker recognition. In *RLA2CESCA Speaker Recognition and its Commercial and Forensic Applications*, pages 72–75.

[Kurtic et al., 2013] Kurtic, E., Brown, G. J., and Wells, B. (2013). Resources for turn competition in overlapping talk. *Speech Communication*, 55(5):721 – 743.

[Lamel et al., 1981] Lamel, L., Rabiner, L., Rosenberg, A., and Wilpon, J. (1981). An improved endpoint detector for isolated word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(4):777–785.

[Lapidot and Bonastre, 2012] Lapidot, I. and Bonastre, J.-F. (2012). Integration of LDA into a telephone conversation speaker diarization system. In *Electrical Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–4.

[Laskowski, 2010] Laskowski, K. (2010). Modeling norms of turn-taking in multi-party conversation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 999–1008, Uppsala, Sweden. Association for Computational Linguistics.

[Laskowski and Burger, 2007] Laskowski, K. and Burger, S. (2007). Analysis of the occurrence of laughter in meetings. In *Interspeech*, pages 1258–1261.

[Laskowski et al., 2007] Laskowski, K., Osterdorf, M., and Schultz, T. (2007). Modeling vocal interaction for text-independent classification of conversation type. In *8th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 194–201, Antwerpen, Belgium.

[Laskowski et al., 2008] Laskowski, K., Osterdorf, M., and Schultz, T. (2008). Modeling vocal interaction for text-independent participant characterization in multi-party convrersation. In *9th ISCA/ACL SIGdial*, pages 148–155, Columbus, USA.

[Laskowski and Schultz, 2006] Laskowski, K. and Schultz, T. (2006). Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *ICASSP*, pages 993–996, Toulouse, France.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Lu et al., 2002] Lu, L., Zhang, H.-J., and Jiang, H. (2002). Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10(7):504–516.

[luc Gauvain et al., 1998] luc Gauvain, J., Lamel, L., and Adda, G. (1998). Partitioning and transcription of broadcast news data. In *ICSLP'98*, pages 1335–1338.

[Mccowan et al., 2005] Mccowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*.

[Meignier et al., 2001] Meignier, S., Bonastre, J. F., and Igounet, S. (2001). E-HMM approach for learning and adapting sound models for speaker indexing. In *Odyssey Speaker and Lang. Recognition Workshop*, pages 175–180, Chania, Creete.

[Mirghafori and Wooters, 2006] Mirghafori, N. and Wooters, C. (2006). Nuts and flakes: A study of data characteristics in speaker diarization. In *IEEE ICASSP*, pages 1017–1020.

[Misra et al., 2003] Misra, H., Bourlard, H., and Tyagi, V. (2003). New entropy based combination rules in hmm/ann multi-stream asr. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, pages 741–744.

[Moraru et al., 2003] Moraru, D., Meignier, S., Besacier, L., Bonastre, J.-F., and Magrin-Chagnolleau, I. (2003). The ELISA consortium approaches in speaker segmentation during the nist 2002 speaker recognition evaluation. In *Acoustics, Speech, and Signal Processing,*

2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, volume 2, pages II–89–92 vol.2.

[Moraru et al., 2004] Moraru, D., Meignier, S., Fredouille, C., Besacier, L., and Bonastre, J.-F. (2004). The ELISA consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, pages I–373–6 vol.1.

[Morgan et al., 2005] Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cretin, O., Bourlard, H., and Athineos, M. (2005). Pushing the envelope - aside [speech recognition]. *Signal Processing Magazine, IEEE*, 22(5):81–88.

[Mostefa et al., 2007] Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S., Tyagi, A., Casas, J., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., and Rochet., C. (2007). The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. In *Language resources and evaluation*, pages 389–407.

[Motlicek et al., 2013] Motlicek, P., Povey, D., and Karafiat, M. (2013). Feature and score level combination of subspace gaussians in lvcsr task. In *ICASSP*, pages 7604–7608, Vancouver, Canada.

[National Institute of Standards and Technology, 2003] National Institute of Standards and Technology (2003). Rich transcription. http://nist.gov/speech/tests/rt/.

[Nguyen et al., 2002] Nguyen, P., Rigazio, L., Moh, Y., and Junqua, J.-C. (2002). Rich transcription 2002 site report panasonic speech technology laboratory (PSTL). In *2002 Rich Transcription Workshop*.

[Nwe et al., 2009] Nwe, T. L., Sun, H., Li, H., and Rahardja, S. (2009). Speaker diarization in meeting audio. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:4073–4076.

[Otterson, 2008] Otterson, S. (2008). *Use of Speaker Location Features in Meeting Diarization*. PhD thesis, University of Washington, Seattle.

[Otterson and Ostendorf, 2007] Otterson, S. and Ostendorf, M. (2007). Efficient use of overlap information in speaker diarization. In *ASRU*, Kyoto, Japan.

