Automatic social role recognition and its application in structuring multiparty interactions

THÈSE N^{O} 6642 (2015)

PRÉSENTÉE LE 29 JUIN 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

Ashtosh SAPRU

acceptée sur proposition du jury:

Prof. D. Atienza Alonso, président du jury
Prof. H. Bourlard, directeur de thèse
Prof. F. Pianesi, rapporteur
Dr J.-M. Vesin, rapporteur
Dr A. Vinciarelli, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

To my parents…

Acknowledgements

The four years of my PHD have come to an end. The work involved in completion of this thesis has been one of the most informative and challenging experience of my academic life. Without the guidance, expertise and encouragement of advisers, family and friends my PHD would not have been possible.

First and foremost, I would like to thank my advisers Hervé Bourlard and Fabio Valente for giving me an opportunity to pursue research at Idiap. Hervé pushed me to set and achieve challenging goals in my research and I could not have finished this PHD without his support and encouragement. His eye for detail and constructive criticism has improved the content and quality of this thesis. I am also thankful to Fabio for getting me started on this research. Without Fabio I would not have known where to begin. His expertise and knowledge were very helpful during the initial part of my PHD.

I thank my committee members Prof. Fabio Pianesi, Dr. Alessandro Vinciarelli, Dr. Jean-Marc Vesin and Prof. David Atienza Alonso for their valuable comments on the thesis. I am grateful to the committee members for taking time from their schedule to review this thesis. I would also like to thank Sree Harsha and Samuel Kim for their collaboration and technical discussions of my work. I thank Mathew for his support and several insightful comments of my work which helped in improving presentation of this thesis. I owe special thanks to David Imseng and Raphael Ullmann for their help in translating the English abstract of this thesis into German. Working in speech group has broadened my understanding and I am thankful to researchers past and present for their feedback. My gratitude to Hasler Foundation for funding my PHD through SESAME grant.

Idiap is a great place to work and none of my experiments would have yielded any results without the excellent computational and storage infrastructure provided here. I thank the system administrators at Idiap, Louie-Marie, Frank, Bastien and Norbert for smooth running of computational resources and their constant help in fixing the system related issues during my PHD. I would also like to thank the administrative staff at Idiap. In particular, thanks to Nadine for her help in finding a suitable accommodation during my stay in Martigny and Sylvie for her help with work permit, insurance and visa. Thanks also to Corinne, Chantal, Claude and everyone at EPFL who helped in the organization of my defense.

Besides my work at Idiap, I also had lot of fun. Thanks to Sree Harsha, David, Blaise, Ramya,

Dinesh, Hari, Anindhya, Venky, Deepu, Gokul, Jagan, Darshan, Chidhansh, Murali, Lakshmi, Saheer, Mathew, Francina, Srikanth, Pranay, Marc, Sriram, Anirudha, Subhadeep, Dhananjay, Raphael, Afsaneh, Mohammad, Pierre-Edouard, Sucheta, Alexandros, Phil, Petr, Daira, Dimitri, James, Laurent, Samuel and many others for making my stay at Idiap a memorable experience.

Finally, I want to thank my family and friends in India for their understanding and support. Most of all, I want to thank my parents, Ashok and Pushpa, for their sacrifices, care and unconditional love without which I would not have reached this place in my life.

Martigny, 28th May, 2015 Ashtosh Sapru.

Abstract

Automatic processing of multiparty interactions is a research domain with important applications in content browsing, summarization and information retrieval. In recent years, several works have been devoted to find regular patterns which speakers exhibit in a multiparty interaction also known as social roles. Most of the research in literature has generally focused on recognition of scenario specific formal roles. More recently, role coding schemes based on informal social roles have been proposed in literature, defining roles based on the behavior speakers have in the functioning of a small group interaction. Informal social roles represent a flexible classification scheme that can generalize across different scenarios of multiparty interaction. In this thesis, we focus on automatic recognition of informal social roles and exploit the influence of informal social roles on speaker behavior for structuring multiparty interactions.

To model speaker behavior, we systematically explore various verbal and non verbal cues extracted from turn taking patterns, vocal expression and linguistic style. The influence of social roles on the behavior cues exhibited by a speaker is modeled using a discriminative approach based on conditional random fields. Experiments performed on several hours of meeting data reveal that classification using conditional random fields improves the role recognition performance. We demonstrate the effectiveness of our approach by evaluating it on previously unseen scenarios of multiparty interaction. Furthermore, we also consider whether formal roles and informal roles can be automatically predicted by the same verbal and nonverbal features.

We exploit the influence of social roles on turn taking patterns to improve speaker diarization under distant microphone condition. Our work extends the Hidden Markov model (HMM)-Gaussian mixture model (GMM) speaker diarization system, and is based on jointly estimating both the speaker segmentation and social roles in an audio recording. We modify the minimum duration constraint in HMM-GMM diarization system by using role information to model the expected duration of speaker's turn. We also use social role n-grams as prior information to model speaker interaction patterns. Finally, we demonstrate the application of social roles for the problem of topic segmentation in meetings. We exploit our findings that social roles can dynamically change in conversations and use this information to predict topic changes in meetings. We also present an unsupervised method for topic segmentation which combines social roles and lexical cohesion. Experimental results show that social roles improve performance of both speaker diarization and topic segmentation.

Keywords: Multiparty Interactions, Social Roles, Formal Roles, Meetings, Turn Taking, Speaker

Diarization, Topic Segmentation, Conditional Random Fields, Hidden Markov Model, Latent Dirichlet Allocation, Distance Dependent Chinese Restaurant Process

Zusammenfassung

Die automatische Verarbeitung von Gruppeninteraktionen ist ein Forschungsbereich mit wichtigen Anwendungsmöglichkeiten wie Content-Browsing, Indexierung von Daten, Zusammenfassungen, und Informationsabruf. In den letzten Jahren waren verschiedene Arbeiten dem Auffinden von regulären Verhaltensmustern von Sprechern in einer Gruppeninteraktion, auch Rollen genannt, gewidmet. Die meisten dieser Forschungsarbeiten bezogen sich auf die Erkennung von formalen, Szenario-spezifischen Rollen. Vor kurzem wurden Rollenkodierungsschemas gestützt auf informalen sozialen Rollen untersucht. Diese Rollen basieren auf dem Verhalten von Sprechern in einer kleinen Gruppe. Informale soziale Rollen bilden ein flexibles Klassifikationsschema, welches sich auf mehrere verschiedene Gruppeninteraktionsszenarien verallgemeinern lässt. Diese Dissertation befasst sich mit der automatischen Erkennung von informalen sozialen Rollen und nutzt ihren Einfluss auf das Sprecherverhalten um die Gruppeninteraktionen zu strukturieren.

Um das Sprecherverhalten zu modellieren, erforschen wir verschiedene verbale und nonverbale Hinweise, welche von Sprecheraktivitätsmustern, vokalen Ausdrucksformen und Sprachstilen extrahiert werden. Der Einfluss von sozialen Rollen auf das Sprecherverhalten wird dabei mit einer diskriminativen Methode basierend auf Conditional Random Fields modelliert. Experimente mit mehreren Stunden von Sitzungsdaten zeigen, dass die Klassifikation mittels Conditional Random Fields die Rollenerkennung verbessert. Wir belegen die Wirksamkeit unserer Methode durch eine Evaluierung auf unbekannten Gruppeninteraktionsszenarien. Wir untersuchen auch, ob das Rollenerkennungssystem gleichermassen für die Vorhersage von formalen und informalen Rollen verwendet werden kann.

Wir nutzen den Einfluss von sozialen Rollen auf Sprecheraktivitätsmuster um die Sprechersegmentierung von Aufnahmen aus grösserer Distanz zu verbessern. Diese Dissertation erweitert den Hidden Markov-Modell (HMM) – Gausssche Mischverteilungs-Modell (GMM) Ansatz zur Sprechersegmentierung und beruht auf der gemeinsamen Schätzung von Sprechersegmenten und sozialen Rollen in Audioaufnahmen. Wir verändern die Bedingung der Minimaldauer in HMM-GMM basierten Sprechersegmentierungssystemen, indem wir mit Hilfe der Rolleninformation die Sprechzeit des Sprechers zu schätzen. Wir benutzen auch n-Gramme von sozialen Rollen als a-priori Information um Sprecherinteraktionsmuster zu modellieren. Wir zeigen auch auf, wie soziale Rollen für die Themensegmentierung von Sitzungen benutzt werden können. Dabei benutzen wir die vorhergesagten sozialen Rollen als charakteristische Merkmale in einer überwachten Klassifizierung und zeigen, dass soziale Rollen Themenwechsel in Konversationen erfassen. Wir führen auch einen neuen Modellierungsrahmen zur Themenseg-

mentierung ein, welcher Informationen über soziale Rollen mit latenter Themenmodellierung kombiniert. Experimente zeigen, dass die Berücksichtigung von sozialen Rollen die Leistung von Systemen zur Sprechersegmentierung als auch Themensegmentierung verbessert.
Schlüsselwörter: Gruppeninteraktionen, soziale Rollen, formale Rollen, Sitzungen, Sprecheraktivitätsmuster, Sprechersegmentierung, Themensegmentierung, Conditional Random Fields, Hidden Markov Modell, Latent Dirichlet Allocation, Distance Dependent Chinese Restaurant Process.

Contents

List of Figures

List of Tables

1 Introduction

One of the most fundamental aspects of our lives is the fact that we spent considerable amount of time and energy in social interaction with other human beings. Our constant engagement with others keeps us aware of our surroundings and helps us comprehend the world in which we live. Roles form an important concept in understanding human social interactions. The activities involved in our daily life can be viewed as a consequence of different roles we assume, and the role playing mechanism is even imitated by children when they pretend at being adults [69]. The concept of roles has been studied extensively in social psychology, and roles have been used to explain a range of phenomena like gender differences, status, leadership and social position.

In small group multiparty interactions, roles can be broadly categorized as formal and informal [46]. Formal role is a designated position that is directly assigned by an organization or a group. Designations such as chairperson and secretary are examples of formal roles. In comparison to formal roles, informal social roles are not designated as positions in a group. Informal social roles naturally emerge as a result of interactions between group members. These roles emphasize functions that usually assist the group in accomplishing its goals [8, 14, 105].

The study of social roles in small groups is an important area of research in social psychology and several studies have shown that participants exchange information with each other, through both verbal and nonverbal communication [8, 57, 40]. Effective communication requires that participants alternate between listening and speaking states and organize the conversation by taking turns [90]. Natural language is a fundamental mechanism to represent the semantic content of speech and is frequently used by group participants to communicate task related goals [8]. Participants also display non verbal behavior characteristics through vocal expression, body and facial gestures and language style [57, 40].

In recent years social computing has emerged as a key area of research for automatic analysis of social interactions. There are various issues that need to be considered while applying computational techniques for analysis of multiparty interactions, including reliable data annotation in absence of ground truth; feature extraction using standard tools like automatic

speech recognition, speaker diarization and prosody extraction; and computationally efficient models that combine those features. Several recent studies have applied computation models to various phenomena studied in multiparty interactions like dominance, roles, engagement and hot-spots [127, 54]. Our work complements the existing research, focusing on automatic recognition of social roles and its application in structuring multiparty interactions.

In the context of this thesis, we interchangeably use the terms social roles, emergent roles, informal roles to refer informal social roles which emerge naturally in small group meetings.

1.1 Objectives of the thesis

The principal objective of this thesis is to investigate and design computational models for analysis of multiparty interactions. This is a challenging area and requires integration of knowledge from diverse research areas, including social psychology, signal processing and machine learning. Our analysis is performed in the context of small group meetings and we investigate multiple verbal and non verbal cues for modeling speaker behavior. We consider various problems in multiparty conversations, such as speaker segmentation, social roles of interacting participants, and discourse segmentation of conversation into different topics. Our primary focus are the social roles exhibited by participants in meetings. We explore various feature groups that can be extracted automatically and develop role recognition system for joint modeling of those features. Furthermore, by exploring the influence of social roles on speaker behavior, we aim to improve the performance of current state-of-the-art systems for automatic analysis of multiparty interactions. We explore the application of social role information to improve both short segmentation of meeting into different speakers and long term discourse segmentation into different topics. As social roles emerge naturally in multiparty interactions, we expect the improvements to generalize across different scenarios of interaction.

1.2 Motivations for the thesis

Advances in multimedia compression and digital storage technologies have resulted in several archives of multiparty interactions in a variety of domains, including television and radio broadcast news, lectures and meetings. Majority of research for analyzing multiparty interactions has focused on standard domains, such as television and radio broadcast news and telephone conversations. Broadcast news represent a special category of social interaction, in which planned and well prepared conversation is moderated by professional speakers. In comparison, meetings are a form of social interaction in which groups of humans can spontaneously interact and exchange information in order to accomplish a common goal. In recent years, there has been an upsurge of interest in automatic analysis of meetings and large projects were established to record human behavior in meetings [23, 19, 53, 21].

Meetings have been actively studied in social psychology and a large body of research has

focused on multiparty interactions in small group meetings [8, 68]. An important approach for analysis of meetings is to identify regular and predictable pattern of human behavior, embodied in the concept of role. "People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability" [111]. Status and the related concept of formal role can change depending on the scenario of interaction, e.g., professional meeting, faculty meeting or informal discussions. In comparison, informal social roles emerge naturally in meetings and characterize the way participants interact with each other. This suggests that a system for automatic recognition of informal social roles can possibly generalize across different scenarios of multiparty interaction.

The principal application of this thesis is for problem of organizing and indexing multimedia content from audio recordings of meetings. As the size of multimedia archives grows, it becomes a very challenging task for a user to retrieve relevant information. On the other hand, prior research reveals that information extracted from meetings can be used to improve future plans and actions [76]. In [11], it was shown that meeting browsers annotated with speaker roles and topic segments were very effective for answering user queries. Role information can also be used to segment topically homogeneous segments in conversation discourses [121] and summarization of spoken documents [120]. Another motivation for our work is its application in social psychology. Analysis of small group interactions, extensively studied in social psychology typically relies on human observers for coding speaker behavior. However, manual annotation is expensive in terms of both cost and time. In this context, computational models can automate the process of cue extraction and behavior modeling. Social scientists can benefit by using computational tools to process large amount of data recorded in a natural environment.

1.3 Contributions of the thesis

The contributions of this thesis can be summarized as follows:

- Automatic role recognition in multiparty interactions
  - We systematically investigate various verbal and non verbal features extracted from turn taking patterns, vocal expression and linguistic style for predicting informal social roles that emerge in small group interactions. We consider various feature groups individually and in combination, to understand the relative influence of each and the benefits of using them jointly. The framework of conditional random fields (CRF) is extended to develop a automatic role recognition model, which integrates features extracted at multiple time scales in a single representation. The classification model based on CRF offers the benefits of discriminative learning and flexibility to include multiple non-independent features. Experimental results

on several hours of AMI corpus meetings demonstrate that social role recognition improves by combination of verbal and nonverbal features. We also compare the model against standard methods like support vector machines (SVM) and show that CRF achieves a higher role recognition accuracy. Parts of our contribution have appeared in [93, 94] and journal article [95].

  - Previous studies in literature have investigated recognition of formal roles in meetings, however no study has systematically compared the relation between features, formal roles and informal roles on the same set of meetings. We demonstrate that, while prediction of informal roles improves from combination of verbal and nonverbal features, verbal features are best predictors of formal roles. Furthermore, we also show that latent topic models can be applied to speaker utterances to automatically infer formal roles of speakers. Parts of our results have appeared in [96, 98].

- Social role based speaker diarization under distant microphone condition

  - Speaker diarization is the task of identifying "who spoke when" in a multiparty interaction. This is a challenging task as the number of speakers and their associated speaking times are initially unknown. Our analysis shows that social roles influence the turn taking patterns of speakers. We exploit our findings to improve current state-of-the-art Hidden Markov model (HMM)- Gaussian mixture model (GMM) speaker diarization system. In particular, we focus on two limitations of HMM-GMM system, i.e, a speaker independent minimum duration constraint and a uniform prior on speaker interaction patterns. Our work extends the HMM-GMM speaker diarization system, by including social roles as prior information in the speaker segmentation step. Experiments on several meeting corpora demonstrate that social role information improves the performance of HMM-GMM speaker diarization system. This work was published in [97].

- Application of social roles for topic segmentation in meetings

  - Topic segmentation consists of dividing a multiparty interaction into several locally coherent topic segments. Our analysis demonstrates that social roles of participants can dynamically change in a meeting. We exploit this finding to relate changes in social roles of participants with shifts in conversation topics. We apply the automatic role recognition system to estimate the social role posterior probabilities of multiple speakers. We train a supervised classifier (Boosting) on role posteriors to demonstrate the relevance of social roles for topic segmentation. We also develop and test an unsupervised topic segmentation method based on distance dependent Chinese Restaurant Process which combines social roles and latent topic models. Experimental evaluation of both supervised and unsupervised methods show the effectiveness of social roles for topic segmentation in meetings.

1.4 Organization of the thesis

The thesis is organized as follows:

- Chapter 2: We review the related work for analysis of multiparty interactions using automatic systems. The related work includes sections on role recognition, topic segmentation and speaker diarization. We then present various publicly available corpora (AMI, ICSI) of small group meetings that will be used for experimental evaluation.
- Chapter 3: This chapter presents the corpus and data annotation for study of social roles in meetings. We describe various verbal and nonverbal features that are automatically extracted from turn taking, acoustic and linguistic behavior of speakers. We then present a supervised classifier based on conditional random fields for automatically recognizing the social roles of speakers from the extracted features.
- Chapter 4: We consider automatic recognition of formal roles in meetings. We compare recognition of social roles and formal roles using the same set of features and over same set of meetings. Using experimental results we show that nonverbal features are weakly influenced by formal roles. We then present an unsupervised feature extraction method to predict the formal roles of speakers from the verbal content of their speech.
- Chapter 5: In this chapter, we exploit the influence of social roles on turn taking patterns of speakers to improve HMM-GMM speaker diarization system. We present modifications to HMM-GMM diarization system and describe the use of social roles to incorporate prior information about speaker turn duration and speaker sequence distribution.
- Chapter 6: We demonstrate that social roles can be used to segment the audio recording of a meeting into different topics. We first train a supervised classifier using social role information and then apply the supervised classifier on meeting data to demonstrate the applicability of social roles for topic boundary detection. We also present an unsupervised method which combines social roles with latent topic models for topic segmentation.
- Chapter 7: In this chapter, we summarize the main contributions of this thesis and discusses future directions.

2 Related work

2.1 Introduction

Recent years have seen a lot of interest in automatic analysis of multiparty interactions. Several contributions have been made to structure unlabeled audio recordings. A survey of relevant literature reveals that most of the initial studies for labeling spoken conversations, including research on speaker diarization, role recognition and topic segmentation were developed for broadcast news recordings. However, as the technology in broadcast domain has started maturing, the focus of research community has shifted to more challenging case of meeting analysis. The main areas of research in meetings includes short term segmentation of audio into speaker homogeneous regions, identifying the roles of different speakers and segmentation of meeting into different topics. In literature, each of these areas has been considered as a separate research problem and the general architecture of current state-of-the-art systems ignores sharing of information between these problems. In this context, distinct set of approaches have evolved in literature on role recognition, topic segmentation and speaker diarization.

In this chapter, we present the related work in the context of automatic analysis of multiparty interactions. The chapter is organized as follows. In Section 2.2 and Section 2.3, we review the literature on roles in social psychology and social computing related to our work. Section 2.4 presents the related work for topic segmentation. In Section 2.5, we review the literature on speaker diarization. In Section 2.6, we present the existing publicly available corpora used in this work.

2.2 Roles in social psychology

The concept of social roles has been a subject of analysis for over 80 years [69]. In social psychology literature, roles have been defined as characteristic behavior patterns of one or more persons in a context [15]. According to [15], role theory presumes that "persons are members of social positions and hold expectations for their own behaviors and those of other

persons." The expectations are regarded as role generators and can be differentiated into three modalities: norms are prescriptive expectations, and express demands or requests of a person; beliefs are descriptive expectations, and represent opinions, assertions or social perceptions of a person; preferences express feelings, evaluations or values. All the three modes of expectation are responsible for role generation, and persons often conform to expectations that are held by others, are attributed to others, or are held by the person for his or her conduct [15].

From the viewpoint of this work, we are interested in informal roles that emerge naturally in small group interactions. These roles generalize across any type of multiparty interaction and are defined in terms of communicative functions that group members perform as they lead the group towards its goal.

In [14], authors formulated a list of functions that participants perform based on their observations of group interactions. They divided this list into three categories: (1) group task roles, (2) group maintenance roles, and (3) individual roles. The first category of roles focus on the set of tasks that the group members perform, and include roles such as the coordinator (coordination function for the group). Group maintenance roles focus on keeping the group together, and include roles such as the harmonizer (lessen discord in a group). Task and maintenance roles are positive function roles and help the group in reaching its goal. In contrast, individual roles are negative functional roles and participants assuming these roles attempt to satisfy their own needs and work against the groups needs. Examples include role of an aggressor. According to [14], successful groups follow a flexible role structure which allows same person with multiple talents to assume different roles.

According to [8, 105], decision making in small groups results in emergence of two specialized roles: one related to task needs of the group and other related to socio-emotional needs of the group. In [8], Bales presented a coding scheme of 12 functions that can be used to analyze the communications which occur during group meetings. Six of these functions are related to socio- emotional balance in the group. These functions can, in turn, be divided into positive reactions (solidarity, agreement, satisfaction), that are responsible for group cohesion and negative reactions (tension, disagreement, hostility), that endanger group cohesion. In general, this study suggests that satisfied groups have a greater proportion of positive statements as compared with negative statements. The other set of six functions are related to management and solution of problems that the group is addressing. These functions are also complementary, such that one set is responsible for asking suggestion, information and opinion and mirror set is responsible for giving suggestion, information and opinion.

2.3 Automatic role recognition

Previous research in social computing area can be broadly classified based on the domain of group interaction, i.e., roles in news broadcast and roles in spontaneous interactions. On broadcast data, speakers generally derive their roles by confirming to specific norms of

behavior. In comparison roles in spontaneous interactions mostly refer to positions in a social system, such as managers, designers, students etc.

2.3.1 Formal role recognition in broadcast domain

One of the first studies to investigate speaker roles in broadcast data was presented in [12]. This work considered the use of speaker role information for inferring the structural summary of broadcast news (BN). The news recordings were manually segmented into speaker boundaries and each segment was automatically labeled into one of three roles: Anchor, Journalist or Guest. The features used in this work were influenced by the structure of news program transcripts. Several features were extracted like signature phrases, explicit speaker introductions, duration of speaker segments and labels from surrounding segments. They reported an accuracy of 80.5% for role classification when features were extracted from manual transcripts and 77% when an ASR system was used. A similar study for segmentation of mandarin BN into three role labels was reported in [62]. Word N-grams were extracted from about 170 hours of speech data to train supervised classifiers. The authors compared two different classifiers, a Hidden Markov Model (HMM) and a maximum entropy model (Maxent). Interestingly, while both models reached a similar accuracy of 77%, the performance is different for individual roles. Maxent performs better in identifying reporters, compared to HMM. An improvement in accuracy, from 77% to 80%, was reported by combining the two models.

Recent studies [129, 24, 50] have also considered the BN roles on broadcast conversations (BC), such as talkshows. In [129], authors investigated role recognition on BC data using a Dynamic Bayesian network (DBN). Four categories of roles were considered: Host, Guest, Audience and Journalist. This contribution highlights the influence of speaking styles in broadcast conversations. They reported an accuracy of 77% for the HMM system. The second contribution was that the current role of a speaker is correlated with the role in immediate past. This information was modeled using a DBN system and the accuracy of the recognition system improved to 82%. More recently, in [24], authors proposed a set of novel features derived from word confidence measures in ASR generated transcripts to recognize three role categories: Anchor, Reporter and Other. They reported accuracy ranging from 88% on segments of pure speaker turns and 75% on turns with multiple speakers. In comparison to previous studies that are based on supervised classification, an unsupervised approach for role labeling was presented in [50]. Like most works in broadcast domain, three different roles were considered: Host, Guest and Soundbites. Several clustering algorithms were applied to a set of structural and lexical features and results reached an accuracy of 86% for role labeling task.

For the methods described above, role assignment is done at the level of speech turn. In [91], a speaker's role was predicted by considering its behavior for the entire length of the recording. Six different roles were considered in radio broadcast news: anchorman, second anchorman, guest, headline reader, weather man, and interview participant. This work leverages the fact that radio programs have a compact structure where a central speaker is usually in direct

interaction with other speakers. A social network for each speaker was constructed based on their immediate interaction with other speakers. Using a combination of social network analysis (SNA) and duration modeling, the authors report an accuracy of 85% in correctly labeling roles. SNA based approaches have also been applied to identify roles in movies and TV shows. In [122], leading roles, such as hero, heroine and their respective friends were identified based on co-occurrences of faces of individuals in the same scene.

One of the main limitations of SNA approach is that it requires a higher number of interacting participants (more than 8-10 persons), to build meaningful social networks. To avoid this limitation a modification of SNA approach was presented in [91]. Here instead of constructing speaker-speaker networks, affiliation networks are constructed based on temporal proximity of speakers. This method reached an accuracy of 86% for labeling six speaker roles in radio shows.

2.3.2 Formal role recognition in meetings

While most studies have explored recognition of formal roles in BN and meeting environments, there are many differences in the nature of data between the two domains. BN data is usually characterized by planned speech while meeting interactions have more spontaneous speech. Furthermore, speaker turn changes occur less frequently in BN data and average length of speaker turns is longer. In comparison, meeting interactions contain more overlapping speech and speaker turns are of shorter duration.

The study in [32] compared the performance of a HMM based automatic role recognition system on BN data and meeting recordings. The BN roles were the same as described in [91], while the meeting roles reflect the position of speakers in an organization. Four categories of formal roles wee considered: Project Manager, Marketing Executive, User Interface Designer and Industrial Designer. It was observed that the perplexity of the role sequence can be used as measure of role formality. Broadcast roles sequences have lower perplexity, which suggests that roles are more formal and speaker interaction is constrained by the program format. In comparison, meeting interactions do not impose explicit constraints on behavior of people, and these roles were harder to model. The recognition algorithm reached an accuracy of 86% for recognizing BN roles, while the accuracy was only 52% on meeting roles.

Several other studies have investigated formal role recognition in meetings and role categories in these studies are dependent on the scenario of interaction. In [10], the authors proposed a simple taxonomy of participant roles (presenter, information provider, participator and information consumer). Simple features like count of speaker changes, number of active meeting participants and overlap duration were computed within a meeting window. The window size was kept as a tunable parameter. Using decision tree classifiers, and a window size of 20 seconds, they reported the best accuracy of 53% for recognizing four speaker roles. Similar speech activity based features were extracted in [60] to recognize roles based on education level of participants (graduate,professor and PHD). They reported an accuracy of

61% for recognizing three speaker roles. Formal role recognition in professional meetings was investigated in [37]. The dataset and roles used in this study are same as described in [32]. Their analysis revealed that combination of verbal and nonverbal features significantly improve the accuracy of role recognition system to (68%) over the system which models only nonverbal information (44%).

In summary, most works on role recognition for BN data have exploited features derived from audio data to classify three main role categories: Anchor/Host, Reporter/Journalist and Guest/Other. The feature extraction is heavily influenced by the structure of broadcast format and both verbal and non verbal (SNA, structural) features have been used to achieve recognition accuracies in excess of 80%. In comparison, formal roles investigated in meetings are influenced by scenario of group interaction and role categories can change from corpus to corpus. Meeting data is also characterized by spontaneous conversation and recognition systems based on nonverbal information perform much lower in meetings compared to BN data. However, recognition systems which combine both nonverbal and verbal information perform significantly better than systems which rely only on nonverbal information.

2.3.3 Social role recognition in meetings

For the studies mentioned above, participants role was formal and considered to remain constant over the duration of entire audio recording. Formal roles are generated due to normative expectations of behavior or from positions in an organizational system. Informal social roles, as discussed in [8, 14] emerge naturally to serve needs of the group. All the studies discussed next, attribute to each participant in the group a role in between Protagonist, Supporter, Neutral, Gatekeeper or Attacker.

Social role recognition in problem solving sessions was considered in [133]. A support vector machine (SVM) classifier was used to discriminate between social roles using features expressing participants activity from both audio and video. They reported an accuracy above 65% for role recognition task. In [28], SVM, HMM, and influence model approaches were compared on the same dataset. In addition to audio and video activity features, speaking rate of participants were also extracted over multiple time windows. The authors use influence models to exploit constraints on the dynamics of social roles and report better performance compared to SVM and HMM models. However, an analysis of classification results revealed a wide difference between accuracy 80% and average recall 55%. This shows that, while the classifier performs well on highly populated roles, results are much worse on less populated roles.

Other studies [124, 115] have also investigated role recognition in professional meetings using the same social role coding scheme proposed in [133]. An HMM based approach was used to model turn statistics and prosody (fundamental frequency, energy) for role recognition in [115]. The authors report an accuracy of 59% for HMM model. This model was then extended to explicitly account for dependencies between speakers yielding an accuracy of 65%. In [124], speech activity features were combined with linguistic subjectivity and expressive prosodic

features for role recognition. There analysis revealed that, while the linguistic features and expressive prosodic features were informative for role recognition, feature combination did not result in a statistically significant improvement in performance for most roles. However, as the feature set used in this study was limited, a more extensive set of features might be more informative for role recognition task.

2.4 Topic segmentation

In multiparty interactions speakers assume different social roles, however, as the conversation progresses topics or stories also evolve. Topic segmentation aims at identifying the topic boundaries in a multiparty interaction. In the context of broadcast news, topics correspond to individual new stories or reports. In comparison, meetings involve spontaneous conversation between participants and display a hierarchical structure, in which coarse grained topics can be further divided into fine grained sub topics. Topic segmentation in meetings is challenging task even for human annotators, especially when fine grained sub topics are also considered [42]. Several previous studies in literature have approached the task of automatic topic segmentation. These approaches can be broadly grouped in three different categories that are described next.

2.4.1 Changes in lexical similarity

The earliest approaches for topic segmentation focused on identifying changes in lexical content. These approaches identifying regions of discourse which are marked by sudden change in vocabulary used by the participants. For each unit of discourse, e.g., turn in a conversation or sentence in a text, a lexical cohesion score is computed. The topic change points are identified by comparing the lexical cohesion score against a threshold that is determined automatically or by using heuristic rules.

Texttiling [47], is one of the earliest algorithms for automatic topic segmentation. The algorithm was initially proposed for locating topic shifts in scientific articles, however, it has also been applied to BN data. The algorithm divides the spoken document into groups of psuedosentences, and each psuedosentence contain a fixed number of words. A sliding window runs across the document and groups a fixed set of psuedosentences into blocks of text. For each window location, a lexical vector is calculated whose elements are the raw frequency values of a words in the text block. The lexical similarity for a pair of adjacent windows is calculated based on the cosine distance between their word frequency vectors. If the cosine distance is plotted for the entire document then the locations of minima in this graph can be considered as potential topic changes. Other works [107, 88], have extended the basic Texttilling algorithm by also considering word bigram statistics.

