Automatic Recognition of Emergent Social Roles in Small Group Interactions

Ashtosh Sapru, *Student Member, IEEE,* and Hervé Bourlard, *Fellow, IEEE*

***Abstract*—This paper investigates automatic recognition of social roles that emerge naturally in small groups. These roles represent a flexible classification scheme that can generalize across different scenarios of small group interaction. We systematically investigate various verbal and non verbal cues extracted from turn taking patterns, vocal expression and linguistic style to model speakers behavior. 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 social role recognition using conditional random fields achieves an accuracy of 74% in classifying four social roles and outperforms the baseline method on all social role categories. Furthermore, we also demonstrate the effectiveness of our approach by evaluating it on previously unseen scenarios of small group interaction.**

***Index Terms*—Small group interactions, Social roles, Crowdsourcing, Conditional random fields**

I. Introduction

Roles are one of the most important concepts in understanding human social behavior. 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 [1]. 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 interactions social roles can be broadly categorized as formal and informal [2]. 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. 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. In comparison to formal roles, informal roles are not designated as positions in a group.

The focus of this work are the informal social roles that emerge in small group interactions. These roles are related to communicative functions that participants perform in the group. 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 [3]. Verbal cues are not the only mechanism used in group communication, participants also display non verbal behavior characteristics through vocal expression, body and facial gestures and language style [4], [5]. Effective communication also requires that participants alternate between listening and speaking states and organize the conversation by taking turns [6]. From a computational modeling perspective, communicative functions of a speaker need to be represented as behavioral patterns that can be automatically extracted using standard tools like automatic speech recognition (ASR), speech activity detection (SAD) and prosody extraction.

Automatic recognition of social roles in meetings is a key area of research in the emerging domain of social computing. It is also complementary to other phenomena studied in meetings like social dominance, engagement and hot-spots [7], [8]. Computational modeling of roles in meetings is challenging, especially compared to broadcast domain, due to presence of disfluency in speech, frequent overlaps and short speaker turns. Furthermore, speaker segmentation and ASR systems used for extracting relevant features produce significantly higher errors in meetings. However, role recognition in meetings is worth further investigation as it can serve as an important tool for structuring and analyzing social interactions [9]. Speaker role information can be used in applications like multimedia data indexing, enhancing media browsers for meeting recordings, segmenting topically homogeneous segments in conversation discourses [10] and summarization of spoken documents [11].

Automatic social role recognition in small group meetings has received increasing attention in recent years. Previous studies [12], [13] have explored formal role recognition on large corpus of meetings. However, roles in this case were imposed by the scenario of meetings, making it difficult to apply the learned model on other scenarios of small group interaction. Other approaches, such as [14], [15] have investigated recognition of emergent social roles in meetings. While these approaches reveal that it is possible to automatically recognize emergent social roles, databases used in these studies were relatively smaller and the performance of recognition models on less populated roles was far from convincing. Furthermore, previous approaches have explored a limited set of features for role recognition and a systematic comparison of different features is still lacking.

This work presents a detailed study on automatic recognition of emergent social roles in small group meetings and contains several new contributions. The corpus we annotated for recognizing emergent social roles in meetings has four times as many speakers compared to similar investigations [16]. 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. The proposed approach models speaker behavior in terms of linguistic, turn taking and acoustic information. In comparison

Authors are affiliated with Idiap Research Institute, Martigny, Switzerland and Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. e-mail: asapru@idiap.ch, bourlard@idiap.ch

to earlier approaches [15], [14], 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 novel classification framework 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 paper is organized as follows. Section II reviews the literature on social roles in social psychology and social computing related to our work. In Section III, we discuss the dataset, description of social roles used in this work, and describe the process for annotation of roles. Section IV presents the various features that are automatically extracted from turn taking, vocalic and linguistic behavior of speakers. In Section V, we propose the supervised learning model for automatically recognizing the social roles of speakers from the extracted features. The experimental methodology for role classification is presented in Section VI, where we also compare and discuss the performance of proposed method. The paper is then concluded in Section VII.

II. Related Work

In the next subsections we summarize the most relevant work in social psychology and social computing related to our own.

A. Roles in Social Psychology

The concept of social roles has been a subject of analysis for over 80 years [1]. In social psychology literature, roles have been defined as characteristic behavior patterns of one or more persons in a context [17]. According to [17], 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 [17]. From the viewpoint of this work, we are interested in functional roles that emerge 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 [18], the 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 focuses 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 [18], successful groups follow a flexible role structure which allows same person with multiple talents to assume different roles.