[Pardo et al., 2007] Pardo, J., Anguera, X., and Wooters, C. (2007). Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Trans. Comput.*, 56:1189–1224.

[Pardo et al., 2006] Pardo, J. M., Anguera, X., and Wooters, C. (2006). Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. In *ICSLP*, pages 2194–2197.

[Peleconas and Sridharan, 2001] Peleconas, J. and Sridharan, S. (2001). Feature warping for robust speaker verification. In *2001: A Speaker Odyssey-The Speaker Recognition Workshop*.

[Pfau et al., 2001] Pfau, T., Ellis, D., and Stolcke, A. (2001). Multispeaker speech activity detection for the icsi meeting recorder. In *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, pages 107–110.

[Rentzeperis et al., 2006] Rentzeperis, E., Stergiou, A., Boukis, C., Pnevmatikakis, A., and Polymenakos, L. C. (2006). The 2006 athens information technology speech activity detection and speaker diarization systems. In *Proceedings of the Third International Conference on Machine Learning for Multimodal Interaction*, MLMI'06, pages 385–395, Berlin, Heidelberg. Springer-Verlag.

[Sacks et al., 1974] Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). simplest semantics for the organization of the turn-taking in conversation. *Language*, 50:696–735.

[Schegloff, 2000] Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

[Seltzer et al., 2004] Seltzer, M., Raj, B., and Stern, R. (2004). Likelihood-maximizing beamforming for robust hands-free speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 12(5):489–498.

[Shriberg, 2005] Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Eurospeech*, pages 1781–1784.

[Shriberg et al., 2001] Shriberg, E., Stolcke, A., and Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Eurospeech*, pages 1359–1362, Aalborg, Denmark.

[Shum et al., 2011] Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D. A., and Glass, J. R. (2011). Exploiting intra-conversation variability for speaker diarization. In *Interspeech*, pages 945–948, Florence, Italy.

[Siegler et al., 1997] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *DARPA speech recognition workshop*, pages 97–99.

[Sinclair and King, 2013] Sinclair, M. and King, S. (2013). Where are the challenges in speaker diarization? In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7741–7745.

[Sinha et al., 2005] Sinha, R., Tranter, S. E., Gales, M. J. F., and Woodland, P. C. (2005). The cambridge university march 2005 speaker diarisation system. In *Interspeech*, Lisbon.

[Slonim et al., 1999] Slonim, N., Friedman, N., and Tishby, N. (1999). Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems*, pages 617–623. MIT press.

[Stolcke et al., 2000] Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373.

[Stolcke et al., 2007] Stolcke, A., Shriberg, E., Ferrer, L., Kajarekar, S., Sonmez, K., and Tur, G. (2007). Speech recognition as feature extraction for speaker recognition. In *Signal Processing Applications for Public Security and Forensics, 2007. SAFE '07. IEEE Workshop on*, pages 1–5.

[Sturim et al., 2002] Sturim, D., Reynolds, D., Dunn, R., and Quatieri, T. (2002). Speaker verification using text-constrained gaussian mixture models. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages 677–680, Orlando, USA.

[Sun et al., 2009] Sun, H., Nwe, T. L., Ma, B., and Li, H. (2009). Speaker diarization for meeting room audio. In *Interspeech*, Brighton, UK.

[Temko et al., 2007] Temko, A., Macho, D., and Nadeu, C. (2007). Enhanced svm training for robust speech activity detection. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1025–IV–1028.

[Tranter and Reynolds, 2006] Tranter, S. and Reynolds, D. (2006). An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565.

[Tranter, 2005] Tranter, S. E. (2005). Two-way cluster voting to improve speaker diarisation performance. In *ICASSP*, pages 753–756.

[van Leeuwen and Huijbregts, 2006] van Leeuwen, D. and Huijbregts, M. (2006). The AMI speaker diarization system for nist rt06s meeting data. In Renals, S., Bengio, S., and Fiscus, J., editors, *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 371–384. Springer Berlin Heidelberg.

[van Leeuwen and Konecny, 2008] van Leeuwen, D. and Konecny, M. (2008). Progress in the AMIDA speaker diarization system for meeting data. In Stiefelhagen, R., Bowers, R., and Fiscus, J., editors, *Multimodal Technologies for Perception of Humans*, volume 4625 of *Lecture Notes in Computer Science*, pages 475–483. Springer Berlin Heidelberg.

[Vijayasenan, 2010] Vijayasenan, D. (2010). *An Information Theoretic Approach to Speaker Diarization of Meeting Recordings*. PhD thesis, Ecole polytechnique fédérale de Lausanne.

[Vijayasenan and Valente, 2012] Vijayasenan, D. and Valente, F. (2012). Diartk : An open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In *Proceedings of Interspeech*, Portland, USA.