While lexical cohesion based methods are unsupervised and relatively faster to implement, these approaches ignore statistical dependencies between words that are related to the same

topic segment. Techniques, such as Latent Semantic Analysis (LSA) [59] have been applied in several text analysis tasks to model co-occurrence relationship between words. LSA is a word document matrix factorization technique, related to singular valued decomposition (SVD). The application of LSA to high dimensional sparse word frequency vectors maps them to low dimensional space. In [75], authors showed that Texttilling can be improved by measuring the lexical similarity in the latent vector space.

While Texttilling based approaches have been applied on BN domain, these approaches have not been as successful on meetings. LCseg algorithm proposed in [36] extends the idea of lexical cohesion for topic segmentation in meetings. However, instead of computing word term frequencies, the hypothesized topic shifts were determined by lexical chains. Lexical chains begin at the first occurrence of a word and end at its last occurrence and include all the word repetitions in between. Using tf-idf criterion, chains were weighted based on their word term frequency (frequent terms are weighted higher) and length (shorter chains receiving higher weights). Like Texttilling, a cosine distance between lexical chain vectors in adjacent windows was used as the distance measure to determine topic shifts. In comparison to other lexical cohesion approaches, LCseg has shown better performance for topic segmentation in meetings [36].

2.4.2 Generative methods

Instead of exploiting the low cohesion at topic shifts, generative models cluster neighboring discourse units belonging to the same topic segment. Generative models assume that a conversation can be expressed as a sequence of topic segments and each topic segment can be represented as a distinct probability distribution over words in the vocabulary. The segment boundaries are inferred by the positions where the vocabulary associated with a topic changes.

Hidden Markov Model(HMM) is one of the most commonly used generative models for sequence data. The states of a HMM can be used to represent the hidden topic segments and each state can be represented with a probability distribution from which words are generated. HMM also assumes that probability of current topic depends only on the previous topic. A topic change occurs during a transition from one topic segment to another, otherwise the model does a self looping transition. In [130], authors applied a HMM model to segment BN stories. Using a large training corpus (CNN news stories), a clustering algorithm was applied to cluster the BN data into a fixed number of topics. The emission probability of each hidden topic was represented using a ngram language model. Topic segmentation was evaluated on a separate testset by applying Viterbi algorithm. In comparison to Texttiling, HMM approach in [130] did not require an explicit distance measure. However, the parameters of the model were trained on a pre-segmented training BN dataset.

The main drawback of HMM approach is the assumption of conditional independence of words given the topic. More recent studies in literature have instead modeled the word co-occurrence statistics in a document. In latent topic modeling framework, a document, i.e,

a topic segment is represented as probability distribution over latent topics and each latent topic is represented as a probability distribution over words. One of the more popular latent topic model approach is based on Latent Dirichlet Allocation (LDA) [16]. The main advantage of LDA is that instead of a fixed topic representation in a HMM, latent topics in LDA can share the underlying syntactic and semantic concepts across multiple documents (topic segments) in a corpus.

In [85], authors presented an unsupervised topic segmentation method (PLDA) based on topic modeling framework. This method assumes a Markov structure over topic segments and associates a binary topic shift variable with each discourse unit. This variable indicates whether a set of consecutive discourse units share the same topic distribution or not. The start of a new topic segment is indicated by a switch in the state of topic shift variable. An alternative Bayesian approach (BayesSeg) for topic segmentation was proposed in [29]. This approach does not require the assumption that documents in a corpus share a set of topics. It can be applied to each meeting individually. This method applies a dynamic programming algorithm to find the most compact set of topic models to segment the data. Evaluations on meeting data have shown that Bayesian methods generally outperform HMMs [85, 103].

2.4.3 Boundary feature based methods

Topic segmentation methods based on lexical information usually ignore the fact that multiparty interactions are instances of social interactions in which speakers can exhibit both verbal (distinct phrases) and non verbal cues to signal topic changes. One of the first studies to investigate the relationship between topic boundaries and cue phrases was presented in [79]. This study showed that speakers often use specific phrases and cue words, such as *so, anyway*, etc at the start of a new topic. In broadcast news, [66] observed that participants use domain specific key phrases such as *welcome back, joining us* and *just in* to signal story changes. Instead of handcrafted cue phrases, other studies [36, 132] applied a supervised system to identify several candidate cue phrases from words that appear near topic boundaries. These cue phrases were then used as features for topic segmentation in meetings.

In [79], authors report that long duration pauses in BN conversations are indicative of topic changes. Prosodic features extracted from F0 and energy contours have also been investigated for topic boundary detection [114]. Other studies have also investigated the potential of motion features extracted from hand and head movement to detect topic boundaries. In [132], it was shown that combination of cue phrases, prosodic, structural and motion features significantly improved topic segmentation performance in meetings compared to using those feature groups individually. This study also revealed that addition of LCseg output with boundary features in a supervised classifier outperforms each of those methods. This suggests that word frequency based approaches and boundary features provide complementary information for topic segmentation in meetings.

2.4.4 Evaluation metrics

In traditional classification tasks evaluation metrics are obtained from aggregating the scores obtained by comparing each instance of the reference class to the output class predicted by the classifier. In the case of topic segmentation, we can assume that speaker turns are base instances for classification, i.e. classifiers predict whether a turn is a potential boundary or non boundary. The performance of the classifier is often summarized using a confusion matrix. The diagonal elements in the matrix represent the number of correct predictions belonging to boundary (true positives) and non boundary (true negative) classes. The offdiagonal elements represent the two types of errors: false positives, i.e., instances where the classifier predicts a boundary when there is no true boundary and false negatives, i.e., classifier predicts a non boundary when there is true boundary.

Table 2.1: Confusion Matrix.

True Positives	False Positives
False negatives	True negatives

The predictive accuracy of a classifier is good evaluation metric when the classes are balanced. However, in applications where the classes are imbalanced accuracy does not result in a reliable measure. This can be serious issue in case of topic segmentation, where the reference boundary marks in the data are much smaller compared to number of turns in the conversation. Classifiers can be optimized to generate high accuracy if they label all the turns as non boundary. Measures, such as precision and recall that consider both types of errors can also be used for evaluation. However, even these measures are not effective for evaluating the performance of segmentation as they are insensitive to near misses. In Table 2.2, we can observe that even though both System1 and System2 produce zero recall and precision, System1 generates a segmentation which is nearly similar to the reference segmentation.

Table 2.2: Comparing the performance of two automatic segmentation algorithms.

Turns	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
Reference	0	0	0	1	0	0	0	0	1	0	0	0
System1	0	0	1	0	0	0	0	1	0	0	0	0
System2	1	0	0	0	0	0	0	0	0	0	1	0

$\mathbf{P_k}$

To resolve the issues inherent in standard evaluation measures, P_k was proposed as an alternative measure in [13]. P_k calculates the probability of segmentation error, such that two turns, drawn randomly from the dataset are incorrectly identified as belonging to same topic segment.

P_k is calculated by considering a window of fixed length k, and moving it across all the turns in the conversation. Let us define t_i and t_j as two turns which correspond to the endpoints of

the window and δ is an indicator function, which evaluates to one if the two turns belong to the same segment and zero otherwise.

$$\delta(t_i, t_j) = \begin{cases} 1 & \text{if } t_i \text{ and } t_j \text{ belong to the same topic segment} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Given a conversation of N turns and ref and hyp as the reference and hypothesized segmentation, P_k is defined according to (2.2),

$$P_k = \frac{\sum_{n=1}^{N-k} \delta_{hyp}(t_i, t_{i+k}) \oplus \delta_{ref}(t_i, t_{i+k})}{N-k} \tag{2.2}$$

The window length k is generally fixed to half of the average topic segment length in the reference segmentation ref. Since P_k estimates the probability of segmentation error a value of 0 results in a perfect segmentation, while higher values result in a significantly worse segmentation.

WD

P_k is one the most widely used measure to evaluate segmentation performance. However, an analysis of (2.2) shows that it only considers whether end points of the window contain a boundary or not, and fails to take into account the number of boundaries between the two end points.In [82], it was shown that this effect can result in scenarios where P_k fails to penalize false positives in the hypothesized segmentation. To mitigate this problem, [82] proposed Windowdiff (WD) as an alternative measure. WD also relies on moving window of fixed length k, however, unlike P_k it corrects for the number of boundaries.

Let $ref(t_i, t_{i+k})$ and $hyp(t_i, t_{i+k})$ denote the number of boundaries in the reference and hypothesized segmentation respectively. WD is defined according to (2.3),

$$WD = \frac{\sum_{n=1}^{N-k} (|ref(t_i, t_{i+k}) - hyp(t_i, t_{i+k})| > 0)}{N-k} \tag{2.3}$$

Like P_k, lower values of WD indicate better segmentation performance and a perfect segmentation results when WD evaluates to 0.

2.5 Speaker diarization

Current state-of-the-art systems for both role recognition and topic segmentation typically include feature extraction as the first stage. While feature extraction using manual speaker segmentation can be useful for initial research, a fully automatic system is dependent on output of speaker diarization system for feature extraction. Speaker diarization aims at identifying "who" spoke "when" in an audio recording. Speech data in meeting recordings is generally acquired using either headset microphones attached to each participant or distant (far-field) microphones. In the case of individual headset microphones, the number of participants in the conversation is known and speaker diarization simply involves segmentation of speech and non speech regions for each participant. Speaker diarization is more challenging in the case of distant microphones, as diarization systems do not assume any knowledge about the number of speakers and speech nonspeech boundaries in the recording. The output of a diarization system is a segmentation of unlabeled audio into speaker homogeneous speech segments. Figure 2.1 shows different stages of a typical diarization system. Next sections summarize the state-of-the-art approaches applied at different stages in speaker diarization.

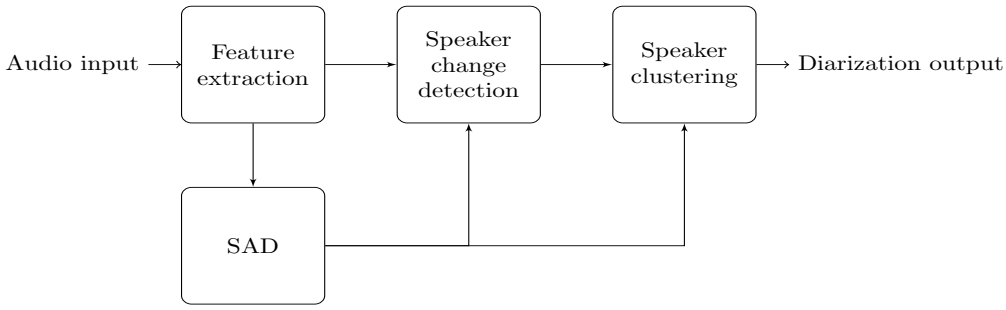

Figure 2.1: Stages in speaker diarization.

2.5.1 Features

In common with many speech processing techniques, such as speech and speaker recognition, speaker diarization systems also model information extracted from short term spectrum of speech. The speech spectrum is obtained by applying a Fourier transform on the windowed audio frames. Generally a hamming window of duration $30ms$ is applied at a frame rate of $10ms$. In most diarization systems, spectral information is represented using parametrized features like Mel-frequency cepstral coefficients (MFCC) [3]. Some studies have also explored alternative features including Perceptual Linear Prediction coefficients [104] and Linear Predictive Cepstral Coefficients [52]. Since speaker specific information is carried in relatively higher frequency bands, a higher number of MFCC (20) are extracted. In comparison to phoneme recognition, the goal of diarization is discrimination between speakers. As a result, delta and double delta features which capture phonetic information are not considered in diarization.

Several recent studies have considered the problem of diarization when meeting room audio

is captured using multiple distant microphones (MDM) [6, 125]. Those studies reveal that speaker discriminative information can be extracted from time delay of arrival (TDOA) of signals in different microphones. TDOA features were extracted as the peaks in crosscorrelation between different microphone channels. Studies have shown that TDOA features perform worse than acoustic features [6]. However, performance of meeting diarization system was improved by combination of MFCC and TDOA features [77].

2.5.2 Speech activity detection

Speech activity detection (SAD) refers to segmentation of speech and nonspeech regions in an audio recording. Depending on the method used to capture audio signal, the output of SAD system yields either the final speaker diarization or further processing by other modules (in Figure 2.1) is required. For the case where audio of each participant is captured using individual headset or lapel microphones, detecting speech and nonspeech regions for each audio channel accomplishes speaker diarization. Otherwise, when audio is captured using distant microphones, SAD system can still influence final speaker diarization. On the one hand, errors made by SAD system such as missed speech and false alarms directly contribute to final diarization error. On the other hand, inclusion of nonspeech segments can corrupt speaker diarization process by weakening inter-cluster discrimination of acoustic models.

Due to spontaneous nature of interaction in meetings, nonspeech regions may include high energy non lexical sounds such as laughter, breathing, coughing, etc., besides silence and ambient room noise. Therefore, comparing short term spectral energy against a threshold is ineffective [51]. In comparison, model based SAD systems have shown better performance for speech activity detection in multiparty interactions. In [126], acoustic models were pretrained on externally labeled speech and nonspeech data. Both speech and nonspeech classes were represented using using Gaussian mixture models. Other approaches based on discriminative classifiers such as Linear Discriminant Analysis and Support Vector Machines have also been investigated [87, 110]. Speech segmentation is obtained by applying Viterbi decoding on unlabeled audio using the pretrained models . Main advantage of this approach is that models can be targeted for specific speech and nonspeech classes. For example, models trained for different genders and channel conditions have been proposed for broadcast data [73]. However, considerable resources need to be spent for labeling training data that might not generalize to new meeting room environments.

Hybrid SAD systems have been proposed as an alternative approach by removing the dependency on labeled training data [125, 5, 74]. Most hybrid systems combine the characteristics of energy based and model based SAD systems. Initially a limited amount of data is labeled automatically using an energy based detector (for which there is high confidence in detection). The labeled speech and nonspeech data is then used to train separate acoustic models that were subsequently applied to yield speech non speech segmentation.

All the systems described so far are applicable to single channel audio recording. However,

in the context of multichannel IHM recordings there is additional source of noise in each participant's microphone due to crosstalk from adjacent participants. In [27], a set of auxiliary features, such as normalized multichannel energy, signal kurtosis and cross correlation between channels was investigated for crosstalk suppression. A Multi Layer Perceptron (MLP) was trained on combination of short term spectrum and auxiliary features to detect speech and nonspeech classes. During evaluation, MLP posterior features were converted to scaled likelihoods and Viterbi decoding was applied to yield final speech nonspeech segmentation.

2.5.3 Speaker segmentation

Multiparty interactions involve speech from more than one speaker. However, the output of SAD system under distant microphone conditions is unable to identify speech segments uttered by different speakers. Instead, diarization systems involve a speaker segmentation stage to mark the positions where speaker changes occur in the audio. Speaker changes are usually identified by comparing a threshold against a distance measure, that is computed between two speech segments around a hypothesized speaker change point. This can be considered as a hypothesis testing problem. At any hypothesized change point, if two adjacent speech segments belong to the same speaker then computed distance measure should yield a low value, otherwise distance measure should yield a higher value.

One of the most commonly used distance measure in speaker diarization is based on Bayesian Information Criterion (BIC) [89]. BIC was first introduced in [101] as a model selection criterion. Given a model $\mathcal{S}^i$ and acoustic feature set X_i, BIC measures the efficiency of the model in predicting the acoustic data. It is defined as:

$$BIC(\mathcal{S}^i) = log\ L(X_i|\mathcal{S}^i) - \frac{\lambda}{2}\#\mathcal{S}^i\ log(N_i) \tag{2.4}$$

where $L(X_i|\mathcal{S}^i)$ is the likelihood of data, $\#\mathcal{S}^i$ denotes the number of parameters in model $\mathcal{S}^i$ and N_i is size of feature set X_i. A free parameter λ controls the trade-off between likelihood and model complexity. Consider two adjacent speech segments represented using feature sets X_i and X_j. If these segments belong to the same speaker, then a model $\mathcal{S}^{ij}$ fits the combined data $X_i \cup X_j$, otherwise two different models $\mathcal{S}^i$ and $\mathcal{S}^j$ better explain the data. The change in BIC between the two alternatives is computed as:

$$\Delta BIC(i,j) = log\ L(X_i \cup X_j|\mathcal{S}^{ij}) - log\ L(X_i|\mathcal{S}^i) - log\ L(X_j|\mathcal{S}^j) - \lambda(\#\mathcal{S}^{ij} - \#\mathcal{S}^i - \#\mathcal{S}^j)log(N_i + N_j) \tag{2.5}$$

The speaker change points are detected whenever ΔBIC value falls below zero. Although this method does not require a separate threshold for comparison, the free parameter λ still needs to be tuned on the development data [113]. Alternatively, an analysis of (2.5) reveals that if sum of parameters $\#\mathcal{S}^i$ and $\#\mathcal{S}^j$ is identical to number of parameters $\#\mathcal{S}^{ij}$ in the combined model, λ gets eliminated from the equation. In [2], a method was proposed to adjust the model parameters so that λ gets canceled out.

The search of all speaker change points using BIC distance measure is computationally expensive. To reduce the computation cost a two pass strategy has been proposed [25, 63]. In the first pass simpler distance measures can be used. Simpler measures that have been proposed include log likelihood ratio test [25] and Kullback-Leibler distance [63]. In the second pass BIC can be used to refine the potential speaker change points identified in the first pass.

2.5.4 Clustering

The output of speaker segmentation is a sequence of speaker homogeneous speech segments. However, the number of distinct speakers and which speech segments are uttered by any given speaker are not known. Speaker clustering algorithms aim to gather all the speech segments corresponding to a given speaker into one cluster and ideally the number of clusters found automatically should equal the number of speakers. Most diarization systems follow a hierarchical approach and iteratively split or merge speaker segments until the desired number of speakers is reached. Since actual number of speakers in the recording is unknown, clustering algorithms halt the merge/split of clusters based on an automatic stopping criterion. In bottom up clustering, the algorithm is initialized with a large number of clusters (number of clusters can be equal to number of segments). The closest clusters are then iteratively merged until a stopping criterion is reached. The second approach is top down clustering. Initially all the segments belong to a single cluster which is then recursively split until an optimal number is reached. A distance measure between clusters and evaluation of stopping criterion are common subroutines in these clustering algorithms.

Similar to speaker change detection algorithm, the most common distance measure for bottom up clustering is also based on BIC. Initially each segment is regarded as a separate cluster and BIC distance is calculated between each pair of clusters. The cluster pair which is most similar, i.e., one with highest BIC is merged into one cluster. The earlier diarization systems performed speaker segmentation and clustering sequentially. However, one main drawback of sequential approach is that clustering can be adversely effected by the errors made in speaker segmentation stage. A possible approach to mitigate this problem is by iteratively performing segmentation and clustering. Viterbi realignment can be used to segment the audio based on current cluster models and subsequently new cluster models can be retrained on the resegmented audio. Viterbi algorithm produces an optimal segmentation taking into account data from all the speakers, while only local information is considered in speaker change detection. After several iterations of this process, the influence of initial segmentation errors on speaker clustering is reduced.

Current state-of-the-art speaker diarization systems implement iterative segmentation and clustering in an HMM-GMM framework. An agglomerative diarization system based on this framework was initially proposed in [52]. Since then several variants of this approach have been implemented [125, 78]. In HMM-GMM diarization systems, each HMM state represents a hypothesized speaker cluster which is modeled using a GMM. GMM parameters are trained

using all the data in each cluster by either using Expectation Maximization (EM) algorithm or MAP adaptation from a background speech model. The audio is then segmented using Viterbi algorithm and the new segmentation is used to update GMM parameters. Several iterations of this approach are performed. The clusters which produce the highest BIC score are then merged. In [52], the number of model parameters before and after each merge is kept constant. This eliminates the need for tuning a threshold and clusters can be merged as long as BIC distance between any cluster pair is positive. Otherwise, clustering stops and final speaker segmentation is produced.

Iterative segmentation and training in HMM-GMM system is computationally demanding. Recent studies have proposed an alternative approach based on non parametric methods that are computationally much faster than HMM-GMM diarization system. The information bottleneck (IB) diarization system [118] is bottom up system based on IB clustering framework [112]. In IB method, a set of relevance variables is introduced and objects are clustered based on their similarity with respect to relevance variables. In speaker diarization, the relevance variables are defined as components of a background GMM estimated over speech regions in the audio recording. IB clustering takes as input a set of speech segments X and set of relevance variables Y. The output of algorithm is a clustering representation C which simultaneously maximizes the mutual information $I(Y,C)$ of a set of relevance variables Y and a set of clusters C, while minimizing the mutual information $I(C,X)$ of the set C and X. This idea is expressed in (2.6):

$$\mathscr{F} = max[I(Y,C) - \frac{1}{\beta} I(C,X)] \tag{2.6}$$

Here β is a Lagrange multiplier representing the trade off between compression of initial representation $I(C,X)$ and amount of relevant information preserved $I(Y,C)$. The speaker diarization system [119] adopts a greedy agglomerative approach for maximizing (2.6). At the start the agglomerative algorithm assumes each segment in X is a cluster. Subsequently, clusters are iteratively merged such that decrease in $\mathscr{F}$ is minimized at each step. The clusters are merged until a stopping criterion based on normalized mutual information falls below a certain threshold. In comparison to HMM-GMM diarization system, IB diarization system is computationally less expensive since models are not retrained after merging of two clusters.

2.5.5 Evaluation metric

National Institute for Standards and Technology (NIST) has organized several Rich Transcription (RT) evaluation campaigns to benchmark the advances in the state-of-the-art speaker diarization systems. NIST has specified Diarization Error Rate (DER) as the measure to evaluate and compare the performance of different diarization systems. DER quantifies the mismatch between the output of the automatic system and the reference, in terms of the fraction of time that is not attributed correctly to a speaker or to nonspeech.

DER is computed in two stages. In the first stage, an one-to-one mapping of reference speaker labels and system generated speaker labels is constructed such that the overlap time between corresponding speaker labels is maximum. In the second stage, DER is computed as a sum of false alarm time, miss time and speaker error time. To quantify the different errors, we consider a audio recording in which a given speech segment seg has a duration $dur(seg)$. Let us denote by $N_{ref}(seg)$ and $N_{hyp}(seg)$ as the number of speakers in reference and hypothesis segmentation corresponding to speech segment seg. We also denote by $N_{correct}(seg)$ as the number of speakers which are correctly matched between the reference and hypothesis segmentation. The different components which constitute DER represent three sources of errors and are defined as follows:

- False Alarm time: fraction of scored time when a hypothesized speaker is labeled as nonspeech in the reference. It can be expressed as:

$$Err_{Fa} = \frac{\sum_{seg:N_{hyp}>N_{ref}} dur(seg)[N_{hyp}(seg) - N_{ref}(seg)]}{\sum_{seg} dur(seg)N_{ref}(seg)} \tag{2.7}$$

- Miss time: fraction of scored time when reference speaker label is not matched with any hypothesized speaker label. Overlapped speech is usually included in this error. It can be expressed as:

$$Err_{Miss} = \frac{\sum_{seg:N_{hyp}<N_{ref}} dur(seg)[N_{ref}(seg) - N_{hyp}(seg)]}{\sum_{seg} dur(seg)N_{ref}(seg)} \tag{2.8}$$

- Speaker error time: fraction of scored time when hypothesized speaker labels do not match reference speaker labels. It can be expressed as:

$$Err_{Spkr} = \frac{\sum_{seg} dur(seg)[min(N_{ref}(seg), N_{hyp}(seg)) - N_{correct}(seg)]}{\sum_{seg} dur(seg)N_{ref}(seg)} \tag{2.9}$$

DER is calculated as the sum of these errors.

$$DER = Err_{Fa} + Err_{Miss} + Err_{Spkr} \tag{2.10}$$

Reference segmentation used in calculation of DER can have imprecise boundaries around speech segments. NIST therefore recommends a collar of 250 milliseconds around speech segments where the diarization errors are not scored.

2.6 Meeting corpora

The focus of thesis is analysis of spontaneous multiparty interactions, with a focus on meetings. Meetings delimit human interaction in a realistic setting and allow study of complex social

behavior. Due to spontaneous nature of interaction between multiple participants, analysis of meeting conversations differ from previous studies in broadcast news [1], human computer dialogues [30] and telephone conversations [39]. In order to develop automatic tools for analyzing meeting interactions, several corpora of meetings have been recorded, including those at Linguistic Data Consortium (LDC) [23], CMU Interactive Systems Labs (ISL) [19], International Computer Science Institute (ICSI) [53], National Institute of Standards and Technology (NIST) [38]. Meetings were recorded in specially instrumented conference style meeting rooms built at these institutes. Other corpora which have focused on meeting analysis and interpretation, include Computers in Human Interaction Loop (CHIL) [71], Cognitive Assistants that Learns and Organizes (CALO) and Augmented Multiparty Interaction (AMI) [21]. Table 2.3 summarizes the scenarios and characteristics of some of these corpora.

Table 2.3: Meeting Corpora

	AMI	ICSI	NIST
Number of meetings	173 (100 hours)	75 (72 hours)	19 (15 hours)
Meeting type	natural	natural	natural
	e.g. movie club, office relocation	e.g. speech recognition, meeting recording	e.g. party planning, focus group
Meeting type	scripted		
	e.g. product development (introduction, conceptual/detail design, conclusion)		

Starting from the initial focus on speech and speaker recognition, the research in automatic processing of meetings has expanded and several algorithms for a wide range of applications have been developed. These applications include meeting activity recognition, gesture recognition, affective state recognition, speaker role recognition, dialogue act segmentation, topic segmentation and summarization. In order to evaluate and benchmark different algorithms periodic evaluation campaigns have been organized and conducted using the recorded meeting corpora. Starting from NIST's Rich Transcription 2002 (RT 02), NIST has organized a series of campaigns with focus on benchmarking various algorithms for automatic extraction of speech including speaker recognition and diarization. Other initiatives include Cross-Language Evaluation Forum (CLEF) sponsored by European Language Resource Association (ELRA) and Classification of Events, Activities and Relationships (CLEAR). These evaluation campaigns consider various tasks in metadata extraction and define evaluation measures to compare different algorithms on those tasks.

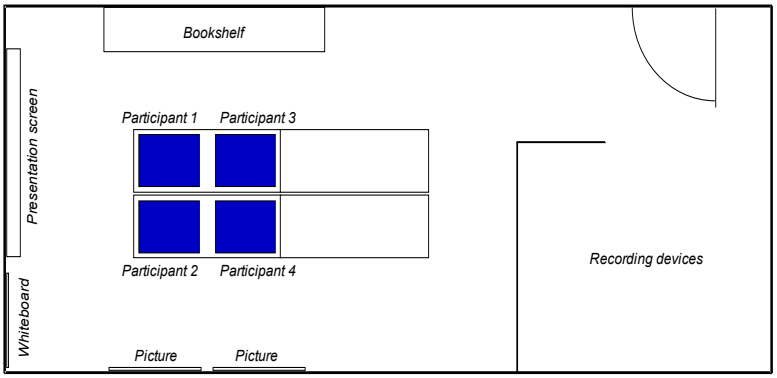

Figure 2.2: AMI meeting room setup.

2.6.1 Corpora used in this thesis

The main corpus used throughout this thesis is AMI meeting corpus. AMI corpus includes more than 100 hours of data collected in three instrumented meeting rooms located at University of Edinburgh (UK), Idiap Research Institute (Switzerland) and TNO Research Institute (Netherlands). Figure 2.2 shows a typical AMI room setup. Each instrumented meeting room was designed to capture behavior of four meeting participants simultaneously using four close up cameras for each participant and two or three room view cameras; headset and lapel microphones for recording clean audio signal for each participant; eight element circular microphone array for recording distant speech; and digital pens. Meeting rooms were also equipped with smart whiteboards, data projection and videoconferencing tools. The corpus also contains shared project specs and manual transcription for each meeting participant. The language used in all the meetings in the AMI corpus was English. However, more than half the participants in the corpus were non native speakers of English. Besides, manual annotation of speech transcripts, AMI corpus also provides ground truth annotations for formal roles, dialog act and topic segmentation.

The ASR transcripts used throughout this thesis are based on the output of AMI-ASR system [45] for IHM channels, with an average word error rate (WER) of nearly 30%. The acoustic models in AMI-ASR system are based on context dependent HMMs with emission probability distributions modeled using GMMs. Acoustic models were initially trained on conversational telephone speech and adaptation on meeting domain was performed using maximum a posteriori (MAP) technique. Speaker level adaptation was then performed using techniques like vocal tract length normalization and maximum likelihood linear regression. A standard trigram language model with a vocabulary of 50000 words was trained on a combination of several resources, including broadcast and conversational telephone speech transcripts, meeting data and related texts found by web search.

The meetings in AMI corpus are used to evaluate the main tasks performed in this thesis including social and formal role recognition, speaker diarization and topic segmentation. In addition, we also use other meeting corpora to evaluate the performance of automatic

systems for diarization and topic segmentation. For both of these problems, the social role recognition model is trained on AMI scenario meetings. Speaker diarization experiments were performed on NIST RT meetings corresponding to evaluation campaigns 2007 and 2009. Topic segmentation experiments were evaluated on 25 meetings selected from ICSI corpus.

2.7 Conclusion

This chapter gave an overview for the different tasks related to automatic analysis of multiparty interactions. We reviewed the literature on automatic role recognition for both formal and social roles. Furthermore, the literature on topic segmentation and speaker diarization was reviewed and evaluation metrics for those tasks were presented. Finally, at the end of this chapter we described the databases that will be used in this thesis. In the following chapters, we will investigate the relationship between social roles and its potential for improving speaker diarization and topic segmentation. We start our investigation in this direction by investigating automatic social role recognition which is the focus of next chapter.

3 Recognition of social roles in multi-party interactions

3.1 Introduction

This chapter presents a detailed study on automatic recognition of informal social roles in small group meetings and contains several contributions. The corpus we annotated for recognizing social roles in meetings has four times as many speakers compared to similar investigations [28]. This is relevant as role annotation is time consuming and relatively expensive, and results obtained on large datasets improve our confidence in the models learned by automatic systems. We model speaker behavior in terms of linguistic, turn taking and acoustic features. In comparison to earlier approaches [115, 133], this is the most extensive feature representation for social role recognition. We consider various feature groups individually and in combination, to understand the relative influence of each and the benefits of using them jointly. Furthermore, we also propose a classification framework based on conditional random fields, which integrates features extracted at multiple time scales in a single representation. Finally, this is the first work, to the best of our knowledge, where experiments are performed on both in domain and out of domain data. This is possible as roles in this work are informal and are not dependent on the specific scenario of multiparty interactions. By evaluating our models on multiple scenarios, we are able to investigate the robustness of the proposed approach.

The chapter is organized as follows. In Section 3.2, we discuss the dataset, description of social roles used in this work, and describe the process for annotation of roles. Section 3.3 presents the various features that are automatically extracted from turn taking, acoustic and linguistic behavior of speakers. In Section 3.4, we propose the supervised learning model for automatically recognizing social roles of speakers from the extracted features. The experimental methodology for role classification is presented in Section 3.5, where we also compare and discuss the performance of proposed method. The chapter is then concluded in Section 3.6.

Figure 3.1: A snapshot of meeting showing four speaker specific closeup cameras and an overview camera.