According to [3], [19], 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 [3], 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.

B. Formal Role Recognition in Broadcast Domain

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. A summary of these works are described now.

One of the first studies to investigate speaker roles in broadcast data was presented in [20]. This work considers the use of speaker role information for inferring the structural summary of broadcast news (BN). The news recording were manually segmented into speaker boundaries and each segment is 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 [21]. 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 [22], [23], [24] have also considered the BN roles on broadcast conversations (BC), such as talkshows. In [22], 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 the observation that current role of a speaker is correlated with its 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 [23], 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 [24]. 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 [25], speakers role is 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 [26], 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 [25]. Instead of constructing speaker-speaker networks, affiliation networks were constructed based on temporal proximity of speakers. This method reached an accuracy of 86% for labeling six speaker roles in radio shows.

C. Formal Role Recognition in Meetings

Several previous studies have explored recognition of formal roles in BN and meeting environments; however, 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 [12] 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 [25], while the meeting roles reflect the position of speakers in an organization. Four categories of formal roles were 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 reaches an accuracy of 86% for recognizing BN roles, while the accuracy is 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 [27], 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 [13] 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 [28]. The dataset and roles used in this study are same as described in [12]. Their analysis revealed that combination of verbal and nonverbal features significantly improves 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, the type of 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.

D. Emergent Social Role Recognition in Meetings

For the studies mentioned above, participants role is formal and considered to remain constant over the duration of entire audio recording. The formal roles are generated due to normative expectations of behavior or from positions in an organizational system. Informal social roles, as discussed in [3], [18] 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 [14]. 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 [16], SVM, HMM, and influence model approaches were compared for 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 [29], [15], [30], [31] have also investigated role recognition in professional meetings using the same social role coding scheme proposed in [14]. An HMM based approach was used to model turn statistics and prosody (fundamental frequency, energy) for role recognition in [15]. 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 [29], 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, the feature set explored in this study was limited and a more extensive set of features might be more informative for role recognition task. In [32], the authors compared the performance of multiple feature groups for recognition of both formal and emergent social roles. This study revealed that, while feature combination improved the performance of social roles, lexical features alone were best predictors of formal roles in meetings.

The influence of social roles on language style and vocal expression was investigated in [30]. Using an SVM classifier, an accuracy of 69% was reported for social role classification. In [31], turn taking sequences were modeled using conditional random fields (CRFs) with a reported accuracy of 70% for social role recognition.

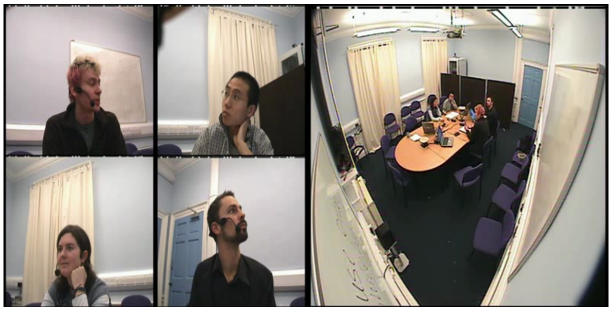

Fig. 1. A snapshot of meeting showing four speaker specific closeup cameras and an overview camera.

Our present work substantially extends previous studies in several ways. We present a detailed study of various features that characterize speaker behavior and describe their importance for individual social roles. In comparison to [32], [31], [30], we analyze the annotation of perceived social roles by human raters and consider its effect on automatically learned role recognition model. Furthermore, unlike [16], [15], [32], [31], [30] we evaluate role recognition model on out of domain data and evaluate its generalized performance on several interaction scenarios.

III. Corpus Description

For the task of annotating social roles, we selected data from AMI meeting corpus [33]. AMI Corpus is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team composed of *Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID)* 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 emergent 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 emergent roles we follow the role coding scheme proposed in [14]. 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 [15] 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 [34]. 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.

A. 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 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.

B. 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 [35], have favored the

TABLE I
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

use of κ statistic to correct for chance agreement between annotators. This idea is expressed in the following equation:

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

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 [36] 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 [14]. This analysis suggests that online raters are fairly consistent in labeling social roles from the point of view of HIT design.

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 [14]. 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 [14] 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.