[Vijayasenan et al., 2009] Vijayasenan, D., Valente, F., and Bourlard, H. (2009). An information theoretic approach to speaker diarization of meeting data. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(7):1382–1393.

[Vinyals and Friedland, 2008] Vinyals, O. and Friedland, G. (2008). Modulation spectrogram features for speaker diarization. In *Interspeech*, Brisbane, Australia.

[Vipperla et al., 2012] Vipperla, R., Geiger, J., Bozonnet, S., Wang, D., Evans, N., Schuller, B., and Rigoll, G. (2012). Speech overlap detection and attribution using convolutive non-negative sparse coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4181–4184.

[Vogt and Sridharan, 2008] Vogt, R. and Sridharan, S. (2008). Explicit modelling of session variability for speaker verification. *Comput. Speech Lang.*, 22(1):17–38.

[Weber et al., 2000] Weber, F., Peskin, B., Newman, M., Corrada-Emmanuel, A., and Gillick, L. (2000). Speaker recognition on single- and multispeaker data. *Digital Signal Processing*, 10(13):75–92.

[Wiener, 1949] Wiener, N. (1949). Extrapolation, interpolation, and smoothing of stationary time series. In *Wiley*, New York.

[Woelfel and McDonough, 2009] Woelfel, M. and McDonough, J., editors (2009). *Distant Speech Recognition*. Wiley, New York.

[Wooters et al., 2004] Wooters, C., Fung, J., Peskin, B., and Anguera, X. (2004). Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In *Fall 2004 Rich Transcription workshop (RT04)*, Palisades, NY, USA.

[Wooters and Huijbregts, 2008] Wooters, C. and Huijbregts, M. (2008). The ICSI RT07s speaker diarization system. In *Multimodal Technologies for Perception of Humans*, pages 509–519. Springer-Verlag, Berlin, Heidelberg.

[Wrede et al., 2005] Wrede, B., Bhagat, S., Dhillon, R., and Shriberg, E. (2005). Meeting recorder project: Hot spot labeling guide. In *ICSI Technical Report TR-05-004*.

[Wrigley et al., 2005] Wrigley, S., Brown, G., Wan, V., and Renals, S. (2005). Speech and crosstalk detection in multichannel audio. *Speech and Audio Processing, IEEE Transactions on*, 13(1):84–91.

[Yella and Bourlard, 2014a] Yella, S. and Bourlard, H. (2014a). Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):1688–1700.

[Yella and Bourlard, 2013] Yella, S. H. and Bourlard, H. (2013). Improved overlap speech diarization of meeting recordings using long-term conversational features. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vancouver, Canada.

[Yella and Bourlard, 2014b] Yella, S. H. and Bourlard, H. (2014b). Information bottleneck based speaker diarization of meetings using non-speech as side information. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Florence, Italy.

[Yella et al., 2014a] Yella, S. H., Motlicek, P., and Bourlard, H. (2014a). Phoneme background model for information bottleneck based speaker diarization. In *Interspeech*, Singapore.

[Yella et al., 2014b] Yella, S. H., Stolcke, A., and Slaney, M. (2014b). Artificial neural network features for speaker diarization. In *IEEE Spoken Language Technologies Workshop*, South Lake Tahoe, USA.

[Yella and Valente, 2011] Yella, S. H. and Valente, F. (2011). Information bottleneck features for hmm/gmm speaker diarization of meetings recordings. In *Interspeech*, pages 953–956, Florence, Italy.

[Yella and Valente, 2012] Yella, S. H. and Valente, F. (2012). Speaker diarization of overlapping speech based on silence distribution in meeting recordings. In *Interspeech*, Portland, USA.

[Zelenák, 2011] Zelenák, M. (2011). *Detection and Handling of Overlapping Speech for Speaker Diarization*. PhD thesis, Universitat Politecnica de Catalunya, Barcelona, Spain.

[Zelenák and Hernando, 2011] Zelenák, M. and Hernando, J. (2011). The detection of overlapping speech with prosodic features for speaker diarization. In *Interspeech*, pages 1041–1043, Florence, Italy.

[Zelenák et al., 2010] Zelenák, M., Segura, C., and Hernando, J. (2010). Overlap detection for speaker diarization by fusing spectral and spatial features. In *Interspeech*, pages 2302–2305, Makuhari, Japan.

[Zelenák et al., 2012] Zelenák, M., Segura, C., Luque, J., and Hernando, J. (2012). Simultaneous speech detection with spatial features for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):436–446.

[Zhu et al., 2006] Zhu, X., Barras, C., Lamel, L., and Gauvain, J. (2006). Speaker diarization: From broadcast news to lectures. In *Machine Learning for Multimodal Interaction*, pages 396–406.