3.2 Corpus description

For the task of annotating social roles, we selected data from AMI meeting corpus [21]. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team and are tasked with designing a new remote control. The meeting is supervised by the PM who follows an agenda with a number of items to be discussed with other speakers.

The formal roles in AMI meetings are scripted and participants know beforehand the overall agenda of the meeting. Each speaker assumes only one formal role that remains fixed for the entire duration of the meeting. Besides formal roles, the speakers also assume informal roles. Informal roles assumed by speakers are influenced by their individual traits, such as personality and interaction with other group members. While the personality of a speaker remains relatively stable across different scenarios, the social roles develop in response to changing dynamics of group interaction. As the meeting progresses different role configurations can emerge and social role of a speaker can change from one type to another.

In order to classify speakers behavior into distinct social roles we follow the role coding scheme proposed in [133]. The underlying motivation behind this approach is that, while same speaker can assume different social roles, its role remains relatively stable over short time windows. Therefore, at each time instant a speaker will have a unique social role which can be defined using a set of acts and behaviors. The attributes of different roles are briefly summarized in the following:

- *Protagonist* - a speaker that takes the floor, drives the conversation, asserts its authority and assume a personal perspective.

- *Supporter* - a speaker that assumes a cooperative attitude, demonstrates attention and acceptance and provides technical and relational support.
- *Neutral* - a speaker that passively accepts ideas from other group members.
- *Gatekeeper* - a speaker that acts like group moderator, mediates and encourages the communication within the group.
- *Attacker* - a speaker who deflates the status of others, expresses disapproval and attacks other speakers.

For the present study a subset of 59 scenario meetings containing 128 different speakers (84 male and 44 female participants) was selected from the corpus. Subsequently each meeting was sliced into short clips (average duration less than 30 seconds). In each slice of meeting, the social role of a speaker was assumed to remain constant. Allocating social roles for short time meeting slices is supported by earlier work. In [115] manual annotations of social roles were smoothed over a one minute long sliding window for training of role recognition models. Furthermore, predicting speaker characteristics over short video clips, referred to as, "thin slices of behavior", is very well documented in social psychology literature [4]. Considering the nature of social role annotation over meeting recordings, this is particularly advantageous since annotators can work on short video slices and need not wait for the entire meeting recording to complete.

From each meeting, a total duration of approximately 12 minutes long audio/video data was selected. Meeting slices were resampled so as to cover the entire length of recording comprising various parts of meeting such as openings, presentation, discussion and conclusions. Using this approach, we generated 1700 meeting slices, corresponding to almost 12.5 hours of meeting data.

3.2.1 Role annotation

In this work, we have used an online environment for social role annotation and the human assessors were selected through the crowdsourcing platform, Amazon mechanical turk (AMT). The online platform allows raters to work on Human Intelligence Task's (HIT's), where they have an option to accept or reject a HIT, and are paid a small amount of money in exchange for providing annotations. The HIT requester can select raters using a set of inbuilt rater qualifications, including raters location and their HIT approval rate, i.e, the fraction of completed tasks that were accepted by other HIT requesters in the past. The requester can also specify the number of unique annotations for a set of HITs as well as reward payment for each HIT. All the completed annotations can be downloaded and reviewed by the requester who also has the option to reject any HIT which does not meet the requisite quality.

For the task of social role annotation we prespecified the inbuilt rater qualifications, i.e., location of raters and their HIT approval rate. As the meetings are in English, we decided

to set the location of raters to United states (US), where most people speak English as their first language. Since a large proportion of AMT raters are based in US, this requirement was not considered to adversely effect the quality of annotations. For the second qualification we decided to use raters whose HIT approval rate exceeds 95%.

Before starting each HIT, the raters were asked to follow a set of annotation guidelines. First, annotators were told that each HIT is a sequence of presentations and discussions according to a predefined meeting agenda. Second, attributes of all the five social roles were described. Third, annotators were asked to watch each clip individually and judgments should be based on behavior of participants with the clip, with focus on their interaction and what participants say and how they say it. Fourth, more than one participant can take the same role. Fifth, participants who are silent during a clip should be perceived as neutrals. Along with the annotation guidelines, the HIT also incorporates the video clips which the raters need to view before submitting their judgments. Figure 3.1 shows the snapshot of one of the selected video clips. The video clip for each meeting slice was obtained by merging the four speaker specific closeup cameras and an overview camera with the audio from individual headset microphones that each speaker wears.

To facilitate the annotation process, we grouped together the video clips from a single meeting in one HIT. Pilot studies revealed that a very large number of video clips in a HIT increases the task submission time. As a compromise about 10-11 meeting slices were grouped in a HIT. Annotators were provided with audio and video for each meeting and tasked with assigning a speaker to role mapping for each meeting participant appearing in the clip. We asked 11 annotators to rate each HIT. An analysis of completed annotations revealed that a majority of accepted HITs (70%) were completed by 10 or more than 10 raters and 95% of HITs were completed by 8 or more than 8 raters. Only HITs completed by 5 or more than 5 raters were used for further analysis.

3.2.2 Analysis of annotations

Since social roles described in this study are obtained from human raters, the role annotations were analyzed to investigate whether different raters come to fair understanding of annotation guidelines and produce consistently similar role labels. The simplest measurement of agreement between a pair of assessors is the observed agreement, which is defined as the percentage of instances where the two give the same answer. However, observed agreement is more favorable towards coding schemes with fewer categories and it does not take into account the distribution of instances among different categories. Several studies, such as [20], have favored the use of κ statistic to correct for chance agreement between annotators. This idea is expressed in the following equation:

$$\kappa = \frac{A_O - A_E}{1 - A_E} \tag{3.1}$$

where A_O measures the observed agreement, while A_E is the agreement that can be expected by chance. The κ coefficient yields a value 1 when there is complete agreement between annotators, while the value 0 signifies chance agreement. In this work, we have used Fleiss' kappa coefficient [33] as the measure of reliability as it can be used even when the number of raters is greater than two. It is also more suited for online environment as it does not require a separate chance probability distribution model for each rater.

In our first investigation, we analyzed the consistency of social role annotations by varying the context in which a video clip is presented to raters. Since video clips from the same meeting are grouped in a HIT, we investigated the possibility that raters might just remember faces of meeting participants from the initial clips and repeat the roles later. To check consistency of annotations we asked raters to annotate two sets of HIT's. The first set consists of HITs in which all the video clips are from the same meeting. For HITs in the second set, we randomly selected video clips from different meetings, thereby preventing same speakers to appear more often in the same HIT. In both cases about 11 video clips were grouped in a single HIT. Since we were interested in evaluating the aggregate performance of the annotation process, the social role for each participant in a meeting clip was obtained from majority voting. The interannotator reliability scores between the two sets are: $\kappa = 0.81 (N = 2260, p < 0.0001, \text{confidence interval}(\alpha = 0.05) : [0.78, 83])$. This corresponds to almost perfect agreement according to Landis and Koch's criterion [133]. This analysis suggests that online raters are fairly consistent in labeling social roles from the point of view of HIT design.

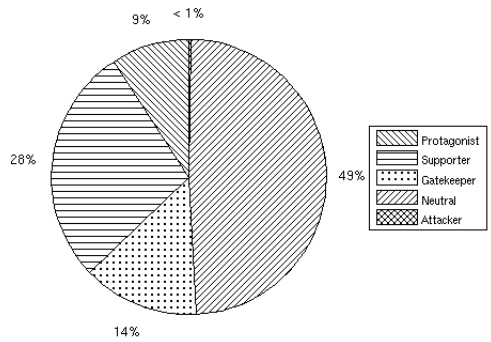

Figure 3.2: Overall distribution of individual social roles in the annotated data. The role label for each instance was obtained by majority voting.

The reliability of overall annotation process, measured using Fliess's kappa statistic, shows a value 0.5 which is considered to have moderate agreement ($0.4 < \kappa < 0.6$) according to Landis and Koch's criterion [133]. Highest level of agreement was observed for neutral role with κ equal to 0.7. An intermediate level of agreement is present for supporter 0.36 and gatekeeper 0.38 roles. This is followed by the protagonist role which shows a fair level of agreement with κ equal to 0.29. One difference from the earlier studies [133] is the higher percentage

of gatekeepers. We observed that the online raters were more likely to associate the role of gatekeeper with project manager, who supervises the overall agenda of the meeting.

Table 3.1: Social role distribution conditioned on speaking state (silence or speech) in a meeting slice.

	protagonist	supporter	gatekeeper	neutral
Speaking	0.15	0.22	0.19	0.44
Silent	0.0	0.02	0.0	0.98

Figure 3.2 shows the distribution of social roles for all the instances in the corpus. Each instance was labeled with the social role obtained by taking a majority vote. The pie chart reveals that role distribution is far from uniform. We observe that very few instances were labeled as attacker. This may be due to collaborative nature of AMI meetings and participants tend to avoid showing hostile attitude. In comparison, neutral label is most prevalent and occupies nearly half of all labeled instances. Further analysis revealed that neutral role is mostly associated with speakers who are completely silent over the duration of meeting slice. In Table 3.1, we compare the role taking behavior of speakers conditioned on the fact whether they speak in the meeting slice or not. We observe that raters were unlikely to label silent speakers with active role like protagonists or gatekeepers. On the other hand, there appears to be a clear association for such speakers and neutral role. This is in accordance with the neutral characteristic of being mostly passive observers.

While Table 3.1 shows the overall distribution of social roles, we also investigated the various group configurations in which the roles appear in meetings. Table 3.2 shows that most frequent group configurations (35% occurrence) have one active speaker who takes the role of either gatekeeper or protagonist, while other three speakers act as neutrals. We also notice that simultaneous appearance of two protagonists or two gatekeepers in a meeting should be a very rare phenomena. This suggests that the active speaker, while assuming these roles, maintains control over the conversational floor. On the other hand, it is likely that more than one speaker can assume a supporters role in the group.

Our investigations also revealed that the raters tend to perceive continuity in role taking behavior of meeting participants. A correlation analysis of the role taking behavior in time revealed a positive correlation ($\rho = 0.46, p < 0.001$) between social roles across adjacent meeting slices. In Figure 3.3, we show the distribution of social roles conditioned on the role assumed in the previous meeting slice. For each previous social role, the probability that the

Table 3.2: Frequency of occurrence of various social role group configurations. Only configuration with a frequency ≥ 1 are reported.

protagonist	1	1	1	0	0	0	0	0	1
supporter	0	1	2	1	2	0	1	2	2
gatekeeper	0	0	0	0	0	1	1	1	1
neutral	3	2	1	3	2	3	2	1	0
occurrence	17	11	5	4	4	18	15	10	2

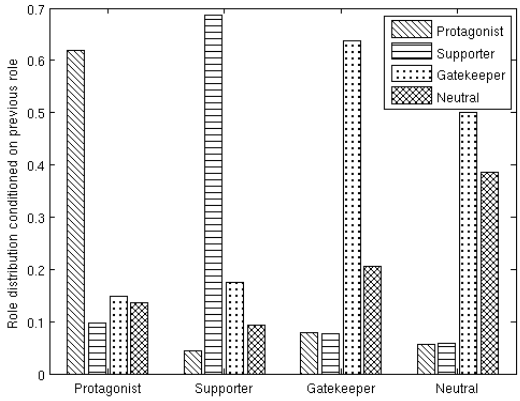

Figure 3.3: Social role distribution in current meeting slice conditioned on participants social role in the previous meeting slice. The vertical axis shows the role transition probability across adjacent meeting slices.

speaker retains the same social role in the current slice is higher compared to the probability that social role changes in the current slice. This suggests that speakers continue to retain the same social role across adjacent meeting slices.

3.3 Feature extraction

Motivated from previous research in automatic role recognition (described in Chapter 2), we extract both verbal and non verbal features from audio data to capture the speakers behavior during the meeting. Other non verbal features, such as hand and body fidgeting extracted from video data can also be modeled for role recognition. However, in this work we focus on audio features as they can be extracted from meetings for which audio track alone is available [53]. In this work, all the speech transcripts were generated using output of AMI-ASR system [45], which has a word error rate of nearly 30%.

3.3.1 Short term features

Turn taking is a basic form of organization for conversations in small group interaction [90]. Not only does it serve as a mechanism for effective communications, but speech activity and speaking time are perceived as indicators of influence and power over other group members in a conversation [9]. In this work, we consider turn taking as a sequence of speech and silence patterns that can be automatically extracted using standard speech processing tools. Intuitively it is also clear that for any given meeting slice, duration of a particular speech or silence region would be of much shorter duration relative to duration of the entire meeting slice.

Audio from the independent headset microphones (IHM) is processed through a speech

segmentation system [44] for obtaining estimated speech/non-speech boundaries for each meeting participant. The output of speech/non speech system for each speaker is a sequence of speech and silence regions in time, which arise due to turn taking in conversations. However, since meeting conversations involve multiple speakers, some activity regions (speech overlaps) will have more than one participant speaking simultaneously. Furthermore, silence regions can be produced due to different phenomena. On the one hand, silence may be produced due to a pause in conversation, when conversation floor changes occur or speakers stop to take breathe. On the other hand, silence can simply be the listening silence from the perspective of some speaker when other speaker(s) is/are speaking.

Each speaker's sequence of speech silence regions are tagged with one of the turn taking states defined as: talkspurts (TS), i.e., a region of speech when only a single speaker speaks; pauses (PA), i.e., regions when all the speakers are silent; overlaps (OV), i.e., regions where multiple speakers are speaking simultaneously; and listening silence (LS), i.e., a region where the current speaker is silent and any other speaker is speaking. A minimum duration criterion (200 ms) is applied to smooth each of these regions. We hypothesize that social roles influence the distribution of turn taking states. For example, it is more likely that a speaker with a more active role will grab the conversation floor after a pause. Similarly, the social role of a speaker can influence whether the speaker retains control of conversation after a speech overlap.

We now describe the extraction of short term features for a turn taking sequence of length N. At each time n, we extract the turn taking state $q_n \in \{PA, OV, LS, TS\}$ and the duration d_n of state q_n. A set of 24 different features were defined from this information. These features are of the type: $\delta(q_1)$ and $\delta(q_N)$, to represent whether the speaker starts or ends a conversation; $\delta(q_n - 1, q_n)$, to represent events like floor grab after a pause or an overlap; and d_n and d_n^2 represent the duration of states. Furthermore, whenever $q_n = TS$, we extract words from speech transcripts. We compile a list of words which speakers use frequently during TS states. The lexical features for each talkspurt were then represented as vector $\mathbf{w}_n$ of unigrams. At time n, we represent the lexical and speech activity information in a sparse feature vector $\mathbf{x}_n$ with dimensionality 636. The complete short term feature sequence of length N is represented using $\mathbf{X}_{\mathfrak{S}}$, where $\mathbf{X}_{\mathfrak{S}} = [\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_N]$.

3.3.2 Long term features

Besides extracting short term turn taking information, we also investigate various long term structural, linguistic and acoustic features extracted from the entire meeting slice. The linguistic and acoustic information is used to capture the speaking style of participants. By speaking style we mean "how participants talk" instead of "what they say". Our definition of speaking style includes both language style, as well as acoustic analysis of vocal expression patterns. The linguistic, acoustic and structural features investigated in this work are described next.

(1) *Linguistic features*: The words used by participants in a group interaction can convey important information about their motives and functions. Existing findings in psychology

- Standard counts:
  - Word count (WC), words per sentence (WPS), type/token ratio (Unique), words captured (Dic), words longer than 6 letters (Sixltr), negations (Negate), assents (Assent), articles (Article), prepositions (Preps), numbers (Number)
  - Pronouns (Pronoun): 1st person singular (I), 1st person plural (We), total 1st person (Self), total 2nd person (You), total 3rd person (Other)
- Psychological processes:
  - Affective or emotional processes (Affect): positive emotions (Posemo), positive feelings (Posfeel), optimism and energy (Optim), negative emotions (Negemo), anxiety or fear (Anx), anger (Anger), sadness (Sad)
  - Cognitive Processes (Cogmech): causation (Cause), insight (Insight), discrepancy (Discrep), inhibition (Inhib), tentative (Tentat), certainty (Certain)
  - Sensory and perceptual processes (Senses): seeing (See), hearing (Hear), feeling (Feel)
  - Social processes (Social): communication (Comm), other references to people (Othref), friends (Friends), family (Family), humans (Humans)
- Relativity:
  - Time (Time), past tense verb (Past), present tense verb (Present), future tense verb (Future)
  - Space (Space): up (Up), down (Down), inclusive (Incl), exclusive (Excl)
  - Motion (Motion)
- Personal concerns:
  - Occupation (Occup): school (School), work and job (Job), achievement (Achieve)
  - Leisure activity (Leisure): home (Home), sports (Sports), television and movies (TV), music (Music)
  - Money and financial issues (Money)
  - Metaphysical issues (Metaph): religion (Relig), death (Death), physical states and functions (Physcal), body states and symptoms (Body), sexuality (Sexual), eating and drinking (Eating), sleeping (Sleep), Grooming (Groom)
- Other dimensions:
  - Punctuation (Allpct): period (Period), comma (Comma), colon (Colon), semi-colon (Semic), question (Qmark), exclamation (Exclam), dash (Dash), quote (Quote), apostrophe (Apostro), parenthesis (Parenth), other (Otherp)
  - Swear words (Swear), nonfluencies (Nonfl), fillers (Fillers)

Figure 3.4: Linguistic categories used in LIWC.

have linked language style with use of simple functional words - pronouns, prepositions, articles and other emotional categories. Language style has been used to analyze personality traits [70]. Recent studies also reveal that quantitative analysis of language style, can be used for understanding social dynamics in small groups, and predicting aspects like leadership [92] and group cohesion [40].

Linguistic Inquiry and Word Count (LIWC) is a psychologically validated state-of-the-art text analysis program that quantifies the language style used by participants in a conversation [81]. LIWC operates by counting the fraction of spoken words that fall into predefined categories,

Table 3.3: Low level descriptors of vocal expression computed from the raw audio file.

Spectral
Zero crossing rate,
Energy in bands 250-600Hz,1-4KHz,
Spectral roll off points at 25%,75%,90%,
Spectral flux and harmonicity
MFCC 1-12
Energy and Voicing Related
RMS energy,
F0, Probability of voicing,
Jitter, Shimmer,
Logarithm of Harmonics to Noise ratio(HNR)

such as function words (pronouns, articles or auxiliary verbs) and psychological (emotion, social words, cognitive mechanism) processes. Figure 3.4 describes the the details of organization of various LIWC categories and subcategories. Speakers convey their emotional and personal preferences by using common words which describe these processes. For example, positive actions and events are often described by emotional words (e.g. nice, good). Similarly assents are often used people to signal agreement or disagreement.

The core part of LIWC program is a dictionary composed of almost 4500 words. There are 80 categories along which word usage can be measured in LIWC. The language categories are overlapping in the sense that a word can belong to more than one category. If a speaker uses a word like *support*, the program increments the current score of both verb category and positive emotion category. The categories can also be hierarchical, for example, positive emotion is a sub category within affect, so for a word like *support*, the counts for both positive emotion and affect categories are incremented. A detailed description of various linguistic categories used for role recognition are presented in [94].

(2) *Acoustic features*: To capture the speaking style information conveyed by vocal expression patterns, we have followed a brute force strategy, based on extracting a very large set of features from acoustic data [94]. We have been motivated in following this approach, as recent studies have revealed that systematically generated large set of acoustic features can capture complex phenomena, like leadership emergence in online speeches [123] and recognizing conflicts [100] in group discussions. Our acoustic features include standard prosodic features like fundamental frequency (F0) and energy, as well as features related to voice quality and spectral information. The feature extraction process works in two passes. In the first pass, acoustic data from IHM is processed at frame rate to extract low level descriptors (LLDs) for each meeting slice. The next pass projects each participant's LLD contour to a fixed size feature vector using statistical and regression functionals.

Table 3.3 shows the various LLDs which were extracted from acoustic data. The LLDs represent traditional prosodic features like F0 and speech energy which have been used for role recognition [94]. Voice quality features like jitter and shimmer were extracted to capture the perception of harshness in voice. We also extracted various spectral and MFCC coefficients.

Table 3.4: Set of functionals used to obtain acoustic features vectors. The functionals were applied to contours generated from lld descriptors in Table 3.3 and the implementation is based on the system presented in [31]

Statistical functionals
arithmetic mean, geometric mean
standard deviation, skewness, kurtosis
range, maximum, minimum
Regression functionals
linear regression slope, intercept and approximation error
quadratic regression coefficients and approximation error

These features are informative for recognizing personality characteristics like openness and conscientiousness [84]. Statistical and regression functionals defined in Table 3.4 were used to obtain features vectors from the contours of LLDs and their first order derivatives. This procedure yields a fixed size feature vector for each participant in the meeting slice, irrespective of the duration they are speaking. In this work, all the acoustic features were extracted from open-source feature extractor openSMILE [31].

(3) *Structural features*: A set of structural features was extracted from speech data. These features represent the total speech time, number of speaker turns in a slice, number of speakers who are active within a slice and total duration of overlapping speech. Also included were statistics like maximum, minimum and mean and standard deviation for these features.

3.4 Automatic social role recognition

In the previous section, we described the features that were used to characterize speakers behavior in a meeting. We now present an approach to automatically predict the role of a speaker using those extracted features.

During the process of feature extraction, we computed features which represent both the turn taking interaction and long term behavior of participants. The short term features capture changes in turn taking patterns and are computed over relatively short time, such as length of a talk spurt (average duration ~ 2 seconds), while long term linguistic, acoustic and structural features are computed over the length of an entire meeting slice (average duration ~ 30 seconds). To represent speaker behavior at multiple time scales, we propose a framework for social role recognition influenced by hidden conditional random fields (HCRFs) [43, 86]. The proposed method offers the benefits of discriminative learning and flexibility to include multiple non-independent features. Also, unlike static methods like support vector machines, the proposed method is capable of directly modeling the relationship between a social role and a dynamic sequence of short term features.

The training data in the corpus is defined for a set of speakers $\mathscr{S}$, who assume social roles in set $\mathscr{R}$, and participate in a set of meetings $\mathscr{M}$. We define $\mathscr{S}_M \subset \mathscr{S}$ as the subset of speakers appearing in meeting M During the annotation process, M is partitioned into a sequence of

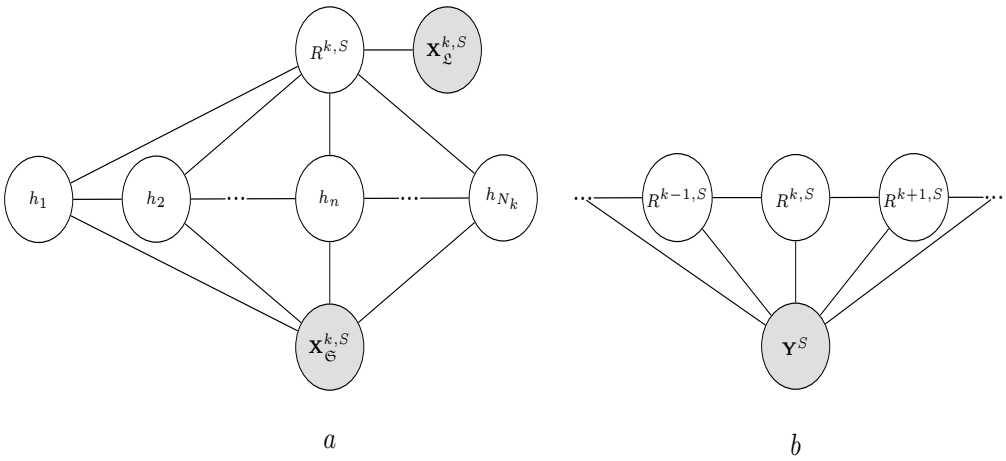

Figure 3.5: Graphical representation of CRFs for social role recognition. (a) Modeling influence of roles on short term and long term observations (b) Modeling sequential dependencies between roles. An open node represents a random variable and the shaded node is set to its observed value.

slices for which social roles are labeled. The variable $k \in \mathscr{K}_M = \{1,..K_M\}$ is used to index the meeting slices. For any speaker $S \in \mathscr{S}_M$, we represent using $R^{k,S} \in \mathscr{R}$ as the role taken by S in slice k. Note that $R^{k_1,S}$ and $R^{k_2,S}$ need not be same for any pair of segments $k_1, k_2 \in \mathscr{K}_M$. We also define the observations for S in k. The dynamics of turn taking are represented using a N_k length temporal sequence $\mathbf{X}_{\mathfrak{S}}^{k,S}$ (see section 3.3.1). The long term features are represented using vector $\mathbf{X}_{\mathfrak{L}}^{k,S}$. The tuple $\mathbf{X}^{k,S} = (\mathbf{X}_{\mathfrak{S}}^{k,S}, \mathbf{X}_{\mathfrak{L}}^{k,S})$ characterizes the participant behavior associated with role $R^{k,S}$.

The problem of automatic role recognition is that of learning a stochastic mapping from the feature space $\mathscr{X}$ to the label space $\mathscr{R}$. In this work, the conditional distribution $P(R^{k,S}|\mathbf{X}^{k,S})$ factorizes according to an undirected graphical model. Figure 3.5a shows the nodes representing the observation and latent variables in the model and the edges that encode the dependencies between these variables. The latent variables are represented by $\mathbf{h} = [h_1, h_2, ..., h_{N_k}]$. The distribution $P(R^{k,S}|\mathbf{X}^{k,S})$ is expressed in terms of product of potential functions:

$$P(R^{k,S}|\mathbf{X}^{k,S}) = \frac{\Psi(R^{k,S}, \mathbf{X}^{k,S})}{Z(\mathbf{X}^{k,S})}, \text{where} \tag{3.2}$$

$$\Psi(R^{k,S}, \mathbf{X}^{k,S}) = \Psi_{\mathfrak{S}}(R^{k,S}, \mathbf{X}_{\mathfrak{S}}^{k,S})\Psi_{\mathfrak{L}}(R^{k,S}, \mathbf{X}_{\mathfrak{L}}^{k,S}) \tag{3.3}$$

The term $Z(\mathbf{X}^{k,S})$ is the partition function that ensures conditional distribution sums to one over all labels. The potential $\Psi_{\mathfrak{S}}$ depends on the short term observations and the potential $\Psi_{\mathfrak{L}}$ depends on the long term observations. We assume that $\Psi_{\mathfrak{S}}$ factorizes according to a set of features $\{f_i\}$ and weights $\{\alpha_i\}$ and $\Psi_{\mathfrak{L}}$ factorizes according to a set of features $\{g_i\}$ and weights $\{\beta_i\}$. The expressions $\Psi_{\mathfrak{L}}$ and $\Psi_{\mathfrak{S}}$ take the form:

$$\Psi_{\mathfrak{S}}(R^{k,S},\mathbf{X}^{k,S}_{\mathfrak{S}}) = \sum_{\mathbf{h}} \exp(\sum_i \alpha_i f_i(R^{k,S},\mathbf{h},\mathbf{X}^{k,S}_{\mathfrak{S}})) \tag{3.4}$$

$$\Psi_{\mathfrak{L}}(R^{k,S},\mathbf{X}^{k,S}_{\mathfrak{L}}) = \exp(\sum_i \beta_i g_i(R^{k,S},\mathbf{X}^{k,S}_{\mathfrak{L}})) \tag{3.5}$$

The real valued weights $\{\alpha_i\}$ and $\{\beta_i\}$ represent the parameters of the model.

The g_i feature function directly model the relationship between long term observations $\mathbf{X}^{k,S}_{\mathfrak{L}}$ and $R^{k,S}$. On the other hand, we define two types of f_i feature functions. The feature function $f_i(R^{k,S},\mathbf{h})$ represent the relationship between $R^{k,S}$ and hidden variable $\mathbf{h}$. This function captures the distribution of hidden states associated with a role label. The observation feature function $f_i(\mathbf{h},\mathbf{X}^{k,S}_{\mathfrak{S}})$ relates the hidden variables with short term observations.

Given a training set of labeled instances the model parameters $\Lambda = (\{\alpha_i\},\{\beta_i\})$ are estimated by maximizing the conditional log likelihood:

$$L(\Lambda) = \sum_{M=1}^{|\mathscr{M}|} \sum_{\forall S \in \mathscr{S}_M} \sum_{k=1}^{K_M} \log P(R^{k,S}|\mathbf{X}^{k,S};\Lambda) \tag{3.6}$$

The objective function can be maximized using an iterative algorithm like stochastic gradient ascent or quasi Newton method like L-BFGS [61]. In this work, we have used L-BFGS algorithm, as it is a scalable method with low memory requirements and has been applied successfully for training HCRFs [43].

The role distribution in (3.2) can be extended to incorporate the continuity in role taking behavior of meeting participants. Figure 3.3 shows that distribution of social roles in the present slice are influenced by the speakers role in the previous slice. Using (3.2), we define a posterior feature vector $\mathbf{Y}^{k,S} = \{P(R^{k,S}|\mathbf{X}^{k,S}),\ \forall R^{k,S} \in \mathscr{R}\}$ for every slice k and speaker S. We note that $\mathbf{Y}^{k,S}$ can be efficiently computed using (3.6). We define the role sequence $\mathbf{R^S} = \{R^{1,S},...,R^{k,S},...,R^{K_M,S}\}$ and feature sequence $\mathbf{Y^S} = \{\mathbf{Y}^{1,S},...,\mathbf{Y}^{k,S},...,\mathbf{Y}^{K_M,S}\}$. A linear chain CRF shown in Figure 3.5b, is applied to estimate the conditional probability of the role sequence.

$$P(\mathbf{R}^S|\mathbf{Y}^S) \propto \prod_k \Phi_k(R^{k,S},R^{k-1,S},\mathbf{Y}^{k,S}) \tag{3.7}$$

where Φ_k is the local potential function for slice k. The potential Φ_k is represented as a linear combination of feature functions $\{\gamma_j\}$ and weights $\{\theta_j\}$. Two types of feature functions were defined: $\gamma_R(R^{k,S},\mathbf{Y}^{k,S})$ which captures relationship between role and posterior features $\mathbf{Y}^{k,S}$ in a slice and $\gamma_{RR'}(R^{k,S},R^{k-1,S})$ which captures role transition information across meeting slices.

$$\Phi_k(R^{k,S},R^{k-1,S},\mathbf{Y}^S) = \exp(\sum_j \theta_j \gamma_j(R^{k,S},R^{k-1,S},\mathbf{Y}^S)) \tag{3.8}$$

The parameters $\Theta = \{\theta_j\}$ of the model are estimated by maximizing the conditional log likelihood of the role sequence.

$$L(\Theta) = \sum_{M=1}^{|\mathscr{M}|} \sum_{S=S_1^M}^{S_M^M} \log P(\mathbf{R}^S|\mathbf{Y}^S;\Theta) \tag{3.9}$$

Since the graphical models in Figure 3.5 have a tree structure, algorithms like forward-backward and Viterbi decoding can be applied to efficiently estimate the model parameters Λ and Θ. The training process of the model is mainly dominated by forward-backward computation to evaluate the log-likelihood and its gradient vector at each iteration. The time complexity at each iteration scales linearly with the number of training instances, feature dimensionality and average length of turn taking sequence and quadratically with number of hidden states.