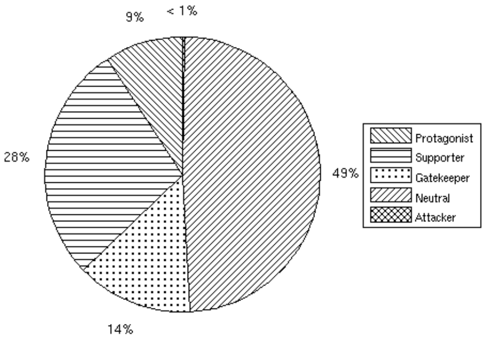

Fig. 2. Overall distribution of individual social roles in the annotated data. The role label for each instance was obtained by majority voting.

TABLE II
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 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 I, 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 I shows the overall distribution of social roles, we also investigated the various group configurations in which the roles appear in meetings. Table II 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 partici-

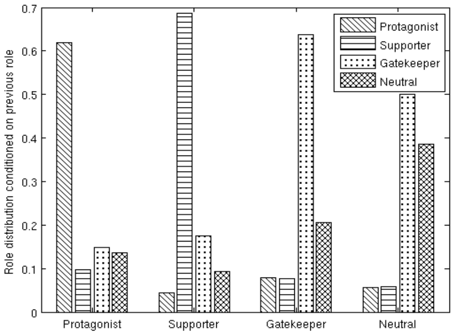

Fig. 3. Social role distribution in current meeting slice conditioned on participants social role in the previous meeting slice. The vertical axis shows the role transition probability across adjacent meeting slices.

pants. 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, 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 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.

IV. Feature extraction

Motivated from previous research in automatic role recognition (described in Section II), 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 [37]. In this work, all the speech transcripts were generated using output of AMI-ASR system [38], which has a word error rate of nearly 30%.

A. Short term features

Turn taking is a basic form of organization for conversations in small group interaction [6]. 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 [39]. 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 [40] 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 $\mathbf{X}_{\mathfrak{S}} = [\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_N]$.

B. 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 have linked language style with use of simple functional words - pronouns, prepositions, articles and other emotional categories.

TABLE III
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)

TABLE IV
SET OF FUNCTIONALS USED TO OBTAIN ACOUSTIC FEATURES VECTORS. THE FUNCTIONALS WERE APPLIED TO CONTOURS GENERATED FROM LLD DESCRIPTORS IN TABLE III AND THE IMPLEMENTATION IS BASED ON THE SYSTEM PRESENTED IN [47]

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

Language style has been used to analyze personality traits [41]. 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 [42] and group cohesion [5].

Linguistic Inquiry and Word Count (LIWC) is a psychologically validated state-of-art text analysis program that quantifies the language style used by participants in a conversation [43]. LIWC operates by counting the fraction of spoken words that fall into predefined categories, such as function words (pronouns, articles or auxiliary verbs) and psychological (emotion, social words, cognitive mechanism) processes. 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 [30].

(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 [30]. 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 [44] and recognizing conflicts [45] 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 III 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 [30]. 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. These features are informative for recognizing personality characteristics like openness and conscientiousness [46]. Statistical and regression functionals defined in Table IV 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, irrespectively of the duration they are speaking. In this work, all the acoustic features were extracted from open-source feature extractor openSMILE [47].

(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.

V. 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 $\sim$ 2 seconds), while long term linguistic, acoustic and structural features are computed over the length of an entire meeting slice (average duration $\sim$ 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) [48], [49]. 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.

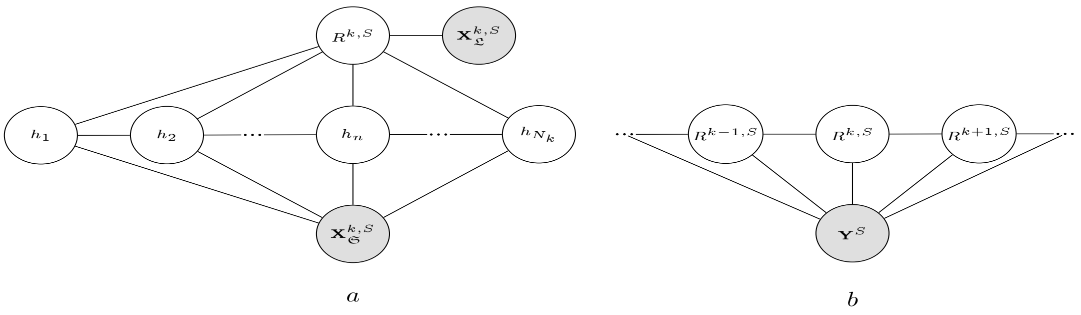

Fig. 4. Graphical representation of CRFs for social role recognition. (a) Modeling influence of roles on short term and long term observations (b) Modeling sequential dependencies between roles. An open node represents a random variable and the shaded node is set to its observed value.