[Zhu et al., 2008] Zhu, X., Barras, C., Lamel, L., and Gauvain, J.-L. (2008). Multimodal technologies for perception of humans. chapter Multi-stage Speaker Diarization for Conference and Lecture Meetings, pages 533–542. Springer-Verlag, Berlin, Heidelberg.

[Zhu et al., 2005] Zhu, X., Barras, C., Meignier, S., and Gauvain, J. (2005). Combining speaker identification and bic for speaker diarization. In *Ninth European Conference on Speech Communication and Technology*.

[Zibert et al., 2009] Zibert, J., Brodnik, A., and Mihelic, F. (2009). An adaptive bic approach for robust speaker change detection in continuous audio streams. In Matousek, V. and

Mautner, P., editors, *Text, Speech and Dialogue*, volume 5729 of *Lecture Notes in Computer Science*, pages 202–209. Springer Berlin Heidelberg.

[Zochová and Radová, 2005] Zochová, P. and Radová, V. (2005). Modified DISTBIC algorithm for speaker change detection. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 3073–3076.

Sree Harsha Yella

303-3, Idiap Research Institute,
Rue Marconi 19, Martigny, Switzerland 1920.
Telephone(office): (+41) 277217798
E-mail: sree.yella@idiap.ch

EDUCATION

PhD (October 2010 - January 2015)
- Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

B Tech. + MS (by Research) dual degree (July 2004 - August 2010)
- Major: Computer Science and Engineering
- International Institute of Information Technology, Hyderabad, Andhra Pradesh, India.

RESEARCH INTERESTS

Speech signal processing, Statistical methods for speech processing, Information retrieval, Applied machine learning.

WORK EXPERIENCE

- Research Intern: Microsoft Research, Moutain View, CA, USA (Apr-June, 2014).
- Research Intern: Telefonica Research, Barcelona, Spain (June-August, 2013).
- Project Associate: Language Technologies Research Center (LTRC), IIIT Hyderabad. (April 2007 - December 2009).
- Teaching Assistant for Linear Programming course, Spring 2008.

PUBLICATIONS

- 'Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations', Sree Harsha Yella and Hervé Bourlard, in IEEE/ACM Transactions on Audio, Speech and Language Processing, Dec, 2014.
- 'Artificial neural network features for speaker diarization', Sree Harsha Yella, Andreas Stolcke and Malcolm Slaney, in IEEE SLT workshop, South Lake Tahoe, Nevada, USA, 2014.
- 'Phoneme Background Model for Information Bottleneck based Speaker Diarization', Sree Harsha Yella, Petr Motlicek and Hervé Bourlard, in Interspeech, Singapore, 2014.
- 'Information Bottleneck based Speaker Diarization of Meetings using Non-speech as Side Information', Sree Harsha Yella and Hervé Bourlard, in ICASSP, Florence, Italy, 2014.
- 'Inferring social relationships in a phone call from a single party's speech', Sree Harsha Yella, Xavier Anguera and Jordi Luque, in ICASSP, Florence, Italy, 2014.
- 'Improved Overlap Speech Diarization of Meeting Recordings using Long-term Conversational Features', Sree Harsha Yella and Hervé Bourlard, in ICASSP, Vancouver, Canada, 2013.
- 'Speaker diarization of overlapping speech based on silence distribution in meeting recordings', Sree Harsha Yella and Fabio Valente, in Interspeech, Portland, Oregon, USA, 2012.
- 'Information Bottleneck Features for HMM/GMM Speaker Diarization of Meetings Recordings', Sree Harsha Yella and Fabio Valente, in Interspeech, Florence, Italy, 2011 **Short-listed for best student paper award**.
- 'Automatic detection of conflict escalation in spoken conversations', Samuel Kim, Sree Harsha Yella and Fabio Valente, in Interspeech, Portland, Oregon, USA, 2012.

- 'Understanding Social Signals in Multi-party Conversations: Automatic Recognition of Socio-Emotional Roles in the AMI Meeting Corpus', Fabio Valente, Alessandro Vinciarelli, Sree Harsha Yella and A. Sapru, in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2011.
- 'Significance of Anchor Speaker Segments for Constructing Extractive Audio Summaries of Broadcast News', Sree Harsha Yella, Vasudeva Varma and Kishore Prahallad, in IEEE SLT workshop, 2010, Berkeley, California, USA.
- 'Prominence based scoring of speech segments for automatic speech-to-speech summarization', Sree Harsha Yella, Vasudeva Varma and Kishore Prahallad, in Interspeech, Makuhari, Japan, 2010.
- 'Statistical Transliteration for Cross Language Information Retrieval using HMM alignment model and CRF', Prasad Pingali, Surya Ganesh, Sree Harsha Yella and Vasudeva Varma, in *P*roc. 2nd workshop on Cross Lingual Information Access (CLIA), 2008.