3.5 Experiments

Evaluation experiments on the scenario meetings of AMI corpus were conducted using k-fold crossvalidation. The annotated dataset was split into k sets, k-1 used for training and the remaining one used for testing. The procedure is repeated k times and each time a different set is left out for testing. For experiments in his study, $k = 22$. Each set comprises of a group of speakers who participate together in a meeting. The partitioning of data into different sets was performed to maintain strict separation between training and test sets in terms of speaker identity. This makes our approach speaker independent as same speaker does not appear simultaneously in both training and testing sets. The ground truth social role label for each instance was derived by taking a majority vote over rater assignments. An initial filtering was done to consider only those instances where a participant is speaking within the meeting slice (see Table 3.1). Furthermore, a few meeting slices where majority voting resulted in participant having an attacker role label were not considered (see Figure 3.2). The performance was measured in terms of overall role recognition accuracy and F-measure/Precision/Recall for individual roles.

3.5.1 Regularization

The number of parameters in the proposed model is large relative to the number of examples available during training. To avoid the problem of model overfitting the training data, a regularization term was added in (3.6). A commonly used technique in CRF training is to add the ridge regularizer, that imposes a zero mean Gaussian prior over model parameters to prevent overfitting. We have applied the same expression during training.

Figure 3.6 shows the effect of regularization on the performance of the model. We observe from the plot that, as the number of training iterations increases, the performance of the unregularized model starts degrading, suggesting that the model is overfitting the training

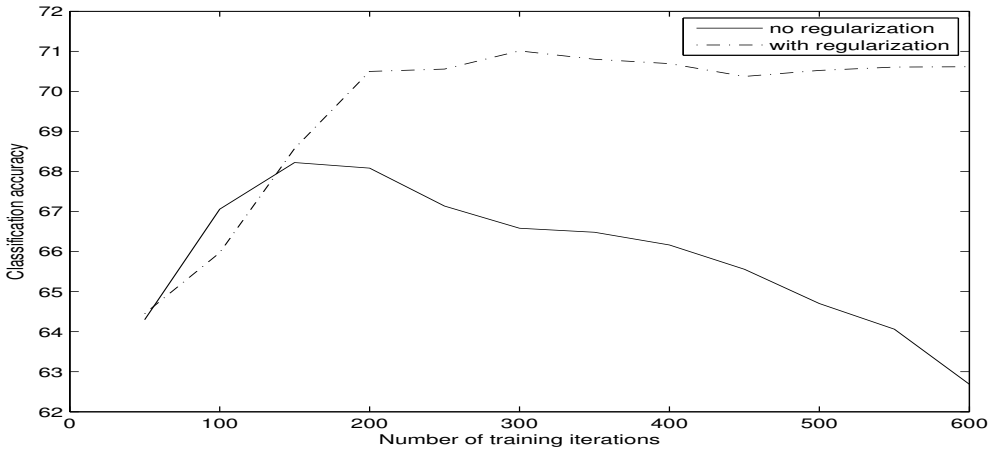

Figure 3.6: Comparison in performance of proposed system when the models are trained with and without adding a regularization term.

data. In comparison, the model trained with regularization converges to a higher classification accuracy showing that overfitting is avoided.

3.5.2 Model and feature selection

We first investigate the performance of various long term features on automatic recognition of social roles. Since the individual features have different scales, we applied a standardization technique to these features, such that each feature is normalized to zero mean and unit variance. Table 3.5 shows the different long term feature groups, the number of features within

Table 3.5: Long term feature groups and their role recognition performance

Feature Group	Number of features	Recognition Accuracy
Voice quality	308	0.60
MFCC	200	0.61
spectral	748	0.60
structural	35	0.62
LIWC	60	0.59

each group and their classification accuracy. The last column in Table 3.5 shows that accuracy of structural features is the best amongst all long term feature groups. We observe that size of a feature group does not explain the difference in their relative performance. The two feature groups with lower size, i.e., structural and LIWC features achieve an accuracy of 62% and 59% respectively. On the other hand, even though acoustic feature groups have larger dimensionality, there performance is lower than that of structural features.

We next explored the effect of different feature combinations on role recognition performance. Table 3.6 illustrates the impact of combining each long term feature group with structural

Table 3.6: Effect of combining different feature groups with structural features. The last column shows the result when all long term features where combined.

LIWC	MFCC	spectral	Voice quality	ALL long term
0.67	0.66	0.66	0.65	0.66

features. The last column in Table 3.6 shows the performance when all long term features are combined. Results show that combining both linguistic (LIWC) features and various acoustic (spectral, mfcc, voice quality) features improves the recognition accuracy. We performed a repeated one way analysis of variance (ANOVA) to determine whether the improvement in performance due to feature combination is significantly better than using structural features alone. ANOVA reveals a significant improvement in performance ($F = 4.7; p < 0.01$) when features are combined. However, Post hoc tests (Tukey HSD) did not reveal any significant difference in performance when all long term features were combined (Column 5) and other feature combinations (Columns 1-4). This suggests that linguistic and acoustic features are complementary to structural features, and it is useful to incorporate some, but not necessary all, of the long term features into the role recognition model. We also note that, while ANOVA analysis reveals significant improvement when features are combined, it is debatable whether the resulting improvement is large enough to be of practical significance.

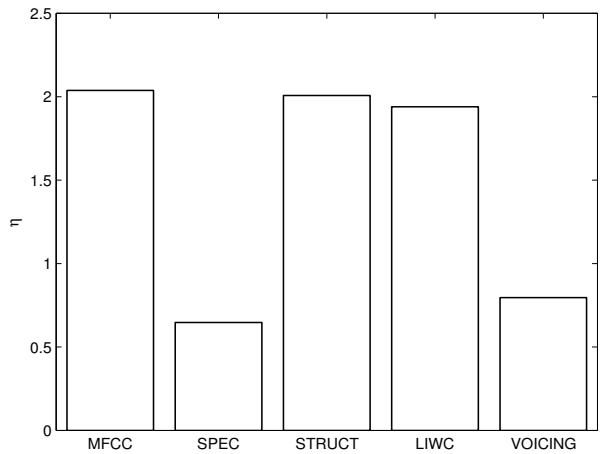

Figure 3.7: Comparison of different long term feature groups after feature selection is applied. η measures the relative importance of each feature group.$\eta > 1$ reveals that distribution of selected features from a group is higher after feature selection is applied compared to their initial distribution.

A feature selection algorithm [80], based on the principle of mutual information, was applied to find the most relevant features in the long term feature set. The feature selection algorithm estimates a scoring criterion that quantifies the relevance of including a specific feature in the set. The algorithm was applied across each cross validation fold and features were ranked. A portion of training data in each fold was used to train the model for different

Table 3.7: Per role F-measure, Precision and Recalls obtained in recognizing social roles for the three considered models. Asterisk besides the accuracy shows that improvement is statistically significant with rejection of null hypothesis at 5%

	Per-role F-measure (Recall/Precision)				Accuracy
Model	Protagonist	Supporter	Gatekeeper	Neutral	
baseline	(0.31/0.57)0.4	(0.84/0.66)0.74	(0.51/0.55)0.53	(0.56/0.71)0.63	0.64
HCRF	(0.52/0.62)0.57	(0.84/0.7)0.76	(0.56/0.63)0.59	(0.57/0.73)0.64	0.69
SVM	(0.49/0.56)0.52	(0.84/0.73)0.78	(0.52/0.58)0.55	(0.69/0.76)0.72	0.70
proposed	**(0.59/0.65)0.62**	**(0.83/0.76)0.79**	**(0.62/0.66)0.64**	**(0.72/0.77)0.75**	0.74*

sizes of ranked feature set and another portion was used to select the accuracy peak. By applying this procedure the median number of selected features across cross validation folds was around 300. We then compared the relative importance of various feature groups after feature selection. We define n_{prior} as the fraction of features belonging to one group before feature selection is applied. For example, $n_{prior} \sim 0.5$ for spectral feature group in Table 3.5. Similarly, we define $n_{selected}$ as the fraction of features from one group after feature selection is applied. We then define η as the ratio of $n_{selected}$ and n_{prior} and it is used to measure the importance given by feature selection algorithm to different feature groups. Figure 3.7 shows η for different long term feature groups. We observe that feature selection procedure selects MFCC, structural and LIWC features with a higher probability compared to their prior distribution. On the other hand, spectral and voice quality features are selected with lower probability compared to their initial distribution. This suggests that majority of acoustic information can be captured by using MFCC features alone and most of the spectral and voice quality features carry redundant information.

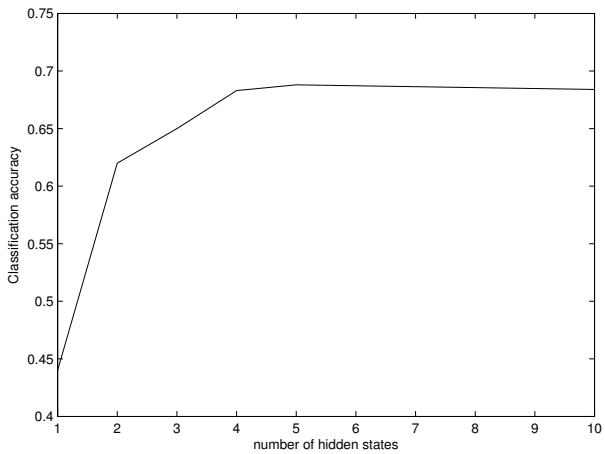

Figure 3.8: Variation in social role recognition accuracy as the number of hidden states is increased in the model.

The latent structure in the turn taking patterns are represented by the hidden states in the model. However, the number of hidden states required to represent the speakers behavior is not obvious. In order to find number of hidden states that best explains the characteristics

of social roles, experiments were performed as this number was varied. The result of this experiment averaged across different crossvalidation folds is shown in Figure 3.8. The model with fewer number of hidden states is not able to capture social role characteristics. The performance saturates around 5 hidden states, while increasing number of hidden states beyond this does not show an increase in performance.

3.5.3 Analysis of classifications results

The baseline system for comparison is based on the method presented in [133]. This system predicts the social roles of speakers from speech activity and fidgeting of each participant in a time window. Since, in all our discussion we have considered information from audio stream alone, for the baseline system too, only audio features were considered. The extracted observation vector in baseline system is composed of speech/non speech activity, as well as, the number of simultaneous speakers in a window of fixed length. The length of the window is a tunable parameter and experiments were performed to find the optimal window length. In [133], a Gaussian RBF kernel support vector machine (SVM) based approach was used for role recognition. SVMs represent the feature vectors as points in a high dimensional space and the algorithm finds a maximum margin separating hyperplane between two classes. For the multiclass classification a one on one strategy was used and each binary classifier was trained using libsvm [22].

In Table 3.7 we compare the performance of baseline classifier with the proposed system. Furthermore, Table 3.7 also shows the performance of proposed approach that simultaneously models both short term and long term speaker characteristics, against systems that only model individual phenomena. The HCRF classifier in [93] is used to model short term features. For long term features we applied linear kernel support vector machine (SVM). The baseline model achieves an accuracy of 64% and the proposed model achieves an accuracy of 74%. The improvement in performance are on all the four role categories. The other two models, HCRF and SVM, show an accuracy which is intermediate between baseline and proposed model. This suggests that joint modeling of multiple features improves performance of social role recognition.

We performed statistical tests to examine the difference between performance of classifiers measured over the same crossvalidation folds. The null hypothesis being tested is that performance of classifiers in Table 3.7 is same and the observed differences are due to random events. We applied Friedman test [26], which is a non parametric method that ranks the performance of each of the classifiers on all crossvalidation folds separately. The classifier which performs best gets rank 1, the second best rank 2, and soon. The average rank of each of the classifiers is used to compute the Friedman statistic, which under null hypothesis is distributed according to F-distribution. For the results in Table 3.7 we reject the null hypothesis ($F(3,63) = 36.7; \alpha = 0.05$). Since the null hypothesis was rejected we performed post hoc (Nemenyi) tests to compare all classifiers with each other. The post hoc tests revealed that

proposed method is statistically significant ($\alpha = 0.05$) compared to both SVM and HCRF.

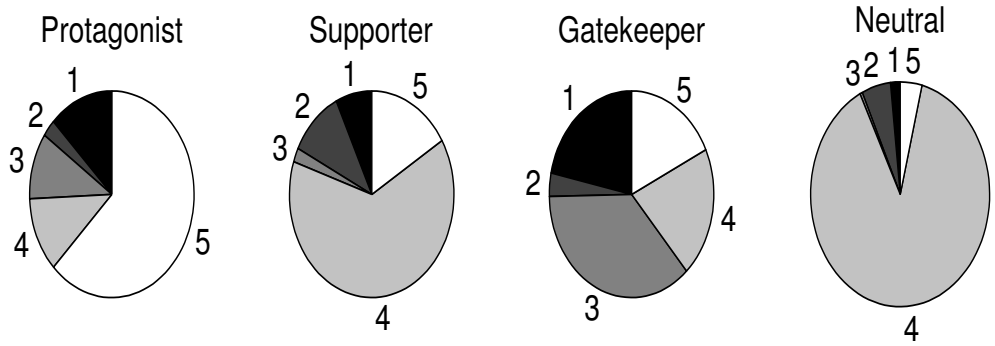

Figure 3.9: Distribution of hidden states learned by the model for each social role category.

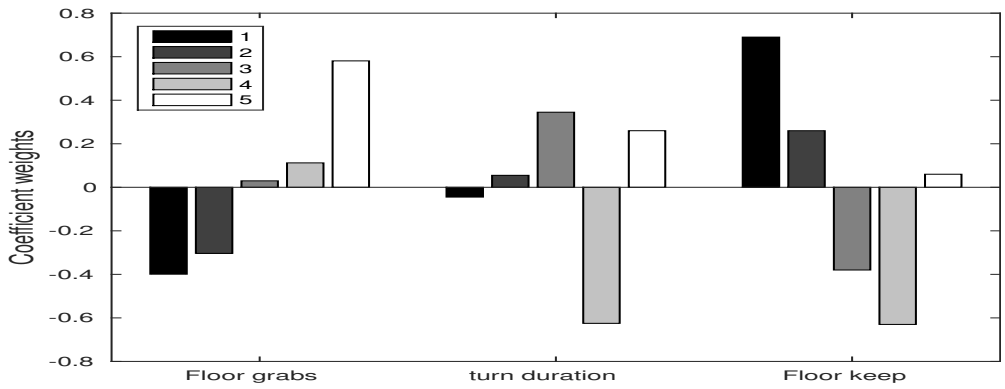

Figure 3.10: Parameter weights α_i corresponding to short term feature functions f_i. The feature functions f_i represent turn taking phenomena, like, floor grabbing, turn duration and floor keeping exhibited by speakers.

The trained CRF model can be used to understand the influence of social roles on the behavior characteristics of the speakers. The parameters of the model, i.e., the hidden states and the weight vector Λ (see 3.6) determine the outcome of the classifier and indicate which features best associate with the raters perception of social roles.

The influence of roles on the turn taking patterns of speakers is determined by the hidden states in the model. Figure 3.9 shows the distribution of hidden states learned by the model for the four role classes. We can observe that while the same hidden states are shared by all the roles, they exploit these hidden states in different proportions. Furthermore, active roles like protagonists and gatekeepers show a relatively more uniform distribution over states compared to neutral speakers.

Figure 3.10 shows the parameter weights $\{\alpha_i\}$ for short term features that were observed after training the classifier. Our analysis considers short term representation of phenomena, such as floor grabbing by a speaker after a silence region or an overlap, duration of speech turns and speaker keeping the conversation floor after an overlap. We observe that features

for floor grabbing have higher weights for the hidden state that is more often associated with protagonists. Furthermore, turn duration features have higher weights for the states corresponding to gatekeepers and protagonists. This is also in line with previous studies [115], where longer turn duration are characteristics of protagonist and gatekeeper speakers. In comparison, the dominant hidden state for neutral has negative weight, which suggests that longer the turn duration, less likely the speaker exhibits a neutral role.

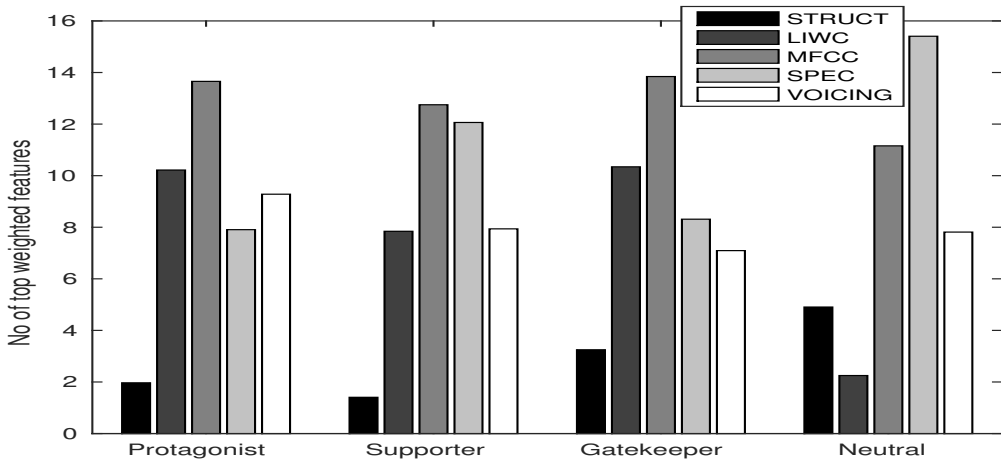

Figure 3.11: Distribution of long term feature groups with largest parameter weights β_i used in predicting each social role.

The relation between social roles and long term features is shown in Figure 3.11. We ranked the long term feature coefficients for each social role label and display the top 15%. We can observe that the feature group distribution is far from uniform and individual social roles exploit various feature groups in different proportions. For supporters and neutrals the acoustic features offer the highest discrimination. In comparison, protagonists and gatekeepers exploit features from both acoustic and LIWC feature groups.

Further analysis revealed that within LIWC features, protagonists have higher weights for processes like causation and inhibition. Gatekeepers have higher weights for positive emotions and social categories. The analysis of " We" words suggests that they are more likely to be used by participants taking the gatekeeper role. This linguistic category is in general associated with feeling of commitment towards the group, as well as maintenance of group longevity [40].

3.5.4 Influence of rater agreement

In this work, the inter-annotator agreement between raters is moderate according to Landis and Koch's criterion. We analyzed the effect of rater agreement on the performance of the learned model. For any instance in the data, we interpret the normalized votes for each role label as the probability of the speaker assuming that role. We compute the label entropy for the instance and use it as the measure of ambiguity associated with the majority label. For instances with a low label entropy, we can infer that the agreement between raters was

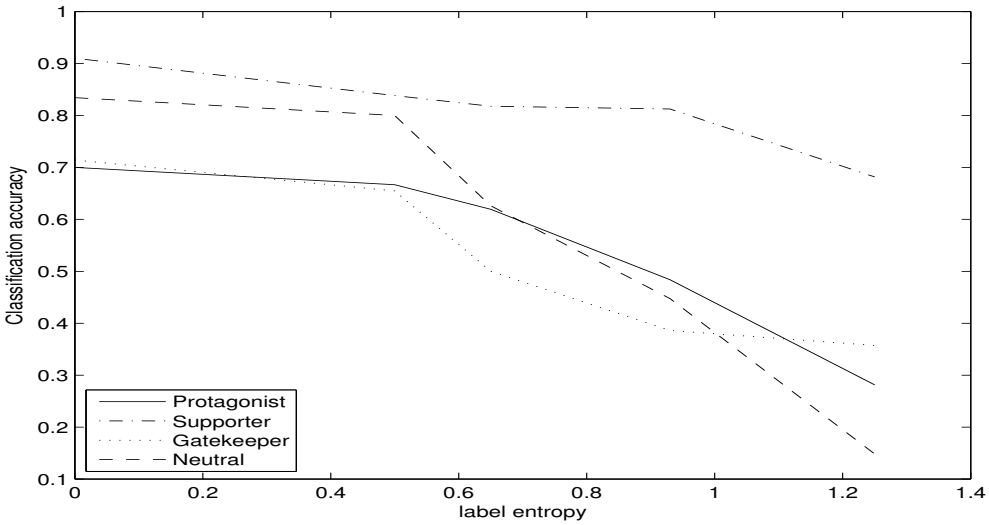

Figure 3.12: Accuracy in recognizing individual role labels as a function of label entropy.

high, while high label entropy instances indicate substantial disagreement between raters. Figure 3.12 shows the classification accuracy for each role as the label entropy is varied. We can observe that the accuracy curves have a negative slope for all social roles. This reveals that the learned model "mimics" the behavior of human annotators in predicting the social role. The instances which where shown to be "hard" for annotators have high label entropy and classification accuracy for those instances tends to be low. On the other hand, labels with low entropy have higher agreement between annotators and the model is likely to predict these instances with higher accuracy.

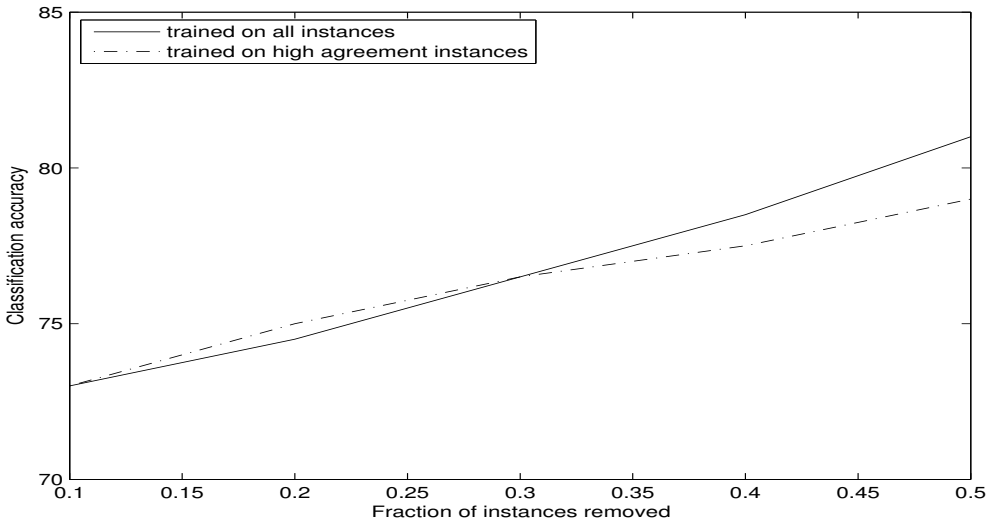

Figure 3.13: Comparison in performance of proposed models when trained on all labeled instances and instances with lower label entropy. In both cases the models are evaluated on low entropy labels.

We next investigated whether classifiers trained only on more confident labels perform better

in comparison to classifiers trained on all instances in the training set. We created various subsets of the corpus by removing increasing proportion of instances with high label entropy. Using crossvalidation we trained new classifiers for each subset of corpus and evaluated their performance. For the same subsets we also evaluated the classifiers trained on all instances in the training set. Figure 3.13 compares the performance of the two cases. We can observe that classifiers trained on all instances do not perform significantly worse than classifiers trained on more confident labels. On the other hand, when half of the labeled instances are removed, the former performs better than the latter. This suggests that proposed classification method is robust against the effect of label noise.

3.5.5 Evaluation on AMI natural meetings

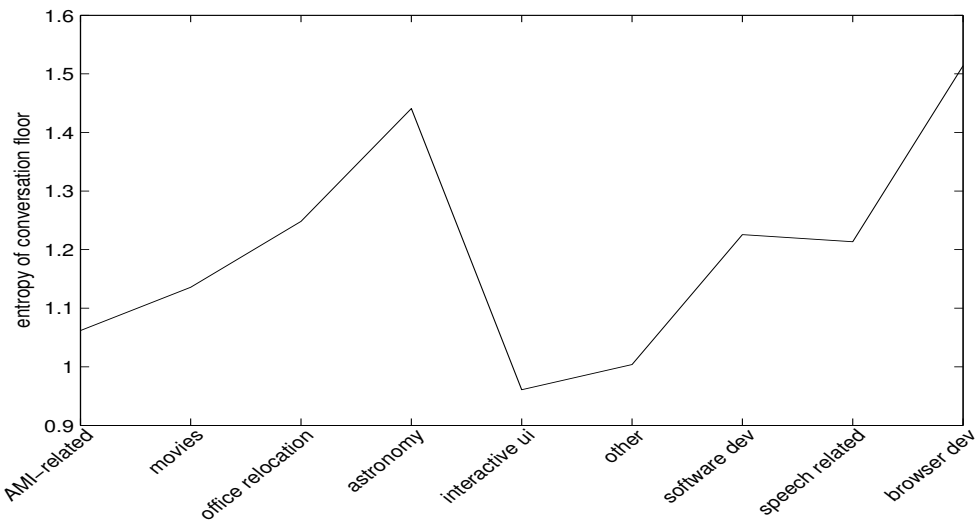

Figure 3.14: Average conversation floor entropy for various scenarios in natural meetings.

In order to investigate the performance of the proposed method on other scenarios of small group interaction, we performed role recognition experiments on the set of natural meetings in AMI corpus. This set includes natural meetings on topics such as speech processing, as well as planning for a fictitious movie club, or office relocation. Compared to scenario portion of the corpus, in natural meetings the participants do not perform roles specific to an organizational system. Moreover, the participants discuss a wide range of topics and the language used is also more diverse and complex.

For this study we annotated almost 5 hours of data from the non scenario portion of the corpus using the procedure described in Section 3.2. All the annotated meetings do not have the same number of participants. While the number of participants in scenario meetings was fixed to four, for natural meetings the participant number can vary between three and four. In terms of speakers gender, we observe that natural meetings have a slightly higher male distribution (70%) compared to scenario meetings (65%).

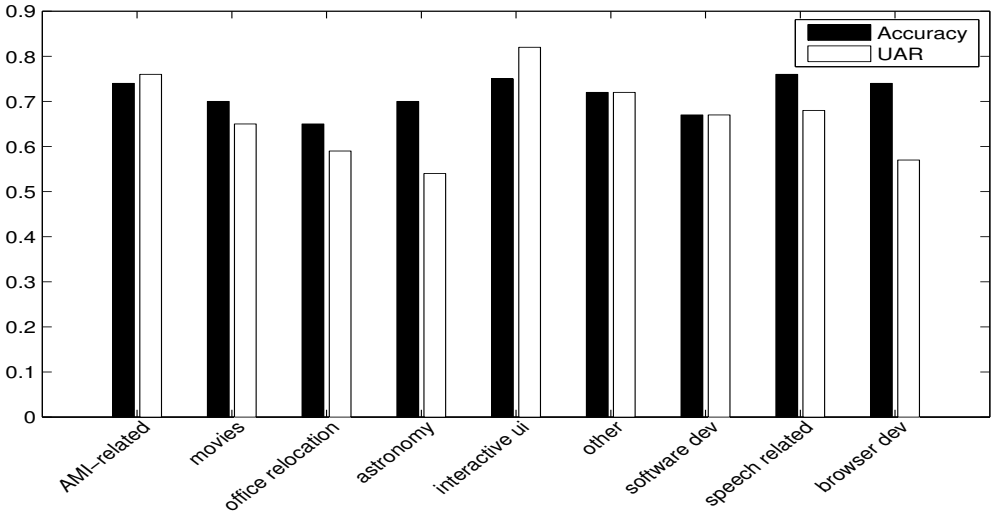

Figure 3.15: Role recognition accuracy and UAR for various scenarios in natural meetings.

We also compared the conversation characteristics of natural meetings against AMI scenario meetings. Our analysis considers the distribution of conversation floor between meeting participants. We interpret the fraction of time each participant is speaking in the meeting slice as the participant's probability of holding the conversation floor. The conversation floor entropy is computed from these probabilities. A high value of floor entropy corresponds to equal participation by speakers and a lower value suggest that conversation is dominated by fewer speakers. In Figure 3.14, we plot the average conversation floor entropy for various topics in natural meetings. The AMI scenario meetings have an average floor entropy equal to 0.92. In comparison, we observe that the natural meetings in general have higher floor entropy, and there is lot of variation between different topics.

We trained the CRF model on scenario portion of the corpus and evaluated the generalization performance on the natural meetings. In order to ensure speaker independent recognition of social roles, the evaluation was done for speakers not present in training data. The trained model achieved a significantly higher recognition accuracy (72%) compared to chance level (39%). This shows that the proposed method learns the relationship between social roles and behavioral cues that are likely to be exhibited in small group interaction.

Since natural meetings cover a range of topics, we evaluated the role recognition performance individually for each topic. To make the comparison independent of the distribution of social roles in different topics, we also measure the performance in terms of unweighted average recall (UAR). The results are shown in Figure 3.15. We observe that role recognition accuracy is higher than chance level and most topics achieve an accuracy of over 70%. Also, for most topics, UAR is quite close to accuracy. However, some natural meetings that include discussion on topics like astronomy and browser development, show higher difference between UAR and accuracy. Our analysis revealed that the observed difference is due to lower recall for protago-

nists and gatekeepers. Furthermore, these topics also have higher than average conversation floor entropy (see Figure 3.14). This suggests that active speakers in these meetings do not exhibit the dominant characteristics associated with these social roles.

3.6 Conclusion

In this chapter we presented an approach for automatic recognition of social roles that emerge in small group meetings. We investigated various short term and long term features for recognition of social roles. Furthermore, we also analyzed the influence of annotator variability and the supervised learning models on role recognition task. The present work has been performed over the largest annotated database for this task, both in terms of number of unique speakers and number of instances. The main conclusions of this chapter are discussed now.

Our analysis revealed that automatically extracted short term and long term features are useful cues for predicting social roles. Experiment results also reveal that combining feature information at multiple time scales in a single representation increases the predictive capabilities of the automatic recognition system. The CRF classifier was able to perform non trivial classification of four social roles, reaching a recognition accuracy of 74% on the scenario portion of AMI corpus.

The social role labels investigated in this chapter are derived from the subjective perception of human raters. Our analysis suggested a relation between classification performance and the variability in perception of different raters. This seems to be a major source of errors in our work, since the accuracy of the recognition system systematically improves, when evaluations are performed on subset of data with higher agreement between raters. While the variability due to subjective experiences of the online raters may be considered as a limitation, the other alternative is time consuming and expensive training of raters, who are then more likely to agree in their judgments.

Our final investigation evaluates the generalized performance of the proposed approach on various scenarios of multiparty interaction. Using the CRF model trained on scenario portion of AMI corpus, we evaluated the social role recognition performance on various topics in natural meetings. Experiments show that the proposed approach reaches a recognition accuracy of (72%) on natural meetings, which is slightly lower than accuracy (74%) observed for AMI scenario meetings. Although, further research on other corpora are needed to reach definite conclusions, results suggest that the our approach captures the influence of social roles on behavioral patterns of speakers in small group interactions.

While this chapter focused on automatic recognition of social roles in a meeting. Our analysis did not consider the influence of formal roles on behavior of participants. It would be interesting to analyze whether the model and features described in this chapter can also be applied for recognition of formal roles. In the next chapter, we consider recognition of formal roles

using the techniques developed in this chapter.