The training data in the corpus is defined for a set of speakers $\mathcal{S}$, who assume social roles in set $\mathcal{R}$, and participate in a set of meetings $\mathcal{M}$. We define $\mathcal{S}_M = \{S_1^M, .., S_M^M\}$ as the subset of speakers appearing in meeting M. During the annotation process, M is partitioned into a sequence of slices for which social roles are labeled. The variable $k \in \mathcal{K}_M = \{1, ..K_M\}$ is used to index the meeting slices. For any speaker $S \in \mathcal{S}_M$, we represent using $R^{k,S} \in \mathcal{R}$ 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 $k_1, k_2 \in \mathcal{K}_M$. 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 IV-A). 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 $\mathcal{X}$ to the label space $\mathcal{R}$. In this work, the conditional distribution $P(R^{k,S}|\mathbf{X}^{k,S})$ factorizes according to an undirected graphical model. Figure 4a 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:

$$P(R^{k,S}|\mathbf{X}^{k,S}) = \frac{\Psi(R^{k,S}, \mathbf{X}^{k,S})}{Z(\mathbf{X}^{k,S})}, \text{where} \tag{2}$$

$$\Psi(R^{k,S}, \mathbf{X}^{k,S}) = \Psi_{\mathfrak{S}}(R^{k,S}, \mathbf{X}_{\mathfrak{S}}^{k,S})\Psi_{\mathfrak{L}}(R^{k,S}, \mathbf{X}_{\mathfrak{L}}^{k,S}) \tag{3}$$

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:

$$\Psi_{\mathfrak{S}}(R^{k,S}, \mathbf{X}_{\mathfrak{S}}^{k,S}) = \sum_{\mathbf{h}} \text{exp}(\sum_i \alpha_i f_i(R^{k,S}, \mathbf{h}, \mathbf{X}_{\mathfrak{S}}^{k,S})) \tag{4}$$

$$\Psi_{\mathfrak{L}}(R^{k,S}, \mathbf{X}_{\mathfrak{L}}^{k,S}) = \text{exp } (\sum_i \beta_i g_i(R^{k,S}, \mathbf{X}_{\mathfrak{L}}^{k,S})) \tag{5}$$

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 $\mathbf{X}_{\mathfrak{L}}^{k,S}$ 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 $f_i(\mathbf{h}, \mathbf{X}_{\mathfrak{S}}^{k,S})$ 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:

$$L(\Lambda) = \sum_{M=1}^{|\mathcal{M}|} \sum_{\forall S \in \mathcal{S}_M} \sum_{k=1}^{K_M} \text{log } P(R^{k,S}|\mathbf{X}^{k,S}; \Lambda) \tag{6}$$

The objective function can be maximized using an iterative algorithm like stochastic gradient ascent or quasi Newton method like L-BFGS [50]. 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 [48].

The role distribution in (2) can be extended to incorporate the continuity in role taking behavior of meeting participants. Figure 3 shows that distribution of social roles in the present slice are influenced by the speakers role in the previous slice. Using (2), we define a posterior feature vector $\mathbf{Y}^{k,S} = \{P(R^{k,S}|\mathbf{X}^{k,S}),\ \forall R^{k,S} \in \mathcal{R}\}$ for every slice k and speaker S. We note that $\mathbf{Y}^{k,S}$ can be efficiently computed using (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 4b, is applied to estimate the conditional probability of the role sequence.

$$P(\mathbf{R}^S|\mathbf{Y}^S) \propto \prod_k \Phi_k(R^{k,S}, R^{k-1,S}, \mathbf{Y}^{k,S}) \tag{7}$$

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.

$$\Phi_k(R^{k,S}, R^{k-1,S}, \mathbf{Y}^S) = \text{exp}(\sum_j \theta_j \gamma_j(R^{k,S}, R^{k-1,S}, \mathbf{Y}^S)) \tag{8}$$

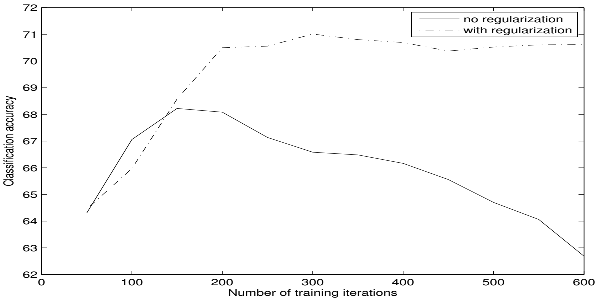

Fig. 5. Comparison in performance of proposed system when the models are trained with and without adding a regularization term.