4 Recognition of formal roles in multi-party interactions

4.1 Introduction

In this chapter, we investigate the problem of identifying formal roles of participants in meeting recordings. As described in Chapter 1, formal role is a designated position that is directly assigned by an organization or a group. In comparison to social roles, formal role of a participant is decided prior to start of a meeting and remains fixed throughout the conversation.

In this chapter, we compare the recognition of formal roles and social roles in meetings using the same set of features. Additionally, we also investigate whether speaker behavior captured from relatively thin meeting slices is sufficient to predict formal roles. While recent studies have proposed automatic models for recognition of formal roles [60, 37], none of those attempts compared recognition of formal and social roles using the same model and features. Our findings show that compared to social roles, non verbal features are weakly influenced by formal roles. On the other hand, verbal features provide the best performance in recognition of formal roles.

We also present an unsupervised model for extracting verbal features from unlabeled data using the framework of latent topic models The probabilistic topic models were trained on speech utterances and we use the estimated latent topic distributions to infer formal roles. Previous studies [85] have applied latent topic models for discourse segmentation, but this is first work to the best of our knowledge that applies latent topic modeling to recognize formal roles. We show that a compact representation of latent topics predicts formal roles in AMI meetings and the latent topics can be used to identify characteristic words used by different formal roles.

In the remainder of this chapter, Section 4.2 describes the corpus and formal roles. In Section 4.3, we compare recognition of formal and social roles using the same data and features. Sections 4.4 and 4.5 provide details of unsupervised feature extraction and classification methodology. The experiments and results are presented in Section 4.6. The chapter is then

concluded in Section 4.7.

4.2 Meeting corpus

In this chapter we investigate the formal roles exhibited by participants in the scenario meetings of AMI corpus. In the scenario meetings, participants are tasked with designing and marketing a new remote control. Each design team passes through four stages, including kick-off meetings, where participants become acquainted with the task; functional design meetings, where participants discuss user requirements and functional design of the remote control; conceptual design meetings involve discussion about component specification, user interface and materials to be used in the remote control; detailed design meetings, where look and feel of the remote control is finalized and results are evaluated. In a design team, each participant acts according to one and only one of the predefined roles described next.

- *Project Manager* (PM): A participant who coordinates the project and is responsible for keeping the project within time and budget limits.
- *Marketing Expert* (ME): A participant who determines user requirements and market trends and is responsible for evaluating the prototype.
- *User Interface Designer* (UI): A participant who is responsible for user interface and technical details of the remote control.
- *Industrial Designer* (ID): A participant who is responsible for internal working and components of the remote control.

4.3 Comparison of formal roles and social roles

This section describes the procedure followed to recognize formal roles in AMI meetings. Our first objective is to investigate how similar or different, is the recognition of formal roles compared to social roles. To make this comparison we consider two cases. First case: formal roles, similar to social roles can be predicted from speakers behavior in a meeting slice and using the same automatically extracted feature groups. Second case: formal role unlike social roles are fixed for a speaker for the entire duration of the meeting. Therefore, speakers behavior over the full span of a meeting should be considered during feature extraction.

4.3.1 Features used

The different feature groups used for formal role recognition are described now.

Linguistic Features : As in Chapter 3, the linguistic features were the scores generated by LIWC text processing module. These features were computed for two different contexts. In the

first context, speaker utterances corresponding to each meeting slice were processed to yield linguistic features. These features are same as those considered in Section 3.3.2. In the second context, we concatenated each speaker's utterances for the entire meeting and processed those using LIWC to generate meeting level linguistic features.

Acoustic features : Table 3.3 and Table 3.4 in Chapter 3 show the various acoustic features, including F0, speech energy, voice quality, MFCC and spectral features extracted from IHM channels. Acoustic features for each speaker in a meeting slice were extracted using the two pass strategy described in Section 3.3.2. Besides slice level features, a second set of acoustic features was computed from the entire meeting. First, low level descriptors (LLDs) were extracted from all speaker utterances. Second, statistical and regression functionals were used to project each LLD contour into feature vectors.

Structural features : Both meeting level and slice level structural features were extracted from speech data. Slice level structural features are the same as described in Section 3.3.2. Structural features extracted per speaker at the meeting level include the total speaking time, duration of overlapping speech, number of speaker turns in a meeting. Also included were statistics like maximum, minimum and mean and standard deviation for these features.

4.3.2 Experimental evaluation

For formal role recognition we repeated the same experimental protocol described in Chapter 3. We split the dataset into $k = 22$ crossvalidation sets, k-1 sets were used for training and remaining one set was used for testing. The procedure is repeated 22 times and each time a different set is left out for testing. Each set comprises of a group of speakers who participate together in a meeting. The partitioning of data into different sets was performed to maintain strict separation between training and test sets in terms of speaker identity. The performance was measured in terms of overall formal role recognition accuracy.

Table 4.1: Formal role recognition performance for different long term feature groups extracted in a meeting slice. For comparison with social roles, the last column in the table repeats the accuracy numbers for social roles previously detailed in Table 3.5

Feature Group	Number of features	Accuracy (Formal)	Accuracy (Social)
Voice quality	308	0.28	0.60
MFCC	200	0.31	0.61
spectral	748	0.31	0.60
structural	35	0.32	0.62
LIWC	60	0.33	0.59
ALL long term	1343	0.32	0.66

Our first experiment is a direct comparison between automatic recognition of formal and social roles. The long term feature groups and recognition model are the same as described

in Chapter 3. In Table 4.1, we show the different long term feature groups, the number of features within each group and their classification accuracy for formal and social roles. Table numbers reveal that there is significant difference in recognition accuracy between formal roles and social roles when feature extraction and modeling is done for each meeting slice. The difference in performance is consistent across structural, linguistic and acoustic feature groups. Furthermore, improvement in accuracy due to feature combination is evident for social roles, but does not ensue in the case of formal roles. Our analysis suggest that predicting formal role from "thin slices of behavior" is much harder compared to social roles.

Table 4.2: Formal role recognition performance for different feature groups extracted over the entire meeting.

LIWC	acoustic	structural
0.60	0.38	0.44

Our first investigation revealed that automatic recognition of formal roles is a challenging task. However, it is not yet clear whether the difficulty in predicting formal roles is due uninformative features or due to extraction of these features over relatively short meeting slices. We note that formal role of a speaker is fixed for the entire meeting. Therefore, for our second investigation we extract various features over the entire meeting. Table 4.2 reports the performance of the various structural, acoustic and linguistic features for recognizing the formal roles at the meeting level. We notice that in this context their is overall improvement in recognition performance for both structural and linguistic features. The best performing feature group consists of linguistic features which show an accuracy of 60%. There is also a slight improvement when structural and LIWC features are combined 61%. These results suggest that language used by participants in AMI meetings is influenced by their formal roles. In the next section, we explore informative features which infer the formal roles of speakers from the verbal content of conversation.

4.4 Features extracted from verbal content

4.4.1 Latent topic model

Topic models are probabilistic generative models that have been extensively used for natural language processing. In Latent Dirichlet allocation (LDA) [16], the corpus is generated from a fixed underlying mixture of K topics, and each topic is modeled as a multinomial distribution over all the possible words. The latent topics discover patterns based on the co-occurrence of words in the documents.

Let D be the set of documents in the corpus and V be the set of unique words. Each document d is represented by a bag of N_d words chosen from V. We also assume a fixed number of topics K for the entire corpus. In LDA, the word distribution for a given document is represented

as a mixture of K topics $P(w) = \sum_{k=1}^{K} P(w|z = k)P(z = k)$, where z is the latent variable from which the word w is drawn. The distribution of words conditioned on z is given by a multinomial $P(w|z = k) = \phi_k(w)$ and the latent variable z in a document d is also sampled from a multinomial distribution $p(z = k) = \theta_d(k)$. In LDA, the variables ϕ_k and θ_d are Dirichlet distributed with hyperparameters α and β respectively. To infer the latent variables of the model, Gibbs sampling [64] is applied by sequentially drawing samples for each latent variable while all the variables are fixed.

For each participant in a given meeting, we extracted the spoken words from speech transcripts generated using output of AMI-ASR system [45]. For the task of speaker role recognition, each document d is represented by all the words uttered by a single speaker S during a meeting. Every participant in a meeting has to perform a separate function defined by its role. We hypothesize that the function of a formal role influences the distribution of latent topics used by the speaker to generate the spoken content. We extract the latent topic distribution θ_S by applying LDA on the utterances of a speaker S.

4.4.2 Dialog act tags

In addition to automatic feature extraction using LDA, we also explored a fixed set of dialog act (DA) tags for formal role recognition. Dialog acts aim at capturing the speaker's intention in the discussion. AMI corpus is annotated in terms of several DA classes which includes minor acts (*Backchannel, Stall, Fragment*), acts about information exchange (*Inform, Elicit-Inform*), acts about possible actions (*Suggest, Offer, Elicit-offer*), acts on commenting (*Assess, Comment, Elicit-Asses, Elicit-Comment*) and also social acts (*Be-positive, Be-negative*). We investigate the per-speaker DA counts as features for formal role recognition. Figure 4.1 plots the histogram of the most common Dialog acts tags.

4.5 Classification approach

We consider formal role recognition as a machine learning problems that consist of finding a stochastic mapping from a set of features to a set of class labels. For the task of formal role recognition, the classes correspond to roles in the set $\mathscr{FR} = \{PM, ME, UI, ID\}$.

We applied a support vector machine (SVM) classifier to predict the formal roles. SVM considers each feature vector as a point in multidimensional feature space and the algorithm works by constructing a separating hyperplane between two classes. For the multiclass classification required for role recognition a one on one strategy was used and each binary classifier was trained using libsvm [22].

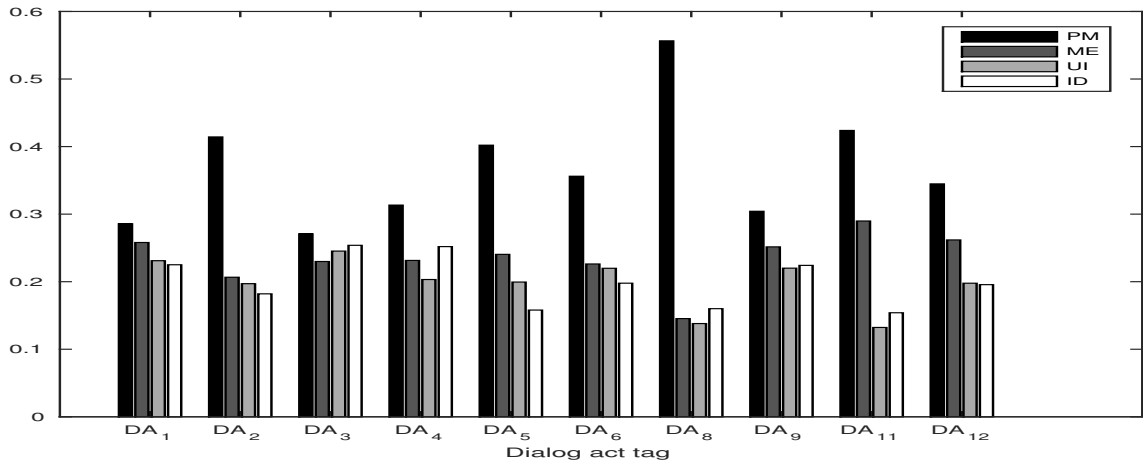

Figure 4.1: Normalized DA tag distribution on data annotated for formal roles for the most common DA tags. DA_1 (Backchannel), DA_2 (Stall), DA_3 (Fragment), DA_4 (Inform), DA_5 (Elicit-Inform), DA_6 (Suggest), DA_8 (Elicit-Offer), DA_9 (Assess), DA_10 (Elicit-Assessment), DA_11 (Comment).

4.6 Experiments

Evaluation experiments were conducted using 35 fold cross-validation wherein one set of meetings (all but four meetings that have the same set of speaker identities) was kept for training/tuning the model parameters while a distinct set (remaining four meetings) was used for evaluation. We used the linear kernel for SVM classifier and the model parameters were selected from a subset of training data.

The posterior distribution over documents and topics, as well as, the hyperparameters of the LDA model were estimated using Gibbs sampling (a Markov chain Monte Carlo method). After a burn in period, the sampling procedure ultimately results in a stationary distribution which corresponds to the topic distribution. For our experiments, we used the mallet implementation [67] of LDA, with symmetric Dirichlet priors.

For evaluation of automatic role recognition, since each CV fold has the same distribution of classes (their being one to one mapping between speakers and roles for each meeting), we use the recognition accuracy as the metric of recognition performance.

Our first experiment evaluates the influence of number of latent topics K on the extracted unsupervised features. Figure 4.2, shows the variation in accuracy for different choices of $K = \{5, 10, 20, 50, 70, 100\}$. The models with fewer number of latent topics are not able to capture all the role information. However, we observe a significant increase in performance with only $K = 20$ topics. Increasing the size of feature set after this does not reveal any increase in performance.

We also considered whether removing stop words during training and evaluation can yield unsupervised features that better capture the functional content of speaker role. A list of stop

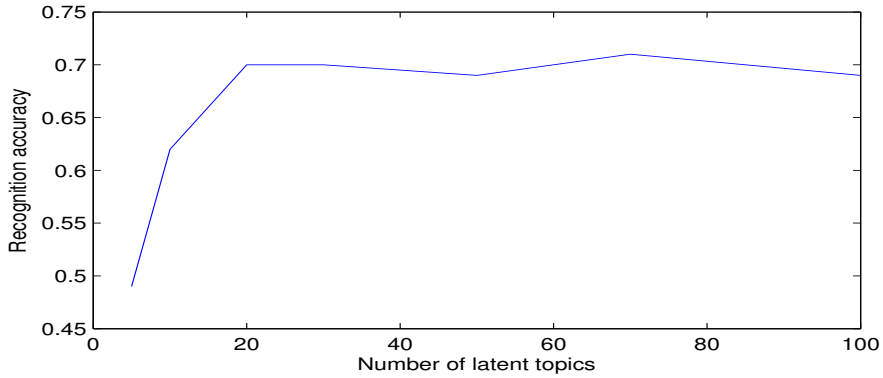

Figure 4.2: Variation of role recognition accuracy as the number of latent topics K in LDA is varied.

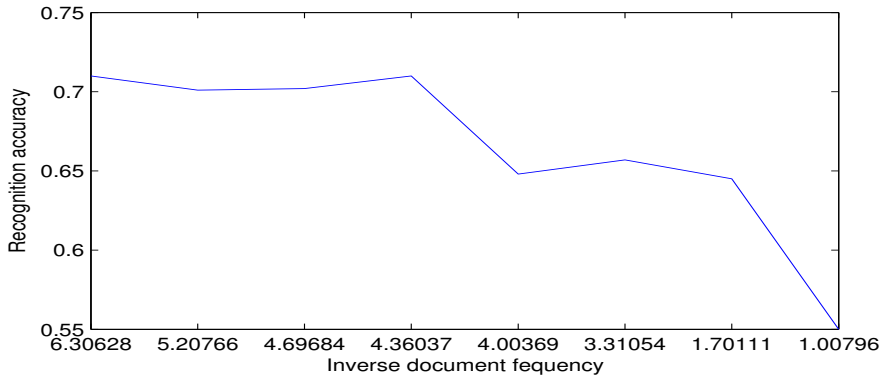

Figure 4.3: Effect of stop words on role recognition accuracy. Horizontal axis shows different values of IDF.

words was prepared based on their inverse document frequency (IDF) scores. We removed the words with low IDF scores, and trained the LDA models on the processed documents. The influence of stop words on the role recognition accuracy is shown in figure 4.3. The plot reveals, that removing stop words (for moderate IDF scores) does not significantly increase the performance of the models. However, performance drops as IDF scores increase, showing that a there exists a limiting size of vocabulary that is needed to express the functions of roles in the corpus.

Table 4.3: Per role accuracy obtained in recognizing roles for various classification approaches

	Overall and per-role Accuracy				
Features	PM	ME	UI	ID	All
LDA	0.84	0.72	0.64	0.55	0.69
dialog acts	0.7	0.4	0.35	0.45	0.48
lex+struct	0.88	0.72	0.62	0.57	0.70
LDA+lex+struct	0.90	0.74	0.69	0.59	0.73

In Table 4.3, we report the accuracy of formal roles for different features. LDA denotes the latent topic features for $K = 20$. We also report results obtained by using combination of raw

word unigrams (denoted by lex) and structural features (denoted by struct). It can be noticed that all the different models in Table 4.3, achieve highest accuracy in labeling PM. Table 4.3 shows that the formal role recognition results obtained using LDA features are better than those obtained using dialog act tags. We also observe that even with a compact latent topic representation (K=20), LDA features perform comparably to a much larger set of lexical and structural features. Furthermore, when the two feature sets are combined the performance increases, suggesting that LDA modeling captures some complementary information in data. The results obtained using LDA modeling also show comparable performance to previous reported in literature [37]. In [37], formal role recognition in AMI meetings was performed using a combination of Social Network Analysis (SNA) and lexical features. They report an role recognition accuracy of 68% (correctly labeled speaking time) . However, it should be notes that the comparison is not entirely fair because in [37] the accuracy is measured in terms of correctly labeled speaking time, while we measure accuracy in terms of correctly labeled speakers.

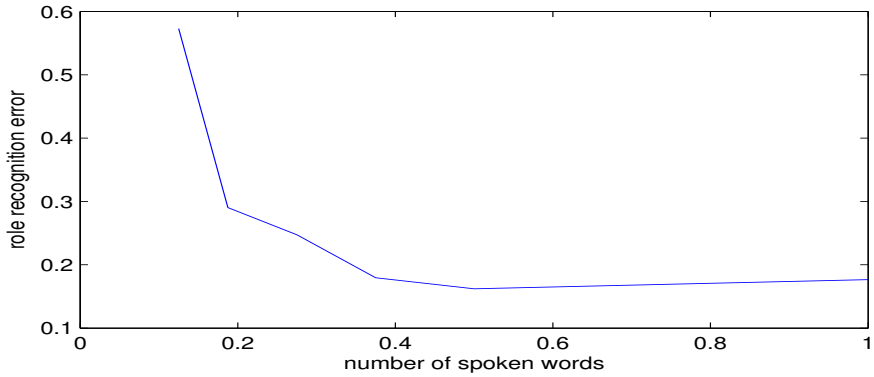

Figure 4.4: Analysis of role recognition errors with respect to number of spoken words in a document.

We can also observe that performance across roles is not uniform. The role of PM is recognized much better by all the models, compared to other roles. An analysis of errors by the classifier revealed that there is a systematic difference in performance that is related to the length of speakers utterances. We observed that average number of words spoken by PM in most meetings was greater than other participants. On the other hand, the design related roles, UI and ID, tend to speak lesser number of words, averaged across the meetings. In Figure 4.4, we plot the observed errors as a function of number of spoken utterances by different speakers. The plot reveals a clear pattern with error decreasing as the length of conversation increases. This explains the higher accuracy for PM, who speaks more often then other formal roles.

Table 4.4 shows the top words associated with LDA topics which are most correlated with various formal roles. The analysis of LDA topics and different roles suggests that top words in these latent topics capture functions of formal role. This shows that unsupervised feature extraction using LDA is effective in automatically clustering word patterns for different roles.

Table 4.4: Top words in latent topics that are most correlated with role labels.

PM	ME
okay meeting design yes will minutes project not what going all your would work uhhuh new but yeah one two three	like spongy fruit remote control important fancy shape banana fashion look feel easy trends maybe innovative remote and people five they
UI	**ID**
buttons then but channel yellow need use see recognition rubber just functions all use shape television one should yes colour volume channel	chip components which infrared titanium signal energy will button design source battery interface circuit working basically power rubber scroll button curved plastic

4.7 Conclusions

This chapter extensively investigated and compared, on the same dataset, how various feature groups perform in the task of labeling formal roles and social roles. AMI scenario meetings are labeled according to both role coding schemes, where the scenario imposes constraints on the participant formal roles during a professional meeting, while speakers spontaneously interact taking on different social roles.

Our analysis revealed that in comparison to social roles, speaker behavior extracted from relatively thin meeting slices is insufficient to predict formal roles. Furthermore, even when feature were extracted from the entire audio recording, verbal features alone were the best predictors of formal roles.

We also demonstrated the effectiveness of unsupervised feature modeling using LDA for detecting formal roles of speakers. By applying LDA on speech transcripts we extracted latent topic features that were able to perform non trivial classification of four formal roles, reaching a recognition accuracy of 69%.

Despite the fact that same features and models were used to predict both formal role and social roles, our investigation revealed that formal roles have weaker relation with nonverbal behavior of speakers. In comparison, social roles have much stronger relation with nonverbal features. In the next chapter, we consider the influence of social roles on turn taking behavior of speakers and investigate whether social roles can be used to improve speaker diarization.

5 Improving speaker diarization using social roles

5.1 Introduction

Speaker Diarization aims at inferring "who" spoke "when" in an audio stream. Most of the advances in this domain have been due to signal processing techniques for enhancing speech signals [102], and statistical modeling of acoustic information [55]. However, speech data used in diarization are instances of multiparty conversations which follow predictable patterns of interaction between participants.

This chapter investigates whether speaker diarization can be improved by modeling the influence of social roles on the interaction patterns of speakers. In particular, we examine two limitations of current state-of-the-art systems [125], i.e., a speaker independent minimum duration constraint and a uniform prior on speaker interaction patterns. Previous studies [56, 117] have shown that speaker sequence modeling can be used to improve diarization. In [56], interaction patterns between speakers where estimated on a per meeting basis and did not consider any role information. In [117], formal roles of speakers were used to estimate the probability of interaction patterns between speakers. However, a limitation of this approach is that formal roles are imposed by specific scenario of a meeting and may not generalize across multiple corpora. In comparison, our analysis in Section 3.5.5 revealed that social roles generalize across different scenarios of multiparty interaction.

Our work extends the standard speaker diarization system based on HMM-GMM modeling [125], by including social role information in the speaker segmentation step. In this chapter, social roles are used to extract prior information about the expected duration of speaker's turn and social role n-grams are used to model speaker sequence distribution. Our analysis is conducted on AMI meetings using the social role annotation describe in Chapter 3. We also show that social role statistics estimated on AMI meetings can be applied on meetings from the Rich Transcription (RT) dataset.

The chapter is organized as follows: Section 5.2 describes the dataset and social roles, Section 5.3 describes the baseline diarization system and Section 5.4 presents the details of social

role based speaker diarization. Section 5.5 describes experiments on the AMI meetings and RT meetings. The chapter is finally concluded in Section 5.6.

5.2 Data description

Let us consider a meeting in which speakers are represented by set $\mathscr{S} = \{S^1, S^2, ...\}$ who assume social roles in set $\mathscr{R}$. The meeting is segmented as a sequence of N turns represented by set $\{(t_1, d_1, s_1), (t_2, d_2, s_2), ..., (t_N, d_N, s_N)\}$, where each turn t_n has a duration d_n and speaker label $s_n \in \mathscr{S}$ associated with it. In this chapter, we consider each turn as the uninterrupted speech by a single speaker. We note that this definition of turn ignores overlapping speech regions. We can use the procedure described in Chapter 3, to associate a social role label for each speaker at each turn.

In this chapter, we apply a reduced version of supervised role recognition model. The lexical information used in Chapter 3 was obtained from speech transcripts generated by ASR system that was trained on audio data recorded from headset microphones. However, in this chapter we consider the more challenging distant microphone audio recordings for which ASR system of Chapter 3 is not applicable. Furthermore, the output of current state-of-the-art speaker diarization systems are unable to identify overlapping speech. Therefore, in this chapter we apply a reduced automatic role recognition model that is trained using combination of structural and acoustic features. To simplify notation used in this chapter, the process of applying the social role recognition model on extracted feature vectors that represents a speaker's behavior in a meeting slice is abstracted as a speaker to social role mapping function Υ. The function $\Upsilon(s_n) -> R_n$ maps a speaker label s_n to a social role $R_n \in \mathscr{R}$. In the next sections, we explain how social role information can be used to improve segmentation of meeting recordings into sequences of speaker turns.

5.3 Baseline speaker diarization

Conventional diarization systems are based on agglomerative clustering framework where each speaker is modeled as a HMM state and each state distribution is modeled using a GMM. The baseline system [125] used in this work achieved state-of-the-art performance in several NIST evaluations. It is initialized by uniformly segmenting a given audio recording into segments representing initial speaker clusters. The number of initial clusters is much higher than the actual number of speakers in the recording. The algorithm iteratively merges the closest clusters until a stopping criterion is met. After each merge, speaker boundaries are realigned based on the estimated speaker models using Viterbi decoding. For the merging and clustering stopping criteria, a modified version of Bayesian Information Criterion (BIC) is used [2].

In the baseline system, each speaker cluster is associated with a HMM model. We denote by $\mathcal{S} = \{\mathcal{S}^1, ..., \mathcal{S}^L\}$, the set of L models, where $\mathcal{S}^l$ is the model associated with speaker S^l. The

individual speaker models can be concatenated to represent a hypothesized speaker turn taking sequence $\{(t_1, d_1, s_1), ..., (t_N, d_N, s_N)\}$. Here N is number of speaker turns and t_n is the nth turn with a duration d_n. $s_n \in \{S^1, ..., S^L\}$ is the speaker label associated with turn n. The speaker sequence representing the turn taking is denoted by $\mathbf{S}_{seq} = \{s_1, ..., s_N\}$.

Consider a acoustic data (MFCC) sequence $\mathbf{X} = \{x_1, x_2, ..., x_{\tilde{N}}\}$ made of $\tilde{N}$ speech frames. The re-estimation step in speaker diarization system aims at finding the optimal speaker sequence $\hat{\mathbf{S}}_{seq}$ that maximizes the $P(\mathbf{S}_{seq}|\mathbf{X})$.

$$\hat{\mathbf{S}}_{seq} = \underset{\mathbf{S}_{seq}}{\arg\max}\, P(\mathbf{S}_{seq}|\mathbf{X}) \tag{5.1}$$

$$\hat{\mathbf{S}}_{seq} = \underset{\mathbf{S}_{seq}}{\arg\max}\, P(\mathbf{X}|\mathbf{S}_{seq})\, P(\mathbf{S}_{seq}) \tag{5.2}$$

Here $P(\mathbf{S}_{seq})$ represents the prior probability of a specific speaker sequence and $P(\mathbf{X}|\mathbf{S}_{seq})$ represents the likelihood of the acoustic data. A minimum duration constraint is imposed by associating a fixed number of states D with each speaker model $\mathcal{S}^l$ and all the states have the same emission probability modeled with a GMM.

5.4 Social roles based speaker diarization

One of the assumptions of diarization systems is that minimum duration D (2.5 seconds in the baseline system), is independent of conversation dynamics, and is fixed for all speakers across all the recordings. Furthermore, a uniform prior on speaker sequence distribution $P(\mathbf{S}_{seq})$ is imposed such that transitions between speakers are equally likely. In this section, we further investigate the validity of these assumptions.

5.4.1 Minimum duration model

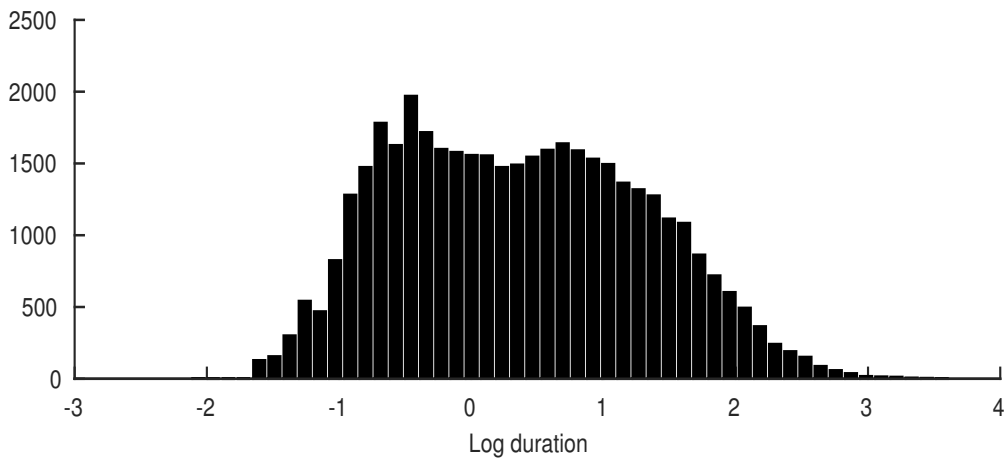

Figure 5.1: Histogram of log duration of speaker turns in AMI corpus meetings.

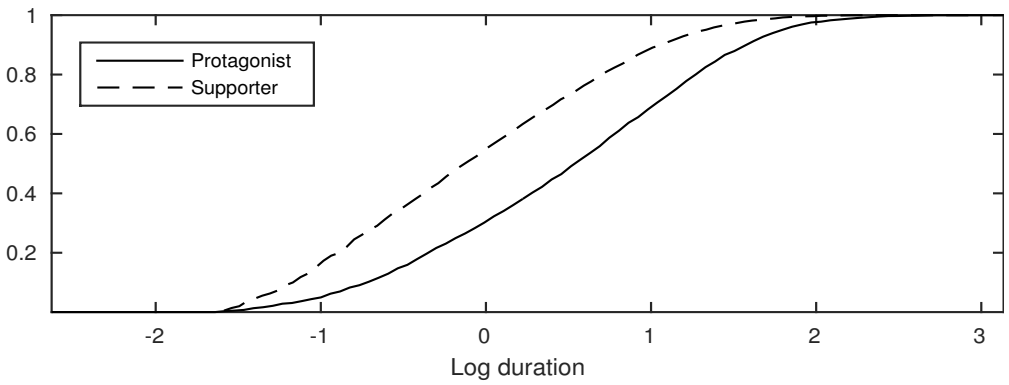

Figure 5.2: Cumulative histogram of log duration of speaker turns for protagonist and supporter roles.

Figure 5.1, shows the histogram of log duration of speaker turns in the IHM audio labeled data (see Chapter 3) . We can see that distribution of speaker turns is not unimodal and speakers are likely to take shorter turns (less than 1 second) as well as longer turns (greater than 2.5 seconds). Our analysis in Chapter 3 showed that turn duration of speakers is influenced by social roles they assume.

To observe this behavior, we plot the cumulative histograms of speaker turn data (IHM conditions) for protagonists and supporter roles as shown is Figure 5.2. The figure reveals that protagonists are more likely to produce longer turns ($> 2.5 seconds$) compared to supporters. Further analysis reveals that for protagonists, 71% of speech time is associated with longer turns, while for supporters this percentage is only 53%. Furthermore, for protagonists, less than 7% of speech time is due to short turns while for supporters it is 21%.

This analysis leads us to propose a minimum duration constraint based on social role of speakers, i.e, minimum duration D^l for speaker model $\mathcal{S}^l$, that is fixed in baseline diarization system, is now made a function of speaker's social role $D^l(R)$. The durations specified by $D^l(R)$ were selected by fitting a probability distribution over the log turn duration for each social role in training data. While earlier studies, such as [41] have assumed a log normal distribution for turn duration, we modeled the log duration of turns using a GMM. Note that log normal distribution is equivalent to a GMM with a single mixture component.

In this work, the number of mixture components was selected based on BIC. Our analysis revealed that the model with two mixture components best explains the variability of turn duration. The mean parameters of the GMM were estimated using maximum likelihood criterion. The minimum duration $D^l(R,c)$ for each speaker is represented as a function of social role and mixture component $c \in \{1,2\}$. $D^l(R,1)$ is the expected duration of short turns while $D^l(R,2)$ is the expected duration of longer turns. The speaker model $\mathcal{S}^l(R,1)$ is formed by concatenation of $D^l(R,1)$ states and speaker model $\mathcal{S}^l(R,2)$ is formed by concatenation of $D^l(R,2)$ states. The models, $\mathcal{S}^l(R,1)$ and $\mathcal{S}^l(R,2)$ have the same emission probability. However, they represent short turns and long turns produced during participant interaction.

5.4.2 Speaker interaction model

Traditional diarization systems also assume a uniform prior on speaker sequence distribution $P(\mathbf{S}_{seq})$, considering all turn taking transitions between speakers equally likely. However, the knowledge of social roles can be used to impose more meaningful information on speaker sequence distribution. As an analogy with ASR systems, we propose a social role n-gram model, that represents the interaction between speakers in terms of their social roles.

Table 5.1: Perplexity of social role sequences for AMI meetings.

	Unigram	Bigram	Trigram
Perplexity	3.4	3.0	2.4

Under a Markov assumption, we can factor the speaker sequence distribution in terms of

$$P(\mathbf{S}_{seq}) = P(s_1, .., s_N) = P(s_1, .., s_p) \prod_{p+1}^{N} P(s_n | s_{n-1}, ..., s_{n-p}) \tag{5.3}$$

Using the mapping $\Upsilon(s_n) -> R_n$, the above can be expressed as:

$$P(\mathbf{S}_{seq}) = P(\mathbf{R}_{seq}) = P(R_1, .., R_p) \prod_{p+1}^{N} P(R_n | R_{n-1}, ..., R_{n-p}) \tag{5.4}$$

The quality of language models in ASR systems is evaluated by computing the perplexity on separate data set. We use the same criterion to select an appropriate language model for conversational turn taking. We calculated the unigram, bigram and trigram estimates on the training data. The observed perplexity on an independent development set is reported in Table 5.1.

Table 5.1 reveals a drop in perplexity when moving from unigram to bigram models and their is further decrease in perplexity for trigram model. This reveals that social role of current speaker conditioned on the social roles of previous speakers produces a large reduction in speaker sequence perplexity. The most probable n-grams correspond to protagonists and gatekeepers and their interaction with supporters. For the rest of this study, we have used trigram models as they showed the best performance. Given the n-gram social role model, we propose a modified speaker re-estimation step,

$$\hat{\mathbf{S}}_{seq} = \underset{\mathbf{S}_{seq}}{\arg\max}\, P(\mathbf{X}|\mathbf{S}_{seq})\, P(\mathbf{R}_{seq}) \tag{5.5}$$

In (5.5), data likelihood $P(\mathbf{X}|\mathbf{S}_{seq})$ is a probability density function (GMM) and role n-gram is a probability distribution. Similar to ASR systems, an insertion penalty and scaling factor are introduced to scale the two values to comparable ranges.

Schematically the social role based diarization system can be summarized as follows:

1. Extract acoustic features from the audio file.
2. Speech/non-speech segmentation and reject non-speech frames.
3. Initialize the model for the initial clusters using linear initialization.
4. Perform iterative merging using the following steps:
   (a) Apply speaker to role mapping $\Upsilon(s_n) = R_n$, for the current speaker segmentation.
   (b) Re-segment the data using role based duration constraints and role trigram model.
   (c) Retrain the speaker models using the Expectation-Maximization (EM) algorithm.
   (d) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.
   (e) If no such pair is found, stop and output the current clustering.
   (f) Merge the pair of clusters found in step (d). The models for the individual clusters in the pair are replaced by a single, combined model.
   (g) Go to step (a).

Step 1 to 3 are the same as in baseline system. First multiple distant microphones are beamformed to produce a single enhanced signal using Beamformit toolkit [128]. Acoustic features representing 19 MFCC coefficients are extracted using a 30ms window shifted by 10ms. After that we run an automatic speech/non speech segmentation and eliminate nonspeech regions to extract frame level acoustic features $\mathbf{X}$. Step 4 introduces new stages which are different from the baseline diarization system. These include a speaker to role mapping and a novel Viterbi segmentation using social role information. During re-training of speaker models, we do not consider frames corresponding to short turns, as previous studies [58] have shown that diarization errors are very high for these segments. The merging and clustering stopping criteria are the same for both the social role based diarization system and baseline diarization system.

5.5 Experiments

In this section, we describe the experiments that were performed to compare performance of the social role diarization system against the baseline diarization system. For experimental evaluation we selected 30 meetings from AMI corpus which are also annotated with social roles. We also evaluated the generalized performance of the proposed method on NIST Rich Transcription meetings.

For evaluation of diarization performance, most commonly used metric is the Diarization Error Rate (DER). DER is the composed of false alarm time, miss time and speaker error time. Since both, social role diarization system and baseline diarization system use the same speech nonspeech segmentation, the difference in performance is evaluated in terms of speaker error.

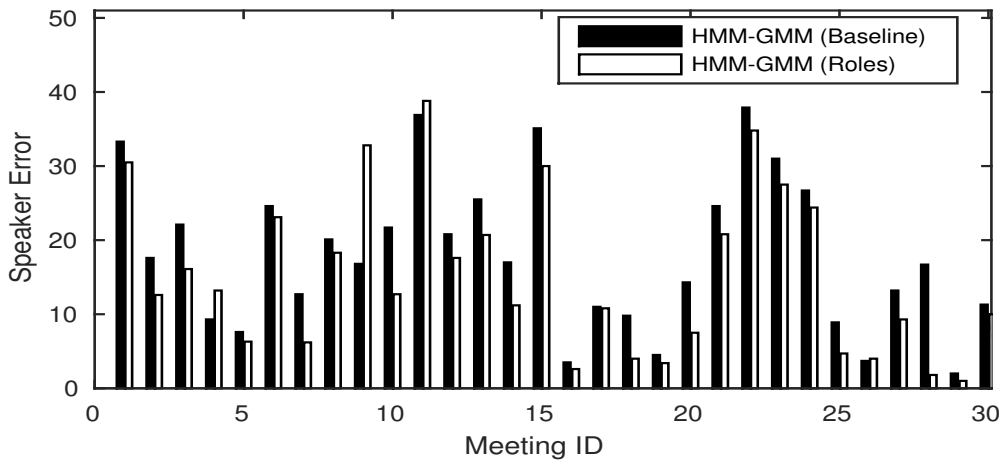

Figure 5.3: Per-meeting speaker error for the 30 meetings of the AMI corpus obtained using the baseline diarization system and social role diarization system.

Similar to NIST evaluations we have used a collar of 0.25 seconds around reference segment boundaries.

The social role recognition system used in this chapter operates in two stages, first each speaker's acoustic and structural behavior patterns are represented using a high dimensional feature vector, then in the next stage supervised model is used to predict a speaker's role. While Chapter 3 also includes linguistic information of speakers, for the task of speaker diarization we have only extracted non verbal features. Social role recognition performance was evaluated using crossvalidation keeping in view that same speaker does not appear in both training and test sets. The social role based trigram language model was estimated from 29 meetings using SRI toolkit [108]. The scaling factor and insertion penalty were tuned on a separate development set. The development set comprised of AMI meetings that were not annotated with social roles.

5.5.1 Evaluation on AMI meetings

For our first experiment, we include automatic speech/nonspeech segmentation and assume an unknown number of reference speakers. The hypothesized number of final speakers is determined by the BIC based stopping criterion. The performance comparison of the social role based diarization system and the baseline diarization system is shown in Table 5.2. Under these conditions, experiments reveal that proposed system, which integrates social role information in the diarization system, results in 16% relative improvement over baseline system. The per meeting performance comparison of the two systems is shown in Figure 5.3. It can be seen that the social role diarization system outperforms the baseline diarization system in most of the cases.

We also compared the role recognition performance from the final output of the two systems,

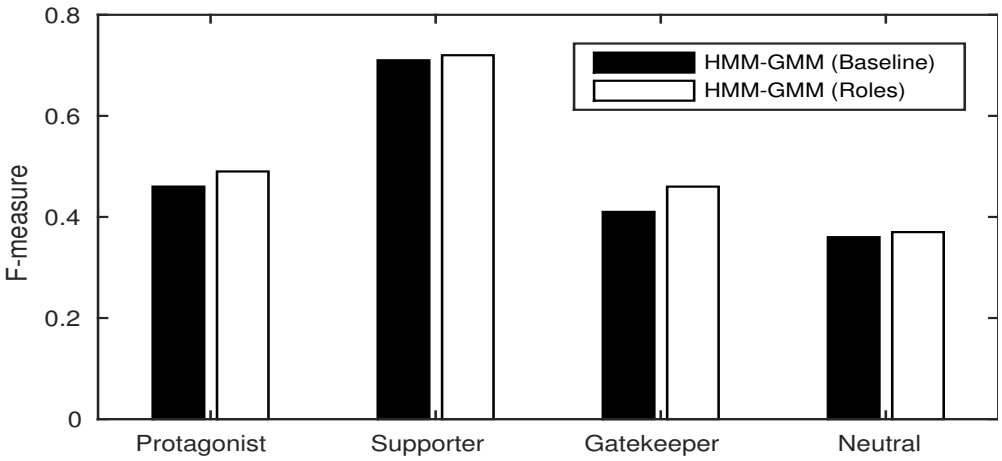

Figure 5.4: Social role recognition performance for each of the four roles using the speaker segmentation from baseline diarization system and social role diarization system.

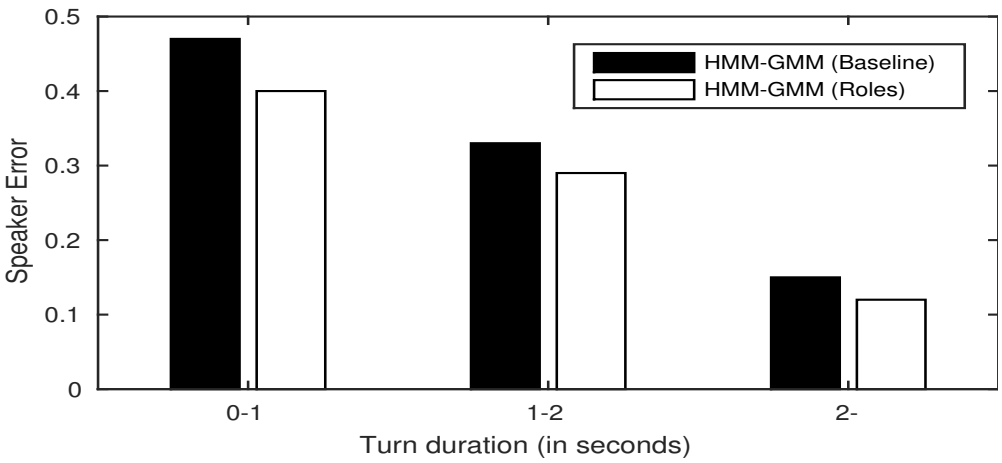

Figure 5.5: Variation in speaker error for various sizes of turn duration for the baseline diarization system and social role diarization system.

shown in Figure 5.4. It can be seen that the social roles of speakers are recognized more accurately for the HMM-GMM diarization system which incorporates social roles as prior information. The improvement in performance is spread across all the four social roles. The features extracted using baseline diarization system yield an accuracy of 56% and the features extracted using social role based diarization system yield an accuracy of 58%. However, if we compare the role recognition results obtained in Table 3.6, we observe that there is significant difference in performance between IHM conditions and distant microphone condition. Our results in Figure 5.4 show that degradation in performance is worse for neutral role. This can be due to fact that speakers assuming neutral role produce relatively shorter speech turns, which are incorrectly labeled in diarization output.

Finally, we also performed an analysis of diarization systems performance as a function of

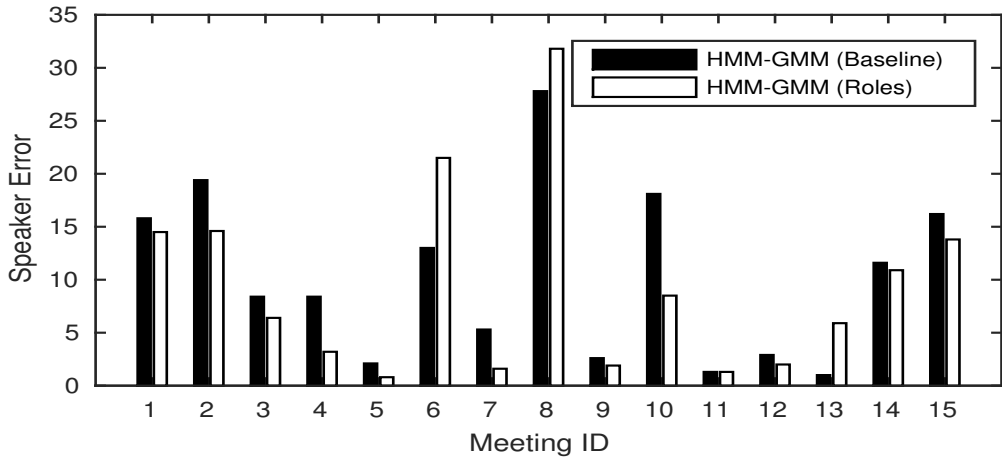

Figure 5.6: Per-meeting speaker error for the meetings of the RT07 corpus and the RT09 corpus obtained using the baseline diarization system and social role diarization system.

speaker turn duration. For this turns in the reference transcripts are partitioned into three bins, short turns with duration less than 1 second, intermediate turns with duration between 1 second and 2 second and long turns with duration greater than 2 seconds. The results for this analysis are presented in Figure 5.5. For both systems similar trends are observed, speaker error decreases as turn duration increases. However, the social role based diarization system shows improved performance for all bins.

Table 5.2: Speaker error obtained from the baseline system and the social role diarization system on AMI testset and RT dataset.

Dataset	Baseline diarization system Speaker Error	Social role diarization system Speaker Error
AMI	17.6	14.8(16%)
RT 07,09	10.2	8.9(13%)

5.5.2 Evaluation on RT meetings

In order to investigate the performance of the proposed method on other meeting scenarios, we also compared baseline and social role diarization systems on NIST Rich Transcription (RT) dataset. RT dataset contains audio recordings, representative of spontaneous conversation. Contrary to AMI meetings, these recordings are not necessarily elicited using a specific scenario. However, this does not represent a significant drawback , since social roles represent a generalized role coding scheme, and can conceivably be adapted to multiple conversation scenarios. For our analysis, we selected 15 meeting recordings comprising the evaluation sets RT-07 and RT-09. The social role information represented using parameters of duration model and n-gram role model were obtained on the AMI training set. Table 5.2 shows the performance of both role based diarization system and baseline system for this dataset. Table numbers reveal that speaker error drops from 10.2% using the baseline system to 8.9% using

the social role diarization system, which represents around 13% relative improvement in performance. This shows that including the influence of social roles on turn taking is effective in reducing the speaker error even on unseen scenarios. A meeting wise comparison of the two systems, shown in Figure 5.6, reveals that social role diarization system performs better in most of the cases. Further analysis revealed that the few meetings where the social role diarization system fails to improve over the baseline results correspond to cases that have higher than average missed speech rate.

5.6 Conclusion

In this chapter, we extended the state-of-the-art speaker diarization framework by using social roles to model the expected duration of speaker's turn. The turn taking interaction between speakers was modeled using role n-gram model. Our experiments conducted on AMI corpus meetings revealed that the inclusion of social roles as prior information in speaker diarization reduces the speaker error by 16% relative to the baseline diarization system.

We also investigated how social role statistics generalizes on a completely different corpus. Meetings from the Rich Transcription campaign, multiparty conversations collected in different sites, were used for this purpose. Results revealed a 13% relative improvement for the social role based diarization system compared to the baseline system.

In this chapter, we considered the difference in turn taking characteristics of speakers when they assume different social roles. However, we did not consider the context in which social roles of speakers can change. In the next chapter, we analyze the change in social roles of meeting participants and investigate whether these changes are related to shifts in conversation topics.

6 Topic segmentation using social roles

6.1 Introduction

Multiparty interactions between participants often involve discussion about different topics which evolve and change as the conversation progresses. Detecting topic changes is therefore an important step towards automatic access and retrieval of information in multiparty interactions. Meetings are characterized by a hierarchical structure which is reflected in the coarse and fine grained segmentation of conversation into multiple topics. In this chapter, we explore the potential of social role information for segmenting the audio recording of meetings into multiple topics.

In Chapter 3, we demonstrated that social roles capture the behavior of speakers and social roles of speakers can change as the meeting progresses. In this chapter, we show that change in social roles of speakers during a meeting is related to shifts in conversation topics. We compute the social role posterior vectors of meeting participants by applying the procedure described in Section 3.4. We consider the role posteriors as informative features for identifying topic boundaries in a meeting. We cast topic segmentation as a boundary/nonboundary detection problem that can be solved by training a supervised classifier on social role posterior features. We apply the trained classifier on meeting data to show the applicability of social roles for topic segmentation.

In addition to modeling social roles for topic segmentation, we show that word frequency distribution in a conversation also provides important information in identifying topic boundaries. Moreover, we also demonstrate that word frequency distribution and social roles provide complementary information about topic segmentation. In this chapter, we present an unsupervised approach which combines social roles with latent topic models in a Bayesian framework. Our work builds on the recent progress in latent topic modeling with non parametric methods [109] and is related to previous unsupervised topic segmentation methods [85, 29]. We use distance dependent Chinese Restaurant Process (ddCRP) [17], which is generalized form of Dirichlet Process, to incorporate prior information about topic boundaries. The prior information in ddCRP framework is based on distance between social role posteriors

associated with different speech turns. We evaluate our approach on two different meeting corpora and compare its performance against state-of-the-art topic segmentation methods.

The chapter is organized as follows. In Section 6.2, we describe supervised topic segmentation using social role information. Section 6.3 introduces a novel approach based on ddCRP, in which we incorporate social roles of speakers as prior information for unsupervised topic segmentation. In Section 6.3, we also outline the inference of latent variables in ddCRP approach. The experimental methodology for both supervised and unsupervised approach is presented in Section 6.4, where we also compare and discuss the performance of proposed approaches against state-of-the-art methods. We conclude our work in Section 6.5

6.2 Supervised topic segmentation using social roles

Supervised methods of topic segmentation are based on extracting a wide range of features to identify topic boundaries. However, most of the previous methods have ignored social role information for topic segmentation. In this section, we use the role recognition model to predict social roles of different speakers and use them as potential features for topic segmentation. State-of-art topic segmentation methods are based on extracting lexical information from manual speech transcripts or ASR speech transcripts generated from headset microphones. To have a meaningful comparison with current baselines and investigate the full effectiveness of social roles for topic segmentation, all the experiments in this chapter are in the context of headset microphones attached to each participant.

We represent each meeting as a sequence of turns $\{t_i\}$. For each turn t_i, we consider a windowed meeting slice (duration $\sim$ 30 seconds) whose starting point corresponds to beginning of turn t_i. We extract various verbal and nonverbal features within the meeting slice associated with t_i (described in Section 3.3) and then apply the social role recognition model on the extracted features. The output of role recognition model for each speaker is a posterior distribution over four social roles in set $\mathscr{R}$. Let the speakers in a meeting be represented by the set $\mathscr{S}$. For each $S \in \mathscr{S}$ and for any turn t_i, we define a posterior feature vector $\mathbf{Y}_R^{t_i,S} = \{P(R^{t_i,S} = R), \forall R \in \mathscr{R}\}$. Here $P(R^{t_i,S} = R)$ denotes the posterior probability of speaker S being assigned a social role R. The social role configuration at t_i is represented by concatenating posterior feature vectors $\mathbf{Y}_R^{t_i,S}$ of the conversation participants in $\mathscr{S}$. This is represented as vector $\mathbf{Y}_R^{t_i} = \{\mathbf{Y}_R^{t_i,S}, \forall S \in \mathscr{S}\}$. In this chapter, we investigate whether social role information of the participants, represented by vector $\mathbf{Y}_R^{t_i}$ is suitable for prediction of topic changes in meetings.

For the task of topic change detection each turn t_i is a possible location for topic change. We hypothesize that neighboring turns t_i and t_j, when they belong to the same topic segment should have similar statistics for $\mathbf{Y}_R^{t_i}$ and $\mathbf{Y}_R^{t_j}$, while statistics of $\mathbf{Y}_R^{t_i}$ and $\mathbf{Y}_R^{t_j}$ for turns on either side of the boundary should be different. We incorporated this information by putting a window of length $2L+1$ around the current turn and concatenating all the features within the window to represent the feature vector $\mathbf{X}(t_i) = [\mathbf{Y}_R^{t_{i-L}}..., \mathbf{Y}_R^{t_i}, ...\mathbf{Y}_R^{t_{i+L}}]$.

Given feature vector $\mathbf{X}(t_i)$, boundary set $\mathscr{B} = \{0, 1\}$ and $b \in \mathscr{B}$, the mapping function $\mathcal{B}(t_i) \mapsto b$ is given as:

$$\tilde{b} = \underset{b \in \mathscr{B}}{\arg\max}\, P(\mathcal{B}(t_i) = b | \mathbf{X}(t_i)) \tag{6.1}$$

We used Boosting as the supervised model [99] to train social role based features. The principle of boosting is to combine many weak learning algorithms to produce a single accurate classifier. The algorithm generates weak classification rules by calling the weak learners repeatedly in series of rounds. Each weak classifier is built based on the outputs of previous classifiers, focusing on the samples that were formerly classified incorrectly. The weak learners are one-level decision trees.

6.3 Unsupervised topic segmentation

The model described in this section partitions the meeting conversation into disjoint segments and in each segment a specific set of semantic concepts are discussed. Our model assumes that each turn in the meeting can be represented as a mixture of latent topics, where each latent topic is represented as a probability distribution over words in the vocabulary. The main idea underlying our approach is that a set of turns which constitute a topically coherent segment should be generated form the same latent topic distribution. However, the partition of turns into topic segments is not know a priori. Instead, our approach jointly models both the segmentation and the latent topic distribution associated with each topic segment.

We assume a distance dependent Chinese restaurant process (ddCRP) as the distribution over partitions of turns. This distribution incorporates social role information to represent changes in dynamics of conversation during a meeting conversation. The next subsections, explain in detail our approach for topic segmentation. We first explain the motivation for proposing a ddCRP prior over turns and then describe the complete generative approach for modeling meeting conversations.

6.3.1 Chinese restaurant process

Dirichlet process mixture model (DPMM) [7] is a data dependent clustering method where the number of clusters is not fixed in advance, but determined directly from the data. A Dirichlet process (DP) prior over the number of clusters is assumed in DPMM. Chinese Restaurant Process (CRP) is alternate representation of the DP, that can be defined in terms of probability distribution over partitions of data [34, 72]. CRP is described using the analogy of a restaurant which has a countably infinite number of tables. Each table in this restaurant is associated with parameters from a family of distributions which generate the data. Customers enter this restaurant in a sequential manner and sit at randomly chosen tables. The first customer sits at the first table. The second customer can sit at the same table as the first customer or sit on a

new table. After N customers have entered the restaurant their seating arrangement describes a partition of data (customers) into different clusters (tables).

Let u_i represent the table assigned to i^{th} customer. We also assume that previous $i-1$ customers have already entered the restaurant and are sitting at U different tables. The probability that the customer i sits at table u is proportional to the number of customers n_u already seated at u. The customer i can also sit at a previously unoccupied table with a probability proportional to a given scaling parameter α_o. CRP defines the probability of assigning table u to customer i as,

$$P(u_i = u | u_{1:i-1}, \alpha_o) \propto \begin{cases} n_u & \text{for } u \leq U \\ \alpha_o & \text{for } u = U+1. \end{cases} \tag{6.2}$$

This process generates a random partition of i customers based on their table assignments.

Although the customers are assigned to tables sequentially, the probability distribution over partitions is invariant to the order in which customers enter the restaurant. This is because, CRP as a representation of Dirichlet process, is an exchangeable model and the ordering of data points does not change their probability distribution. While exchangeability is a reasonable assumption in many clustering applications, it is not applicable for linear segmentation of meeting conversations.

6.3.2 Distance dependent Chinese restaurant process

The distance dependent Chinese Restaurant Process (ddCRP) [17] is an extension of CRP that allows for a non exchangeable distribution over partitions. In traditional CRP, customers are directly assigned to tables in the restaurant. In comparison, ddCRP links a customer with itself or with other customers. The seating arrangement of customers is a byproduct of these customer-customer links. Customers sit at the same table when they are linked with each other. This is illustrated in Figure 6.1. The colored blocks represent the customers and circles represent the tables. The links between customers are shown by arrows connecting the blocks. In Figure 6.1, customer 3 is linked with itself, customers 4, 5 are also directly linked with 3 and customer 6 is linked indirectly to customer 3 via customer 5. No other block is linked with any of these four blocks. Therefore, customers 3, 4, 5 and 6 sit at the same table.

Let $\{c_1, c_2, ...c_N\}$ be a collection of N variables, in which for each customer i, c_i denotes the index of customer with whom the ith customer is linked. Let d_{ij} denote the distance between customers i and j, $\mathbf{D}$ denotes the set of distances between all customer pairs (i, j) and $f(d_{ij}) \in [0, 1]$ is a decay function. Let γ represent the self link parameter ,i.e., $d_{ii} = \gamma$. In ddCRP, the distribution of customer assignments is,

$$P(c_i = j | \mathbf{D}, \gamma) \propto \begin{cases} f(d_{ij}) & \text{for } j \neq i \\ \gamma & \text{for } j = i. \end{cases} \tag{6.3}$$

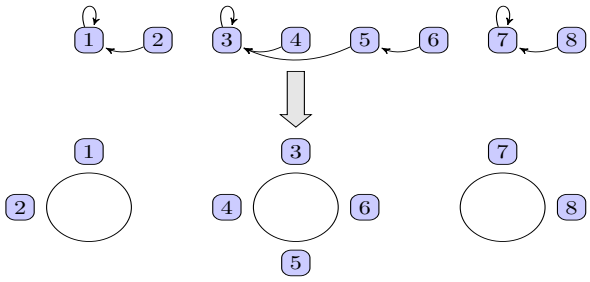

Figure 6.1: The seating arrangement of customers on various tables in ddCRP. The top plot shows customers linked either with themselves or with other customers. Bottom plot shows the table arrangement inferred from those customer assignments.

From (6.3), we can observe that probability of customer assignment depends on the distance between customers. The decay function f relates the influence of the distance between two customers and the probability that they are connected with each other, i.e, they share the same cluster.

In a more general approach a customer can be assigned to any other customer. However, topic segmentation usually requires that speaker turns are clustered sequentially, which implies distance between nearby turns is more important than distance between turns that are widely separated. In this work, we only consider sequential ddCRPs which assume that distance $d_{ij} = \infty$ for $j > i$. We also assume that decay function f takes only non-negative values and satisfies $f(\infty) = 0$. These assumption imply that no customer can be assigned to a later customer.

6.3.3 Topic segmentation as a generative process

Let us consider a corpus consisting of a set of meetings $\mathscr{M}$. Each meeting $M \in \mathscr{M}$ has multiple speakers who take part in the conversation to produce N_M turns. Each turn t is associated with a bag of $N_{M,t}$ words $\{w_{M,t,n}, n \in [1, N_{M,t}]\}$. Furthermore, speakers in the meeting also assume different social roles that can change dynamically during the conversation. We represent by the vector $\mathbf{Y}_R^{M,t}$ (described in Section 6.2) as the social roles of the speakers associated with turn t in meeting M.

We consider each meeting M as a restaurant and customers are turns $\{t\}$ in the meeting. The partition of customers into distinct tables are considered as topic segments in our representation. The tables in ddCRP are a byproduct of customer assignments. Since we consider sequential distances in this work, a new segment is formed when a customer (turn in the meeting) is linked with itself. Furthermore, ddCRP is formulated such that the number of potential topic segments, while unbounded are influenced by the distance between turns. Unlike traditional CRPs, where a constant distance is assumed between current turn and each of the previous turns, we consider distances between pair of turns based on the changes

in social role configuration. The distance between turns t_i and t_j is expressed in terms of $d(\mathbf{Y}_R^{M,t_i}, \mathbf{Y}_R^{M,t_j})$. We hypothesize that change in social roles of speakers convey information about the state of meeting conversation. For example, if a speaker previously acting as a neutral assumes the role of a protagonist then such a transition can be responsible for a shift in conversation topic. In this work, the distance d_{ij} is represented using Kullback–Leibler (KL) divergence between social role posteriors estimated for two adjacent turns t_i and t_j and averaged across all speakers.

The words corresponding to each turn in the meeting are assumed to be generated from a mixture of K latent topics. The distribution over latent topics associated with a turn is denoted by θ_t, which is a multinomial distribution over K topics such that the probability of drawing topic $k \in [1,...,K]$ is $\theta_t(k)$. In each meeting M, if the customer assignment of turn t_i is linked to previous turn t_{i-1}, i.e., $c_{M,t_i} = i-1$, then $\theta_{t_i} = \theta_{t_{i-1}}$. On the other hand, if customer assignment of a turn is linked with itself, i.e., $c_{M,t_i} = i$, then a new θ_{t_i} is drawn from a Dirichlet distribution with parameter α. The customer assignments result in partition of M into L_M tables. We use variable $l_M \in [1, L_M]$ to denote the table representation. A set of turns $\{,...,\theta_{t_i}, \theta_{t_{i+1}},...\}$ assigned to table l_M share the latent topic distribution θ_{l_M} corresponding to table l_M , i.e., $\theta_{t_i} = \theta_{t_{i+1}} = \theta_{l_M}$. We also use variable u_{M,t_i} to denote the table assigned to turn t_i. If t_i sits at table l_M, then $u_{M,t_i} = l_M$.

Following the method proposed in [16], each latent topic k is a probability distribution ϕ_k over a vocabulary of size V. The probability of generating a word $w \in [1, V]$ for a latent topic assignment k is $\phi_k(w)$, where ϕ_k is Dirichlet distribution with a symmetric prior β. The words $\{w_{M,t,n}\}$ corresponding to a turn t are generated by first sampling a topic assignment $z_{M,t,n}$ for each word position n in that turn and then a sampling a word $w_{M,t,n}$. The sampling probability for latent topic is $P(z_{M,t,n} = k|\theta_t) = \theta_t(k)$ and sampling probability for word is $P(w_{M,t,n} = w|\phi_k, z_{M,t,n} = k) = \phi_k(w)$. The complete generative process is as follows:

- For each latent topic k in $\{1...K\}$:
  - Draw multinomial over words $\phi_k \sim \text{Dir}(\beta)$
- For each M and each t_i, draw $c_{M,t_i} \sim \text{ddCRP}(\mathbf{D}, \gamma, f)$
- Connect all c_{M,t_i} to get table representation (topic segments), (e.g., table l_M might consist of set of turns $u_{M,t_i} = \{,.., t_i, ..., t_j, ...\}$)
- For each topic segment l_M:
  - Draw the shared multinomial over latent topics $\theta_{l_M} \sim \text{Dir}(\alpha)$
  - For each turn t_i in l_M:
  - For each observed word position (M, t_i, n):
    - Draw a latent topic $z_{M,t_i,n} \sim Mult(\theta_{l_M}), z_{M,t_i,n} \in \{1,...,K\}$
    - Draw a word $w_{M,t_i,n} \sim Mult(\phi_{z_{M,t_i,n}})$

6.3.4 Inference

We have described the generative process for meeting conversations. The fundamental problem we need solve is to infer the latent variables that best explain the observed data. The variables in our model consists of customer assignments $\mathbf{c} = \{c_{M,i}\}$, word tokens $\mathbf{w} = \{w_{M,t,n}\}$, latent topics $\mathbf{z} = \{z_{M,t,n}\}$, topic segment multinomials $\Theta = \{\theta_{l_M}\}$ and word topic multinomials $\Phi = \{\phi_k\}$. The joint probability of all the parameters in the model is given as,

$$P(\mathbf{c}, \mathbf{w}, \mathbf{z}, \Theta, \Phi | \alpha, \beta, \gamma, \mathbf{D}) = P(\mathbf{c}|\gamma, \mathbf{D}) P(\Theta|\mathbf{c}, \alpha) P(\mathbf{z}|\Theta) P(\mathbf{w}|\mathbf{z}, \Phi) P(\Phi|\beta) \tag{6.4}$$

We can simplify (6.4) by integrating out the latent variables Θ and Φ.

$$P(\mathbf{w}|\mathbf{z}, \beta) = \int P(\mathbf{w}|\mathbf{z}, \Phi) P(\Phi|\beta) d\Phi \tag{6.5}$$

$P(\Phi|\beta)$ is represented using Dirichlet distribution and $P(\mathbf{w}|\mathbf{z}, \Phi)$ is represented using multinomial distribution.

$$P(\Phi|\beta) = \prod_{k=1}^{K} P(\phi_k|\beta) \propto \prod_{k=1}^{K} \prod_{v=1}^{V} (\phi_k(v))^{\beta-1} \tag{6.6}$$

$$P(\mathbf{w}|\mathbf{z}, \Phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} (\phi_k(v))^{I_{kv}} \tag{6.7}$$

I_{kv} denotes the number of times topic k is assigned to word $v \in [1, V]$ in the corpus. Using the property that Dirichlet distribution is conjugate distribution of multinomial distribution and the posterior distribution is also a Dirichlet distribution, we can show that (6.7) can be simplified by marginalizing over Φ. Let us use Γ to denote gamma function. The marginalized expression is defined as:

$$P(\mathbf{w}|\mathbf{z}, \beta) = \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V} \frac{\prod_{v=1}^{V} \Gamma(\beta + I_{kv})}{\Gamma(\sum_{v=1}^{V} \beta + I_{kv})} \tag{6.8}$$

Similar ideas can be applied to integrate out variable Θ form (6.4).

However, exact inference for the latent variables is not feasible. Instead we rely on approximate inference techniques based on Markov Chain Monte Carlo (MCMC) algorithm to generate draws from a complex distribution. We use Gibbs sampling which is a simplified form of MCMC method [64]. The main idea behind Gibbs sampling, is to sample from univarite conditional distributions. All the random variables except the current one are fixed to their previous state and a new value is sampled for the current variable.

The Gibbs sampler is defined over the space of latent variables in our model. These consist of

customer assignments **c** over turns and latent topics **z** corresponding to each spoken word. In each iteration of the sampler, we cover all the turns in all the meetings in our corpus. For each turn t in a meeting M we first sample its customer assignment c and then sample the latent topic z for each word token w in that turn. The details of this procedure are presented next.

6.3.5 Sampling latent topics

One key step to perform inference in our model is to sample latent topic assignments over spoken words, conditioned on all customer assignments **c**. As previously discussed, a given sequence of customer assignments can be used to define the meeting segments. A new segment (table) is formed when the customer assignment for a turn links to itself. If two turns t_i and t_j are linked by an intermediate sequence of customer assignments then they are in the same segment (share the same table). Let $u_{M,t}$ denote the topic segment to which the turn t belongs. Given the customer links **c** and topic assignments $\mathbf{z}_{-m,t,n}$ (for all the other words in the corpus, with the exception for word $w_{M,t,n}$), the conditional equation for assigning latent topic k at the current word position $z_{M,t,n}$ is:

$$P(z_{M,t,n} = k|\mathbf{z}_{-M,t,n}, \mathbf{c}, \mathbf{w}) \propto \frac{I_k^{u_{M,t}} + \alpha}{I_{\cdot}^{u_{M,t}} + K\alpha} \frac{I_{kw_{M,t,n}} + \beta}{I_{k.} + V\beta} \tag{6.9}$$

In (6.9), all the counts are represented by index variable I. $I_k^{u_{M,t}}$ denotes the number of times latent topic k is assigned for words in segment $u_{M,t}$ and $I_{\cdot}^{u_{M,t}}$ is the sum of all the topic assignments in $(u_{M,t})$. $I_{kw_{M,t,n}}$ is the number of times word $w_{M,t,n} \in [1, V]$ is assigned to topic k across the corpus and $I_{k.}$ is the total number of words assigned to k across the corpus. All of these counts exclude the topic assignment for the current position.

6.3.6 Sampling customer assignments

The conditional distribution for iteratively sampling customer assignment $c_{M,t}$ for each turn t conditioned on other customer assignments $\mathbf{c}_{-M,t}$ and latent topic assignments **z** is:

$$P(c_{M,t}|\mathbf{c}_{-M,t}, \mathbf{w}, \mathbf{z}, \gamma, \mathbf{D}) \propto P(c_{M,t}|\mathbf{D}, \gamma) P(\mathbf{z}|\mathbf{c}, \alpha) \tag{6.10}$$

The first term in (6.10) denotes the influence of prior information and second term denotes the distribution of latent topics given the current segmentation. We can factor the second term in (6.10) as:

$$P(\mathbf{z}|\mathbf{c}, \alpha) = \prod_{\forall M \in \mathscr{M}} \prod_{l_M=1}^{L_M(\mathbf{c})} P(\mathbf{z}^{l_M(\mathbf{c})}|\alpha) \tag{6.11}$$

Here $\mathbf{z}^{l_M(\mathbf{c})}$ is the collection of latent topic variables assigned to table l_M. The distribution of latent topics in $l_M(\mathbf{c})$ is given as:

$$P(\mathbf{z}^{l_M(\mathbf{c})}|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{k=1}^{K} \Gamma(I_k^{l_M(\mathbf{c})} + \alpha)}{\Gamma(I_{\cdot}^{l_M(\mathbf{c})} + K\alpha)} \tag{6.12}$$

where $I_k^{l_M(\mathbf{c})}$ is the number of times a latent topic variable in segment $l_M(\mathbf{c})$ is assigned a value k and $I_{\cdot}^{l_M(\mathbf{c})}$ is total number of latent variables in $l_M(\mathbf{c})$. Because of this factorization, we only need to consider terms corresponding to a single meeting M and only those terms which are influenced by reassignment of $c_{M,t}$.

To sample from (6.10), we first remove the customer link of turn t and then we compute the probability of each possible reassignment. If as a result of reassignment customer links to itself, then there is no change in table configuration. Otherwise, reassigning $c_{M,t}$ can result in joining of two tables l_M and r_M.

The conditional sampling distribution for customer assignments is given by,

$$P(c_{M,t}|\mathbf{c}_{-M,t}, \mathbf{z}, \gamma, \mathbf{D}) \propto \begin{cases} P(c_{M,t}|\mathbf{D}, \gamma) \frac{P(\mathbf{z}^{l_M(\mathbf{c})} \cup \mathbf{z}^{r_M(\mathbf{c})}|\alpha)}{P(\mathbf{z}^{l_M(\mathbf{c})}|\alpha) P(\mathbf{z}^{r_M(\mathbf{c})}|\alpha)} & \text{if } c_{M,t} \text{ joins tables } l_M \text{ and } r_M \\ P(c_{M,t}|\mathbf{D}, \gamma) & \text{otherwise.} \end{cases} \tag{6.13}$$

In Equation (6.13), the expression for $P(\mathbf{z}^{l_M(\mathbf{c})}|\alpha)$ can be solved using (6.12). Similar procedure can be used to solve both $P(\mathbf{z}^{r_M(\mathbf{c})}|\alpha)$ and $P(\mathbf{z}^{l_M(\mathbf{c})} \cup \mathbf{z}^{r_M(\mathbf{c})}|\alpha)$.

6.4 Experiments

6.4.1 Topic annotation: AMI corpus

The scenario meetings in AMI corpus were annotated with topic segmentation [131]. Since scenario meetings follow a prearranged agenda several common topics were expected to reappear regularly. Annotators were given a standard set of topic descriptions that act as labels for identifying topic segments. Three categories of topic description were defined.

- *Top level topics* reflect the meeting agenda, e.g., presentation, discussion and issues related to remote design.
- *Functional topics* refer to off-topic discussion between participants, including segments related to opening, closing, chitchat, etc.
- *Sub topics* divide parts of complex top level topics into segments related to budget, market trends, usability, materials, components and energy resources.

Table 6.1: Average intercoder agreement for Top level and subtopic segmentation in AMI meetings

Segmentation	κ	P_k	WD
TOPSEG	0.70	0.11	0.17
ALLSEG	0.60	0.23	0.28

The annotators had freedom to mark any speaker turn (consecutive speech with no pause longer than 0.5 seconds) as a coarse (top level, functional) or fine (sub) topic boundary. The reliability of annotation scheme was measured using κ statistic between pair of coders. The details of this procedure are described in [131]. Table 6.1 shows the κ values for two levels of topic segmentation. In Table 6.1, TOPSEG refers to coarse topic segmentation at the top level and ALLSEG includes all top level and sub level segments. The average number of TOPSEG boundaries in a meeting was 7 while average number of ALLSEG boundaries was roughly 12. Table 6.1 also reports the average segmentation scores using metrics P_k and WindowDiff (WD). Analysis of κ statistic reveals a reliable topic segmentation, with good agreement on top level boundaries and moderate agreement on sub topic boundaries. This is also confirmed from low values of P_k and WD scores.

6.4.2 Baseline results for supervised topic segmentation

The first question which we wish to address is how topic segmentation using social role information compares against existing approaches. Several previous works have explored a wide range of features for topic segmentation in meetings. These features exploit the characteristics of speaker behavior when they initiate a new discussion or end an existing one. For example, long pause between speaker turns is often seen as a marker for a new topic segment. In [132], Maxent classifier was trained using multiple feature groups and a detailed comparison of various conversational, prosodic, motion and lexical features was investigated for topic segmentation in AMI meetings. Those results serve as a benchmark against social roles based method. We now briefly summarize various features described in [132] and then compare their performance against supervised social role based approach for topic segmentation.

Conversational features (CONV): Several speaker interaction features were extracted, including amount of pauses between speech segments, amount of speaker activity change, and amount of overlapping speech. The authors also included predictions of LCseg, such as lexical cohesion at the potential boundary, estimated posterior probability and predicted boundary class as additional conversation features.

Lexical features (LEX): A list of words was complied from the training data. These words occur more often near the top level or sub-topic segment boundaries. The lexical feature vector is

represented as vector space of unigrams from the complied list.

Prosodic features (PROS): The audio data was processed to extract prosodic features which capture speaker information derived from fundamental frequency, energy and intonation. These features were extracted at different points of a speaker turn and include features derived from speech rate, pitch contour, and speech energy.

Motion features (MOT): The motion of various meeting participants was captured from video recordings. All the AMI meetings were recorded using close up and overview video cameras. The head and body movements of speakers were captured using close up cameras and hand movements were captured using overview cameras. The video frames which correspond to the time interval for a speaker turn were processed to extract the average magnitude of movement for each speaker.

Contextual features (CTXT): The features were used to describe the context for each turn based on the formal role of the speaker and the dialog act types used to characterize the turn.

In order to compare the performance of social roles for topic segmentation, we report the performance of feature based method described in [132]. Table 6.2 summarize the performance of different features for topic segmentation, including evaluation at TOPSEG and ALLSEG levels.

Table 6.2: Baseline results showing effect of different feature groups for topic segmentation in AMI meetings.

Features	TOPSEG		ALLSEG	
	P_k	WD	P_k	WD
LEX	0.53	0.72	0.49	0.66
CONV	0.34	0.34	0.37	0.37
PROS	0.35	0.35	0.37	0.37
MOT	0.36	0.40	0.38	0.41
CTXT	0.34	0.34	0.37	0.37
ALL	0.29	0.33	0.35	0.38

6.4.3 Results for supervised topic segmentation using social roles

To evaluate the performance of social roles for topic segmentation we used a set of 100 scenario meetings. The meetings where selected from all the three recording sites, including 50 meetings recorded at Edinburgh, 30 meetings recorded at Idiap and 20 meetings recorded at TNO. For our experiments, we did a site wise evaluation, where meetings from one of the three sites was kept for testing, while social role based classifier was trained on meetings from other two sites. Topic segmentation performance is reported in terms of P_k and WD. Unless otherwise stated all the reported results correspond to ALLSEG case which includes both top level and sub topic boundaries.

The supervised classification performance was also compared against a random baseline. The baseline scores were obtained by taking the turn sequence and randomly marking them as boundary candidates. However, to make effective comparison the boundary marks were made proportional to average number of topic changes in the training set. Evaluation scores were obtained by 10000 iterations of this procedure during testing.

Influence of context size

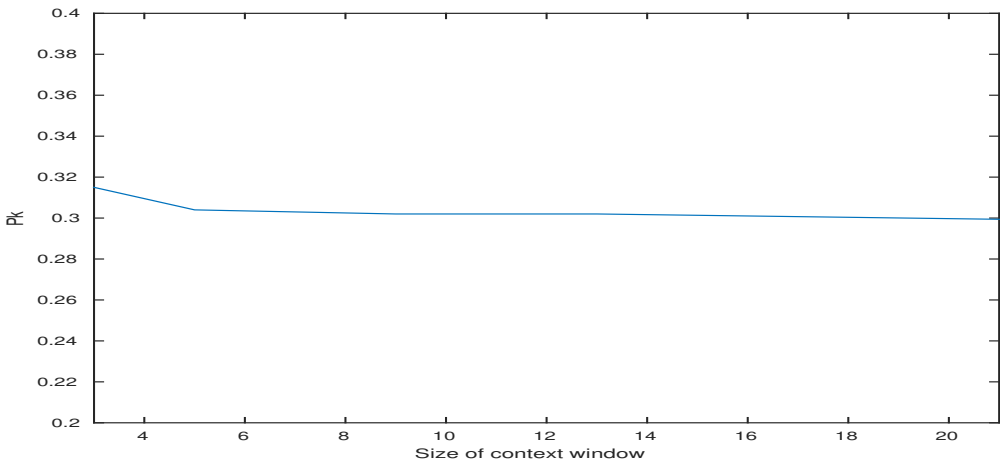

Figure 6.2: Variation in topic segmentation performance (measured in terms of P_k values) as a function of window size.

In order to take into account the change in social role distribution near a topic boundary, we employed context windows of increasing size (described in Section 6.2). For each position of the window, feature set is comprised of social role posterior of each participant from starting turn of the window to the ending turn of window. The size of the feature vector is dependent on number of speakers and width of the window. For each window size, classifiers were trained on different versions of data. In Figure 6.2, we report the performance of the classifier, measured using P_k score, as a function of size of feature window. The shape of the plot reveals that adding some surrounding context to social role posterior features improves performance. However, the performance does not increase proportional to the size of context window. This suggests that changes in social roles of meeting participants occur near topic boundaries.

In Figure 6.3, we show an example meeting which compares the performance of automatic segmentation against the topic segment boundaries marked by human annotators. The vertical lines in the plot, which originate from the top horizontal axis correspond to turn index where reference topic segments start. The vertical lines which originate from the bottom horizontal axis show the hypothesized boundaries predicted using social roles (context size=7). Figure 6.3 reveals that for majority of cases the hypothesized topic boundaries are located near the reference topic boundaries. This holds for both top level topics and sub topics (shown by dark black lines). This suggests that social roles are good predictors for both coarse level and more fine grained topic changes in meetings.

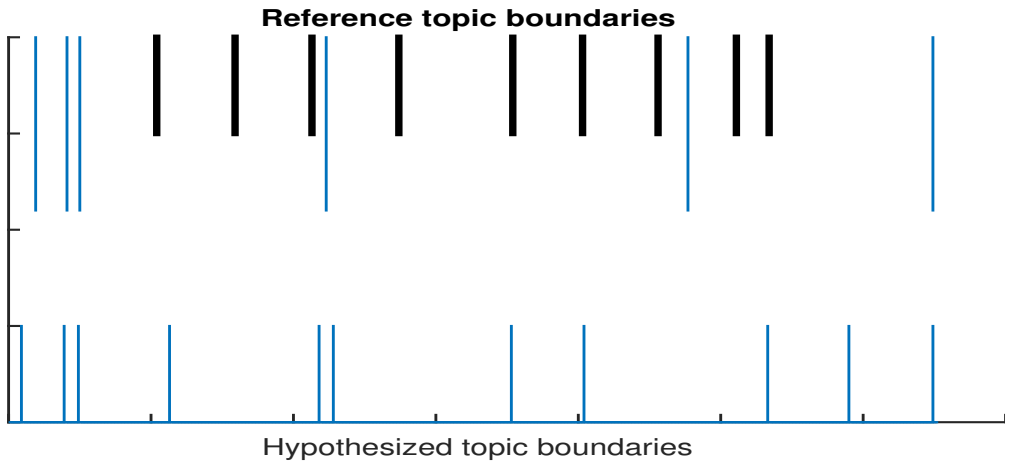

Figure 6.3: Performance of topic segmentation model using social role posterior features over an example AMI meeting. For the top horizontal axis, the longer thin lines represent top topic level boundaries and thick black lines represent additional sub topic level boundaries specified by human annotators. The boundaries predicted automatically using social roles are shown as vertical lines starting from bottom horizontal axis.

Evaluation results

Table 6.3: Topic segmentation results for various social role posterior features evaluated on 100 meetings from AMI corpus.

Features	TOPSEG		ALLSEG	
	P_k	WD	P_k	WD
Random	0.45	0.48	0.46	0.50
Social Role posteriors ("hard")	0.31	0.37	0.33	0.39
Social Role posteriors	0.28	0.34	0.30	0.36

Table 6.3 shows the performance of random baseline and two supervised models for topic segmentation. The difference between the two models is due to the fact, in one case "hard" features were computed by binarizing the social role posteriors to a value in $\{0,1\}$. In this case, we throw away the confidence of estimated role posteriors and retain only the predicted role label. Table numbers reveal that social role based topic segmentation results in significant increase in performance compared to random baseline. We observe that for all the models WD scores are worse than P_k scores. This can be due to fact WD measure penalizes over predictions more heavily compared to P_k measure. We also observe that "hard" social role features perform worse than estimated social role posterior features for topic segmentation. This suggests that supervised topic segmentation classifier extracts information from changes social roles, as well as, the confidence with which those social roles were estimated. Figure 6.4 plots the P_k measure for random baseline, social role posterior features and their hard counterparts. All the values used in Figure 6.4, correspond to ALLSEG (both top level and sub topic boundary) context. Our analysis reveals that for only one meeting recorded at Edinburgh posterior features perform worse than random error, while for the meetings recorded at Idiap

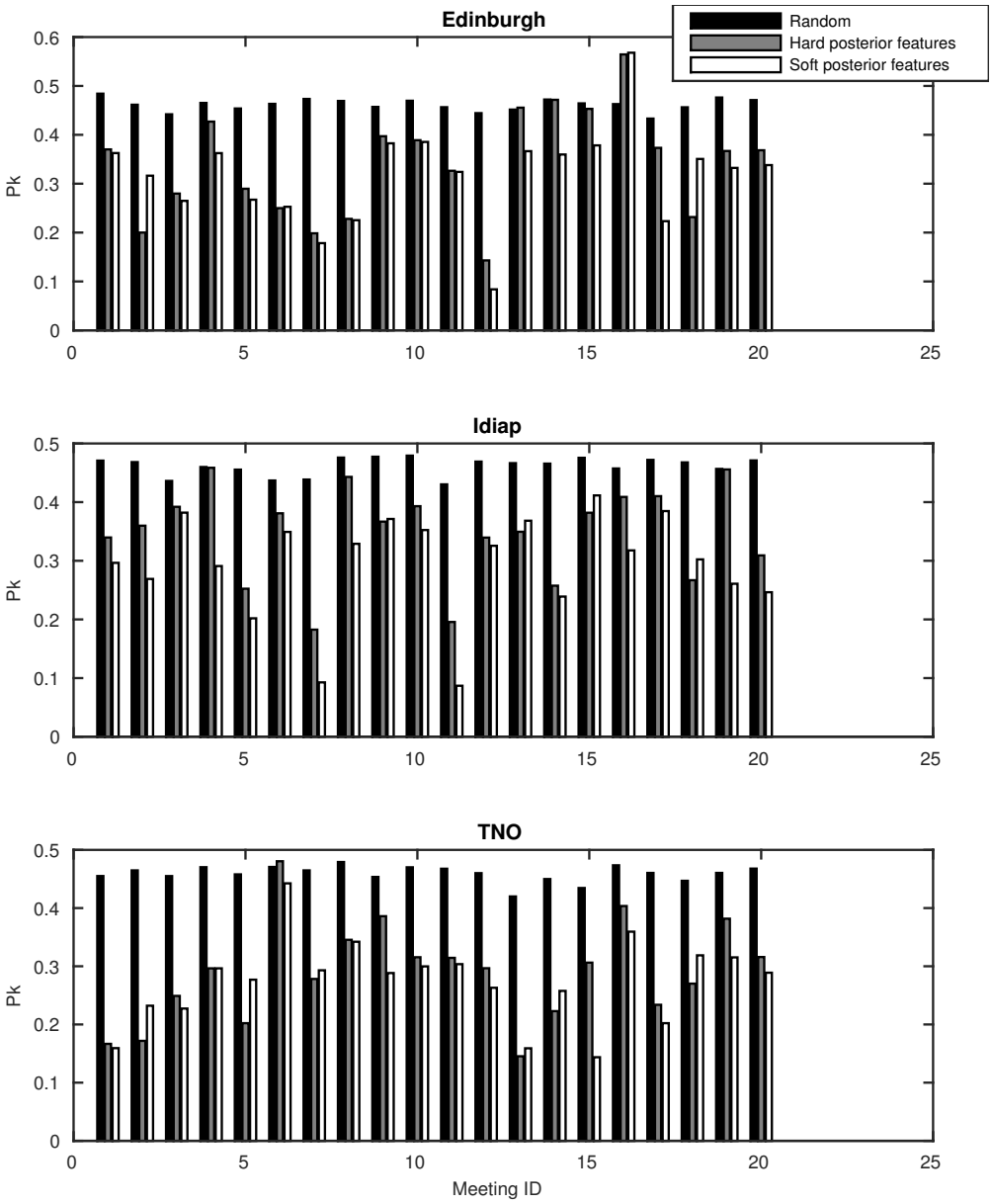

Figure 6.4: P_k values for various meetings in grouped based in terms of recording site, (a.) Edinburgh (b.) Idiap and (c.) TNO.

and TNO no degradation in performance was observed.

We next compared the performance of multimodal features extracted from lexical, prosodic, conversational and motion information against social role posterior features. Table 6.2 reports the topic segmentation results from literature [132] that is closest to our experimental setup. In Table 6.2, conversational features are the best performing feature group. However, as described in Section 6.4.2, conversational features also incorporate lexical information from LCseg algorithm. Comparing the numbers reported in Table 6.2 and Table 6.3 reveals that role

posterior features outperform all the individual feature groups for both TOPSEG and ALLSEG contexts. The TOPSEG performance of role posterior features is comparable to the model trained on combination of all multimodal features. However, in the context of ALLSEG, where performance is evaluated over both top level and sub topic boundaries, social role posterior features are better compared to combined multimodal features. In fact, previous studies [132] have also suggested that feature combination is less helpful in predicting fine grained sub topic boundaries compared to coarser top level topic boundaries. In comparison, our analysis suggests that modeling changes in social role dynamics is also useful for predicting sub topic level changes.

6.4.4 Results for unsupervised topic segmentation using ddCRP

Comparison between CRP and ddCRP priors

The inclusion of prior information for unsupervised topic segmentation is one of the main motivations for investigation ddCRP framework in this work. As described in Section 6.3, traditional CRPs produce exchangeable distributions which ignore prior information about topic segmentation. We analyzed those properties of CRP and ddCRP by simulating draws from the two distributions. Figure 6.5 shows the plot of seating arrangement of customers (turns) at various tables (topic segments) derived from draws from CRP and ddCRP. Figure 6.5a and Figure 6.5b shows the plot obtained from CRP for different hyperparameter settings. Figure 6.5c and Figure 6.5d represent the case of sequential ddCRP with distance between adjacent customers given by average KL divergence between social role posteriors and a logistic decay function. We can observe that top two plots produce dispersive partitions where customers at the end of meeting can be clustered with customers at the start. In comparison, the bottom two plots demonstrate a prior for which nearby customers share the same cluster. The positions where the distance between customers is large are likely candidates for occupation of new tables.

Evaluation results

In order to validate the effectiveness of our proposed ddCRP model for topic segmentation, we applied it on 100 AMI scenario meetings. We also compared ddCRP against three baseline models: lexical cohesion based LCseg [36], Bayesseg [29] and PLDA [85]. LCseg results are those reported in [132]. For Bayesseg we used the optimal configurations described in [29]. The hyperparameters for the PLDA model were set to values ($\alpha = 0.1, \beta = 0.01, \gamma = 0.01$) as described in [85]. To make a fair comparison between PLDA and our approach, we retained the same hyperparameter values while training ddCRP model. For both PLDA and ddCRP, 5 randomly initialized Gibbs chains we used. Each chain ran for 30000 iterations with 25000 for burn-in, then 200 samples were averaged to estimate the posterior probability of topic boundary at each turn. These probabilities were compared against a threshold that was set in order to yield the number of topic boundaries equal to those in reference segmentation

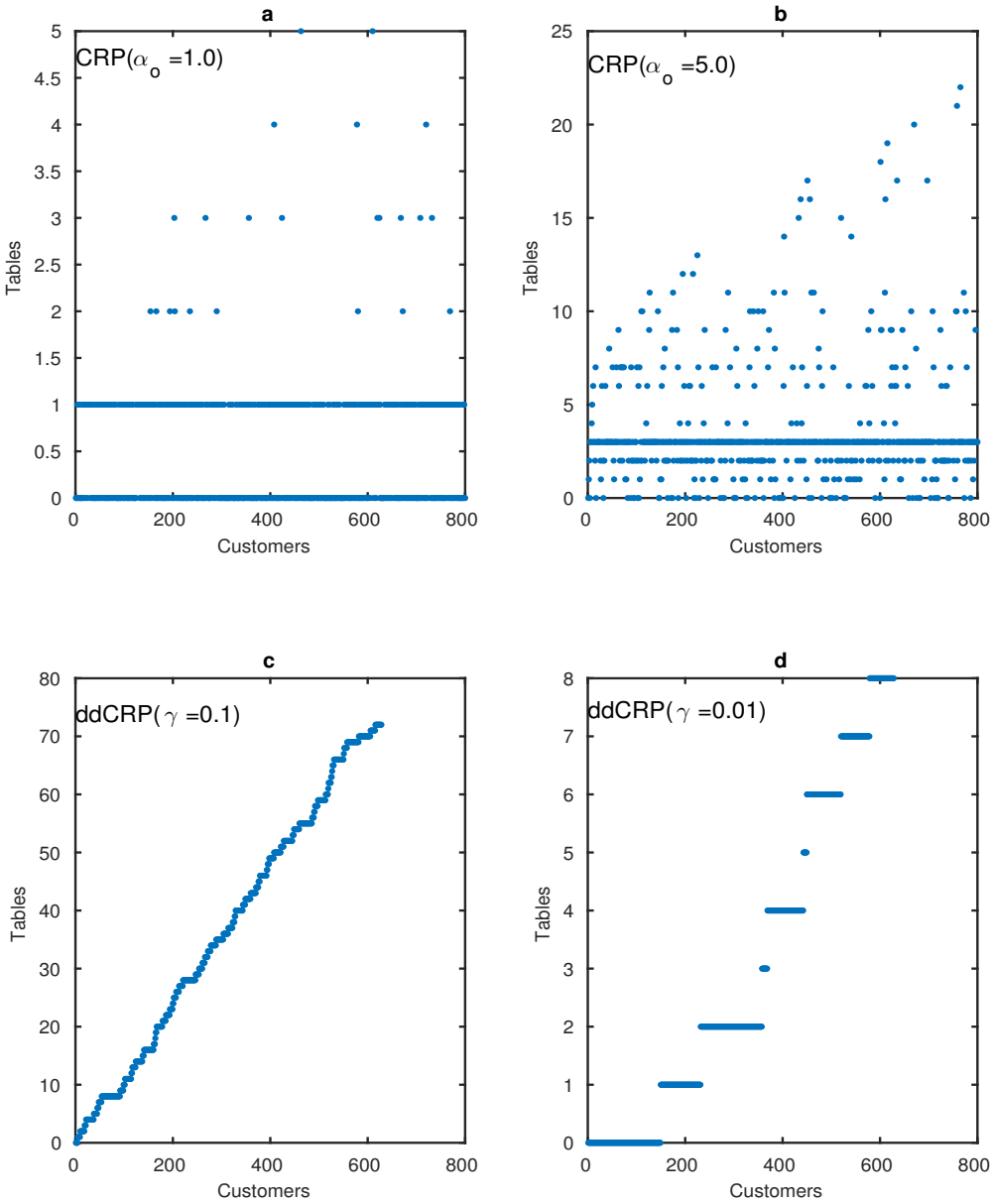

Figure 6.5: Simulated draws from CRP and ddCRP. CRP draws in (a and b) are dispersive. In comparison, ddCRP draws (c and d) show the property of linear segmentation.

(ALLSEG).

We investigated the performance of ddCRP model by comparing the P_k score against random baseline. Figure 6.6 shows the P_k scores for different values of number of latent topics K. We can observe that ddCRP model outperforms random baseline for all conditions. In Figure 6.6, $K = 16$ shows the best performance. Furthermore, when the K is fixed to a lower number (4), the performance of the model is relatively worse compared to higher values of K.

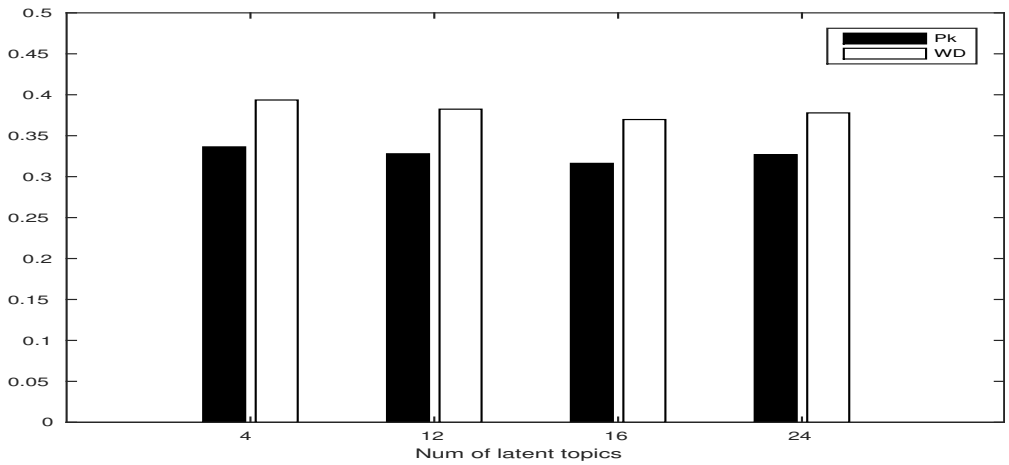

Figure 6.6: (P_k and WD) scores as the number of latent topics in the ddCRP model is varied.

Table 6.4: Topic change detection performance for latent topic based models. The last row in the table shows the results for social role based ddCRP model.

Model	P_k	WD
LCseg	0.40	0.47
Bayesseg	0.34	0.41
PLDA	0.34	0.40
ddCRP	0.32	0.37

Table 6.4 shows the topic segmentation of various models on AMI scenario meetings. We observe that topic modeling methods (PLDA, Bayesseg and ddCRP) outperform those based on lexical cohesion (LCseg), which is consistent with previously reported results [29]. Moreover, among the topic model based methods, ddCRP achieves the best performance on both P_k and WD measures. ddCRP shows 6% and 7.5% relative improvement over PLDA on P_k and WD measure respectively. We performed statistical tests to examine the difference between performance of latent topic based models. The null hypothesis being tested is that performance of models in Table 6.4 is same and the observed differences are due to random events. We applied a non parametric method based on Friedman test [26] to rank the performance of each unsupervised model. The average rank of each model is used to compute the Friedman statistic, which under null hypothesis is distributed according to F-distribution. For the results in Table 6.4 we reject the null hypothesis ($F(2,198) = 3.43; \alpha = 0.05$). Since the null hypothesis was rejected we performed post hoc (Nemenyi) tests to compare three models with each other. The post hoc tests revealed that ddCRP is statistically significant ($\alpha = 0.05$) compared to both PLDA and Bayesseg. We note that similar hyperparameters (α, β) and number of latent topics ($K = 16$) were used while training both PLDA and ddCRP models. This suggests that improvement in performance over PLDA is due to inclusion of social role information as KL distance in ddCRP.

6.4.5 Experiments: ICSI corpus

We also applied the methods proposed for topic segmentation in this work to ICSI corpus meetings [53]. ICSI corpus contains a collection of 75 meetings, out of which 25 meetings were annotated with reference topic segmentation [36]. In comparison to AMI meetings, ICSI meetings were annotated only at a coarser level with an average of 5 – 6 topic segments per meeting.

Table 6.5: Topic segmentation results for social role posterior features and baseline evaluated on 25 meetings from ICSI corpus. The automatic social role recognition model was trained on AMI corpus.

Model	P_k	WD
Random	0.43	0.46
Social Role posteriors	0.32	0.37

The meetings in ICSI corpus were not annotated with social role labels. Instead, we directly applied the social role recognition model trained on AMI scenario meetings. However, training (AMI) and testing (ICSI) meetings are different in several aspects. Compared to scripted scenario of AMI meetings, ICSI corpus contains natural meetings where participants discuss real life issues. Moreover, there are four participants present in AMI scenario meetings. In comparison, the number of participants per meeting in ICSI corpus varies from 3 to 9. We extracted the different features described in Section 3.3 for each speaker over different slices in a meeting. The role recognition model, described in Section 3.4, was then applied over the extracted features to estimate the social role posteriors.

Our objective was to determine whether social roles automatically predicted on ICSI meetings can be used for topic segmentation. To evaluate the performance of our approach, we trained a supervised classifier (Boosting) on ICSI meetings. We performed a 25 fold leave one out crossvalidation on the set of 25 meetings annotated with topic segmentation. Since the number of speakers of speakers across meetings is not constant, the social role posterior feature set was ordered according to four most active speakers in a meeting slice. Table 6.6 reports the result of our evaluation using standard P_k and WD measures. In comparison to random baseline, we observe that social roles posterior features are effective in predicting topic boundaries. From these results we can infer that automatic role recognition model generalizes to natural meetings in ICSI corpus. Moreover, changes in social roles configuration of speakers in natural meetings are informative for predicting topic changes. Figure 6.7 shows the details of comparison between randomly placing segment boundaries and social role based topic segmentation for 25 ICSI meetings. We can observe that for a large majority of meetings our approach compares favorably against random baseline. However, there are 4 noticeable exceptions where the role based approach over predicts the number of topic boundaries.

We next evaluated the performance of unsupervised ddCRP model for topic segmentation. The model was trained on both labeled 25 meetings and unlabeled 50 meetings. The ddCRP model

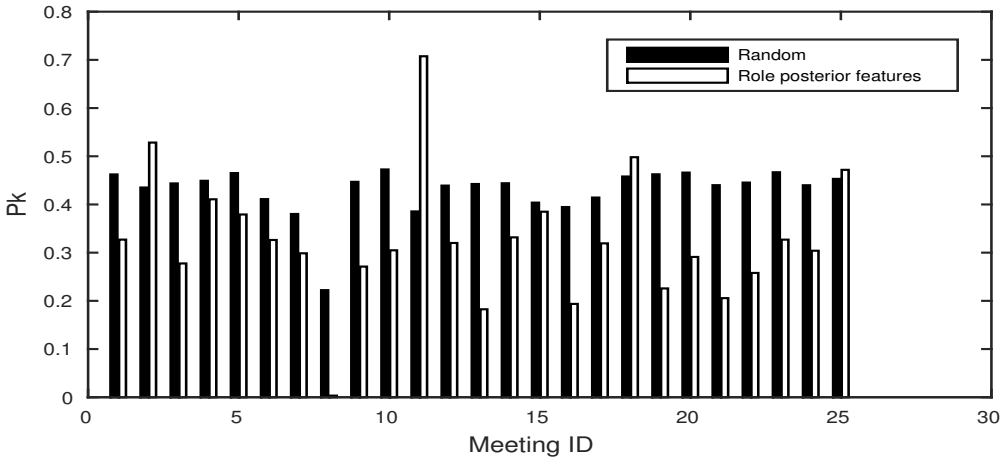

Figure 6.7: P_k values for 25 ICSI meetings.

was trained using the same configuration described for AMI meetings. Table 6.6 reports the results for various unsupervised models. The performance of LCseg and Bayesseg corresponds to those reported in [29]. Compare to results on AMI corpus, Table 6.6 shows that LCseg performs much better. Among the Bayesian models, Bayesseg, PLDA and ddCRP achieve lower P_k score. However, compared to AMI meetings, ddCRP does not result in any improvement over other unsupervised approaches. A plausible reason for this behavior could be because topic segments in ICSI meetings are marked at much coarser level and have on an average much lower number of topic shifts compared to AMI meetings.

Table 6.6: Topic change detection performance of various unsupervised models. The last row corresponds to case of ddCRP with social role distance.

Model	P_k	WD
LCseg	0.31	0.32
Bayesseg	0.26	0.32
PLDA	0.27	0.33
ddCRP	0.27	0.33

6.5 Conclusions

Detecting topic changes is an important step towards automatic access and retrieval of information in multiparty interactions. Meetings are characterized by a hierarchical structure which is reflected in the coarse and fine grained segmentation of conversation into multiple topics. In this chapter, we investigated the potential of group social roles for topic segmentation in meetings. Our results show that a supervised classifier trained using social role posterior features improves on previous work, that uses a combination of lexical, prosodic, conversational and motion boundary features. Furthermore, our results on topic segmentation on ICSI meetings show that social roles posterior features generalize even on previously unseen scenarios of multiparty interaction.

We also presented an unsupervised method which combines latent topic model with social role information. The changes in social role configuration of a group were used as prior information in a ddCRP topic segmentation framework. The lexical content of speech utterances was modeled using LDA. Experimental evaluation of ddCRP model that incorporates social role information, showed that this method improves topic segmentation performance on AMI meetings compared to state-of-the-art unsupervised models.

7 Conclusion

This thesis addresses several challenges related to automatic structuring of spontaneous multiparty interactions, i.e., meetings. Our work demonstrates that social roles capture the behavior of speakers in meetings and we used this finding to improve short term segmentation of meeting into distinct speakers and long term segmentation of meeting into different topics.

We presented an approach for automatic recognition of social roles that emerge in small group meetings. Our work has been performed over the largest annotated database for this task, both in terms of number of unique speakers and number of annotated meetings. We considered various short term and long term features for recognition of social roles. The short term features were computed over short time windows and represent the influence of social roles on turn taking patterns. The long term features were computed over an entire meeting slice and capture the linguistic style and vocal expression of speakers. The role recognition system was modeled by extending the framework of CRFs and integrates feature information at multiple time scales. Experiments revealed that automatically extracted speaker interaction features and long term features are useful cues for predicting social roles. CRF trained using these features was able to perform non trivial classification of four social roles, reaching a recognition accuracy of 74% on the scenario portion of AMI corpus. We also demonstrated that automatic role recognition system generalizes across various scenarios of multiparty interaction.

Our work considered, how various verbal and non verbal features perform in the task of predicting formal and social role on the same dataset. Experiment results revealed that, in comparison to social roles, speaker behavior extracted from relatively thin meeting slices is insufficient to predict formal roles. Furthermore, even when feature are extracted from the entire recording, verbal features alone were the best predictors of formal roles. In comparison, prediction of social roles significantly improves from combination of both verbal and non verbal features. In Chapter 4, we also presented an unsupervised lexical modeling approach based on LDA. This approach extracts a compact representation of verbal information in terms of latent topics and reached an accuracy of 69% in recognizing formal roles.

We demonstrated that social roles can be used to improve speaker diarization when audio is recorded using distant microphones. This was a challenging task as both the number of speakers and their associated speaking times were not known. We examined two limitations of current state-of-the-art speaker diarization systems, i.e., a speaker independent minimum duration constraint and a uniform prior on speaker interaction patterns. Our study extends the commonly used speaker diarization system based on HMM-GMM modeling, by including social role information in the speaker segmentation step. Analysis of social role statistics revealed that turn duration of speakers was influenced by the social roles. Social roles, such as protagonists were more likely to produce longer turns compared to speakers assuming a neutral role. Also, most probable transitions between speakers correspond to protagonists and gatekeepers and their interaction with other roles. This knowledge was used as prior information in speaker diarization system. Results on AMI corpus revealed that HMM-GMM system augmented with social role information achieved a 16% improvement over baseline HMM-GMM system. We also showed that social role statistics generalize on NIST meetings and diarization results revealed a 13% improvement compared to baseline system.

In Chapter 6, we demonstrated the application of social roles for the task of topic segmentation in meetings. Social roles capture the group dynamics in meetings and our investigation considered whether changes in social roles also indicate a shift in discourse of the meeting. For comparison, we selected state-of-the-art supervised classification method based on extracting a wide variety of multimodal features as baseline. Experiment results based on AMI meetings confirmed that supervised classification using social role posterior features is useful for topic segmentation. Our analysis further revealed that topic segmentation performance using social roles is comparable to baseline method when coarse grained top level topics are considered. However, when performance is also measured at fine grained sub topic level, social roles ($P_k = 0.30$) outperform baseline method ($P_k = 0.35$). In Chapter 6, we also developed a novel framework for unsupervised topic segmentation which combines social roles with latent topic models in a ddCRP framework. Experiments demonstrated that social roles improve the performance of lexical generative models for topic segmentation in AMI meetings.

7.1 Future directions

Automatic modeling of social roles in meetings can be extended in several directions. While we explored several verbal and non verbal features for social role recognition, our approach was limited to feature extraction from audio data alone. However, several studies in social psychology [57, 65] have shown that participants display visual cues through gestures, postures and gaze to convey information. Although some initial work [133] has investigated frame level hand and body motion features for social role recognition, other important cues, such as gaze and head orientation have yet to be investigated. Automatically extracting gaze and head orientation information from visual data can be used to understand the focus of visual attention in face to face meetings. Furthermore, recognizing gestures, such as nodding, sharing objects, etc., that indicate consent and cooperation might also be other interesting features

to explore. In terms of speech data, high level features that identify sarcasm, irony, humor, chuckle in a meeting could serve as complementary features to acoustic features considered in this work.

Automatic social role recognition under distant microphone conditions is a challenging task. While social role based diarization system improves the performance over HMM-GMM baseline, there is still lot of scope for improvement in this area. A major limitation in our approach is due to inadequacy of HMM-GMM system and social role based diarization system to handle overlapping speech. Overlapping speech is caused when two or more participants speak simultaneously. Unlike broadcast news conversations, overlapping speech is a common phenomena in spontaneous conversations like meetings. Several previous studies [49, 48, 58] have shown that a major source of errors in state-of-the-art speaker diarization systems is due to presence of overlapping speech. Inability to model overlapping speech can adversely effect automatic social role recognition as the errors in speaker segmentation stage cause corruption in features used by role recognition model. A possible future direction should therefore include overlapping speech detection as an additional problem to explore.

Meetings are characterized by a hierarchical structure, in which coarse grained topics are composed of fine grained sub topics. The supervised topic segmentation system described in this work was trained separately for top topic and sub topic segmentation. However, unsupervised topic segmentation framework considered in this work and other unsupervised state-of-the-art approaches presume a linear structure for topic segments. In future, we could explore alternative Bayesian models [18] with hierarchy of latent variables in which latent variables at lower level are related to latent variables at higher level.

A possible future research direction would be to explore other factors that influence speaker behavior. In social psychology, personality refers to a person's characteristic pattern of behavior [35]. The relationship between personality, social roles and team effectiveness has been explored in [106]. Recent studies [83, 116] in automatic modeling of personality perception have shown that traits like extroversion and introversion can be estimated in multiparty interactions. Since personality remains relatively stable across different situations, an interesting research direction would be to explore the relation between a participant's personality and its influences on role taking behavior.

Bibliography

[1] http://www.nist.gov/speech/tests/tdt/.

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

[3] Jitendra Ajmera. *Robust audio segmentation*. PhD thesis, Ecole polytechnique fédérale de Lausanne, 2004.

[4] N. Ambady and R. Rosenthal. Thin Slices of Expressive behavior as Predictors of Interpersonal Consequences : a Meta-Analysis . *Psychological Bulletin*, 111(2):256–274, 1992.

[5] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando. 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, June 2006.

[6] Xavier Anguera. *Robust speaker diarization for meetings*. PhD thesis, Universitat Politecnica de Catalunya, 2006.

[7] Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6), November 1974.

[8] R. F. Bales. A Set of Categories for the Analysis of Small Group Interaction. *American Sociological Review*, 15(2), 1950.

[9] R.F. Bales. *Personality and interpersonal behavior*. New York: Holt, Rinehart and Winston, 1970.

[10] S. Banerjee and A. Rudnick. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. *Proceedings of ICSLP*, 2004.

[11] Satanjeev Banerjee, Carolyn Penstein Rosé, and Alexander I. Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In Maria Francesca Costabile and Fabio Paternò, editors, *INTERACT*, volume 3585 of *Lecture Notes in Computer Science*, pages 643–656. Springer, 2005.

[12] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. The rules behind roles: Identifying speaker role in radio broadcasts. *Proceedings of AAAI*, 2000.

[13] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1-3):177–210, February 1999.

[14] K. D. Benne and P. Sheats. Functional roles of group members. *Journal of social issues*, 4, 1948.

[15] B. J. Biddle. *Role theory : expectations, identities, and behaviors*. Academic Press, 1979.

[16] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[17] David M. Blei and Peter I. Frazier. Distance dependent chinese restaurant processes. *J. Mach. Learn. Res.*, 12:2461–2488, November 2011.

[18] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, February 2010.

[19] Susanne Burger, Victoria MacLaren, and Hua Yu. The isl meeting corpus: the impact of meeting type on speech style. In John H. L. Hansen and Bryan L. Pellom, editors, *INTERSPEECH*. ISCA, 2002.

[20] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, June 1996.

[21] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41:181–190, 2007.

[22] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[23] Christopher Cieri, David Miller, and Kevin Walker. Research methodologies, observations and outcomes in (conversational) speech data collection. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 206–211, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[24] G. Damnati and D. Charlet. Robust speaker turn role labeling of TV Broadcast News shows. *proceedings of ICASSP*, 2011.

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

[26] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[27] John Dines, Jithendra Vepa, and Thomas Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Ninth IEEE International conference on Spoken Language Processing*, 2006.

[28] W. Dong, B. Lepri, F. Pianesi, and A. Pentland. Modeling functional roles dynamics in small group interactions. *IEEE Transactions on Multimedia*, 15(1):83–95, 2013.

[29] Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 334–343, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[30] Maxine Eskenazi, Alexander I. Rudnicky, Karin Gregory, Paul C. Constantinides, Robert Brennan, Christina L. Bennett, and Jwan Allen. Data collection and processing in the carnegie mellon communicator. In *EUROSPEECH*. ISCA, 1999.

[31] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, MM '10, 2010.

[32] S. Favre, A. Dielmann, and A. Vinciarelli. Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In *ACM International Conference on Multimedia*, 2009.

[33] J.L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[34] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, 2010.

[35] D. C. Funder. Personality. *Annual Review of Psychology*, 52:197–221, 2001.

[36] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 562–569, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[37] N. Garg, S. Favre, D. Hakkani-Tur, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and social network analysis. *Proceedings of the ACM Multimedia*, 2008.

[38] John S. Garofolo, Christophe Laprun, Martial Michel, Vincent M. Stanford, and Elham Tabassi. The nist meeting room pilot corpus. In *LREC*. European Language Resources Association, 2004.

[39] J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520 vol.1, Mar 1992.

[40] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 2010.

[41] John Grothendieck, Allen L. Gorin, and Nash Borges. Social correlates of turn-taking style. *Comput. Speech Lang.*, 25(4):789–801, October 2011.

[42] Alexander Gruenstein, John Niekrasz, and Matthew Purver. Meeting structure annotation – annotations collected with a general purpose toolkit. *Text, Speech and Language Technology*, pages 247–274, 2008.

[43] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. In *INTERSPEECH*, pages 1117–1120. ISCA, 2005.

[44] T. Hain, J. Vepa, and J. Dines. The segmentation of multichannel meeting recordings for automatic speech recognition. *Proceedings of Interspeech*, 2006.

[45] T. Hain, V. Wan, L. Burget, M. Karafiat, J. Dines, J. Vepa, G. Garau, and M. Lincoln. The AMI System for the Transcription of Speech in Meetings. *Proceedings of Icassp*, 2007.

[46] A.P. Hare. Types of roles in small groups: a bit of history and a current perspective. *Small Group Research*, 25, 1994.

[47] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, March 1997.

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

[49] Marijn Huijbregts, David A. van Leuwen, and Franciska M. G. de Jong. Speech overlap detection in a two-pass speaker diarization system. In *Interspeech*, pages 1063–1066, Brighton, United Kingdom, 2009.

[50] G. Hutchinson, B. Zhang, and M. Ostendorf. Unsupervised broadcast conversation speaker role labeling. *Proceedings of ICASSP*, 2010.

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

[52] Ajmera J. and C. Wooters. A robust speaker clustering algorithm. *IEEE Automatic Speech Recognition Understanding Workshop*, 2003.

[53] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The icsi meeting corpus. In *ICASSP*, pages 364–367, Hong Kong, 2003.

[54] D. B. Jayagopi and D. Gatica-Perez. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Trans. on Multimedia*, 2010.

[55] Xavier Anguera Jose M. Pardo and Check Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, 56(9), September 2007.

[56] Han K.J. and Narayanan S.S. Improved speaker diarization of meeting speech with recurrent selection of representative speech segments and participant interaction pattern modeling. In *Proceedings of Interspeech*, 2009.

[57] M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth, 2005.

[58] Mary Tai Knox, Nikki Mirghafori, and Gerald Friedland. Where did I go wrong?: Identifying troublesome segments for speaker diarization systems. In *Interspeech*, Portland, USA, 2012.

[59] T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284, 1998.

[60] K. Laskowski, M. Ostendorf, and T. Schultz. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[61] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)):503–528, 1989.

[62] Y. Liu. Initial study on automatic identification of speaker role in broadcast news speech. *Proceedings of HLT/NAACL*, 2006.

[63] Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10(7):504–516, Oct 2002.

[64] D. J. C. Mackay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, 2003.

[65] Valerie Manusov and Miles L. Patterson. *The SAGE Handbook of Nonverbal Communication*. SAGE Publications, 2006.

[66] Mark T. Maybury. Discourse cues for broadcast news segmentation. In Christian Boitet and Pete Whitelock, editors, *COLING-ACL*, pages 819–822. Morgan Kaufmann Publishers / ACL, 1998.

[67] Andrew Kachites. McCallum. Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu*, 2002.

[68] Joseph E. Mcgrath. Time, Interaction, and Performance (TIP): A Theory of Groups. *Small Group Research*, 22(2):147–174, May 1991.

[69] G. H. Mead. *Mind, self, and society.* University of Chicago Press, 1934.

[70] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. In *Journal of Personality and Social Psychology*, 2006.

[71] Djamel Mostefa, Nicolas Moreau, Khalid Choukri, Gerasimos Potamianos, Stephen M. Chu, Ambrish Tyagi, Josep R. Casas, Jordi Turmo, Luca Cristoforetti, Francesco Tobia, Aristodemos Pnevmatikakis, Vasileios Mylonakis, Fotios Talantzis, Susanne Burger, Rainer Stiefelhagen, Keni Bernardin, and Cedrick Rochet. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407, 2007.

[72] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[73] Patrick Nguyen, Luca Rigazio, Yvonne Moh, and Jean-Claude Junqua. Rich transcription 2002 site report panasonic speech technology laboratory (PSTL). In *2002 Rich Transcription Workshop*, 2002.

[74] Tin Lay Nwe, Hanwu Sun, Haizhou Li, and Susanto Rahardja. Speaker diarization in meeting audio. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:4073–4076, 2009.

[75] Andrew Olney and Zhiqiang Cai. An orthonormal basis for topic segmentation in tutorial dialogue. In *HLT/EMNLP*. The Association for Computational Linguistics, 2005.

[76] Vincenzo Pallotta, Hatem Ghorbel, Afzal Ballim, Agnes Lisowska, and Stéphane Marchand-Maillet. Towards meeting information systems: Meeting knowledge management. In *ICEIS (3)*, pages 464–469, 2004.

[77] Jose Pardo, Xavier Anguera, and Chuck Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Trans. Comput.*, 56:1189–1224, September 2007.

[78] Jose M. Pardo, Xavier Anguera, and Chuck Wooters. Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. In *ICSLP*, pages 2194–2197, 2006.

[79] Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

[80] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, August 2005.

[81] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 2003.

[82] Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, 28(1):19–36, March 2002.

[83] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 53–60, New York, NY, USA, 2008. ACM.

[84] T. Polzehl, S. Moller, and F. Metze. Automatically assessing personality from speech. In *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, 2010.

[85] Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 17–24, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[86] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*, 2004.

[87] Elias Rentzeperis, Andreas Stergiou, Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros C. Polymenakos. 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, 2006. Springer-Verlag.

[88] Jeffrey C. Reynar. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 357–364, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[89] Chen S. and Gopalakrishnan P. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Proceedings of the DARPA Workshop, 1998*.

[90] H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4, Part 1):696–735, December 1974.

[91] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7):1373–1380, 2009.

[92] D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez. Assessing the impact of language style on emergent leadership perception from ubiquitous audio. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, 2012.

[93] A. Sapru and H. Bourlard. Automatic social role recognition in professional meetings using conditional random fields. In *Proceedings of Interspeech*, 2013.

[94] A. Sapru and H. Bourlard. Investigating the impact of language style and vocal expression on social roles of participants in professional meetings. In *Affective Computing and Intelligent Interaction*, page 6, September 2013.

[95] A. Sapru and H. Bourlard. Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5):746–760, May 2015.

[96] A. Sapru and F. Valente. Automatic speaker role labeling in AMI meetings: recognition of formal and social roles. *Proceedings of Icassp*, 2012.

[97] A. Sapru, S.H. Yella, and H. Bourlard. Improving speaker diarization using social role information. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 101–105, May 2014.

[98] Ashtosh Sapru and Hervé Bourlard. Detecting speaker roles and topic changes in multiparty conversations using latent topic models. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2882–2886, 2014.

[99] R. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 2000.

[100] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings of Interspeech*, 2013.

[101] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

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

[103] Melissa Sherman and Yang Liu. Using hidden markov models for topic segmentation of meeting transcripts. In Amitava Das and Srinivas Bangalore, editors, *SLT*, pages 185–188. IEEE, 2008.

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

[105] P. E. Slater. Role Differentiation in Small Groups. *American Sociological Review*, 20(3), 1955.

[106] Greg L. Stewart, Ingrid S. Fulmer, and Murray R. Barrick. An exploration of member roles as a multilevel linking mechanism for individual traits and team outcomes. *Personnel Psychology*, 58(2):343–365, Sum 2005.

[107] Nicola Stokes, Joe Carthy, and Alan F. Smeaton. Select: a lexical cohesion based news story segmentation system. *AI Commun.*, 17(1):3–12, 2004.

[108] A. Stolcke. Srilm – an extensible language modeling toolkit. *Proc. of ICSLP*, 2002.

[109] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[110] A. Temko, D. Macho, and C. Nadeu. 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, April 2007.

[111] H. Tischler. *Introduction to Sociology*. Cengage Learning, 2006.

[112] N Tishby, F Pereira, and W Bialek. The information bottleneck method. In *NEC Research Institute TR*, 1998.

[113] Alain Tritschler and Ramesh A. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *EUROSPEECH*. ISCA, 1999.

[114] Gökhan Tür, Andreas Stolcke, Dilek Hakkani-Tür, and Elizabeth Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Comput. Linguist.*, pages 31–57, 2001.

[115] F. Valente and A. Vinciarelli. Language-Independent Socio-Emotional Role Recognition in the AMI Meetings Corpus. *Proceedings of Interspeech*, 2011.

[116] Fabio Valente, Samuel Kim, and Petr Motlícek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *INTERSPEECH*. ISCA, 2012.

[117] Fabio Valente, Deepu Vijayasenan, and Petr Motlicek. Speaker diarization of meetings based on speaker role n-gram models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

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

[119] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 250–255, Kyoto, Japan, 2007.

[120] A. Vinciarelli. Sociometry based multiparty audio recordings summarization. In *ICPR (2)*, pages 1154–1157. IEEE Computer Society, 2006.

[121] A. Vinciarelli and S. Favre. Broadcast news story segmentation using social network analysis and hidden markov models. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 261–264, New York, NY, USA, 2007. ACM.

[122] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *Multimedia, IEEE Transactions on*, 11(2):256–271, Feb. 2009.

[123] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller. The voice of leadership: models and performances of automatic analysis in online speeches. *IEEE Transactions on Affective Computing*, 2012.

[124] T. Wilson and G. Hofer. Using linguistic and vocal expressiveness in social role recognition. In Pearl Pu, Michael J. Pazzani, Elisabeth André, and Doug Riecken, editors, *IUI*, pages 419–422. ACM, 2011.

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

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

[127] D. Wrede and E. Shriberg. Spotting "hotspots" in meetings: Human judgments and prosodic cues. *Proceedings of Eurospeech*, 2003.

[128] Anguera X., Wooters C., and Hernando H. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), September 2007.

[129] S. Yaman, D. Hakkani-Tur, and G. Tur. Social Role Discovery from Spoken Language using Dynamic Bayesian Networks. *Proceedings of Interspeech*, 2010.

[130] J. P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden markov model approach to text segmentation and event tracking. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 333–336, Seattle, WA, 1998. IEEE.

[131] Pei yun Hsueh and Johanna D. Moore. Automatic topic segmentation and labeling in multiparty dialogue. In *SLT*, pages 98–101, 2006.

[132] Pei yun Hsueh and Johanna D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In John A. Carroll, Antal van den Bosch, and Annie Zaenen, editors, *ACL*. The Association for Computational Linguistics, 2007.

[133] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In Francis K. H. Quek, Jie Yang, Dominic W. Massaro, Abeer A. Alwan, and Timothy J. Hazen, editors, *ICMI*, pages 28–34. ACM, 2006.

Ashtosh Sapru

303-3, Idiap Research Institute,
Rue Marconi 19, Martigny, Switzerland 1920.
Telephone(office): (+41) 277217792
E-mail: ashtosh.sapru@idiap.ch

EDUCATION

PhD (March 2011 - April 2015)
- Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

ME (August 2004 - June 2006)
- Major: Signal Processing
- Indian Institute of Science, Bengaluru, Karnataka, India.

BSc Engineering (August 2000 - July 2004)
- Major: Electrical Engineering
- National Institute of Technology, Jamshedpur, Jharkhand, India.

RESEARCH INTERESTS

Speech processing, Social computing, Information retrieval, Applied machine learning.

Professional Experience

- February 2011 to till Date: Working as a research assistant at Idiap Research Institute, Martigny, Switzerland.
- August 2006 to January 2011: Worked as a Senior Software Engineer at Aricent, Bangalore, India. Was involved with various projects related to speech and audio coding for mobile phones.

PUBLICATIONS

- 'Automatic Recognition of Emergent Social Roles in Small Group Interactions', Ashtosh Sapru and Hervé Bourlard, in IEEE Transactions on Multimedia (in print), 2015.
- 'Detecting speaker roles and topic changes in multiparty conversations using latent topic models', Ashtosh Sapru and Hervé Bourlard, in Proceedings of Interspeech, 2014.
- 'Improving Speaker Diarization using social role information', Ashtosh Sapru, Sree Harsha Yella and Hervé Bourlard, in Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.
- 'Automatic Social Role Recognition In Professional Meetings Using Conditional Random Fields', Ashtosh Sapru and Hervé Bourlard, in Proceedings of Interspeech, 2013.
- 'Investigating the Impact of Language Style and Vocal Expression on Social Roles of Participants in Professional Meetings', Ashtosh Sapru and Hervé Bourlard, in Affective Computing and Intelligent Interaction, 2013.
- 'Automatic Speaker Role Labeling in AMI Meetings: Recognition of Formal and Social Roles', Ashtosh Sapru and Fabio Valente, in Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan, 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 Ashtosh Sapru, in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2011.

- ‘Automotive media player - Efficient design and implementation of essential features’, Vasanth Rajgopal, Ravi Lakkundi, Tony James, Ashtosh Sapru, in IEEE International Conference on Consumer Electronics (ICCE), 2011.
- ‘An efficient Huffman decoding method using concurrent execution in ARM Cortex-A8’, Tony James, Ashtosh Sapru, Vasanth Rajgopal, in in IEEE International Conference on Consumer Electronics (ICCE), 2011.
- ‘Improve detection performance of speech recognizer in an automotive environment’, Ashtosh Sapru, Ravi Lakkundi, Nisar Ahmed, in Asilomar Conference on Signals, Systems and Computers, 2008.
- ‘Gesture Recognition by Line Fitting Over Significant Pixels’, Ashtosh Sapru, Ravi Lakkundi, Nisar Ahmed, in EEE International Symposium on Signal Processing and Information Technology, 2008.