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

$$L(\Theta) = \sum_{M=1}^{|\mathcal{M}|} \sum_{S=S_1^M}^{S_M^M} \log P(\mathbf{R}^S | \mathbf{Y}^S; \Theta) \tag{9}$$

Since the graphical models in Figure 4 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.

VI. 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 I). Furthermore, a few meeting slices where majority voting resulted in participant having an attacker role label were not considered (see Figure 2). The performance was measured in terms of overall role recognition accuracy and F-measure/Precision/Recall for individual roles.

A. 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 (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 5 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 data. In comparison, the model trained with regularization converges to a higher classification accuracy showing that overfitting is avoided.

B. 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 V shows the different long term feature groups, the number of features within each group and their classification accuracy. The last column in Table V 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.

TABLE V
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

TABLE VI
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

We next explored the effect of different feature combinations on role recognition performance. Table VI illustrates the impact of combining each long term feature group with structural features. The last column in Table VI 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.

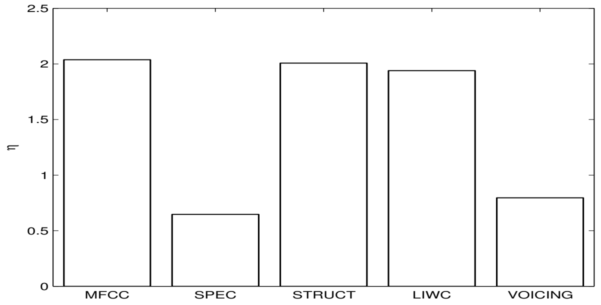

Fig. 6. Comparison of different long term feature groups after feature selection is applied. η measures the relative importance of each feature group.$\eta > 1$ reveals that distribution of selected features from a group is higher after feature selection is applied compared to their initial distribution.

A feature selection algorithm [51], 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 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 V. 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 6 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.

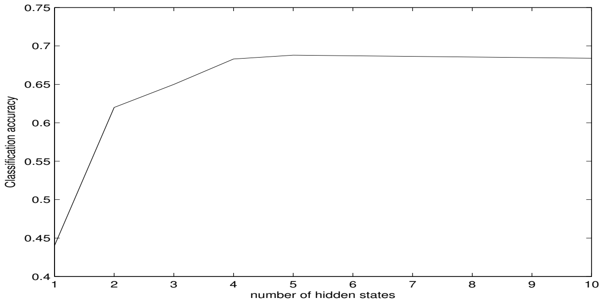

Fig. 7. Variation in social role recognition accuracy as the number of hidden states is increased in the model.

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 7. 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.

C. Analysis of classifications results

The baseline system for comparison is based on the method presented in [14]. 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 [14], 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 [52].

In Table VII we compare the performance of baseline classifier with the proposed system. Furthermore, Table VII 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 [31] 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

TABLE VII
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*

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 VII is same and the observed differences are due to random events. We applied Friedman test [53], which is a non parametric method that ranks the performance of each of the classifiers on all cross-validation 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 VII 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.

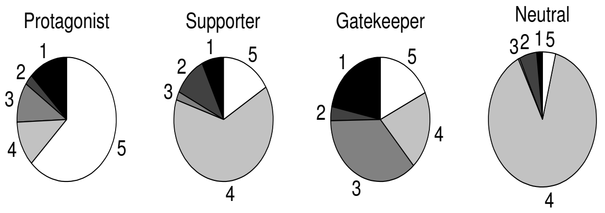

Fig. 8. Distribution of hidden states learned by the model for each social role category.

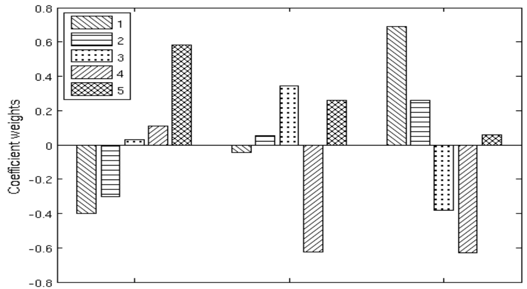

Fig. 9. Parameter weights α_i corresponding to short term feature functions f_i. The feature functions f_i represent turn taking phenomena, like, floor grabbing, turn duration and floor keeping exhibited by speakers.

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 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 8 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 9 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 [15], 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.

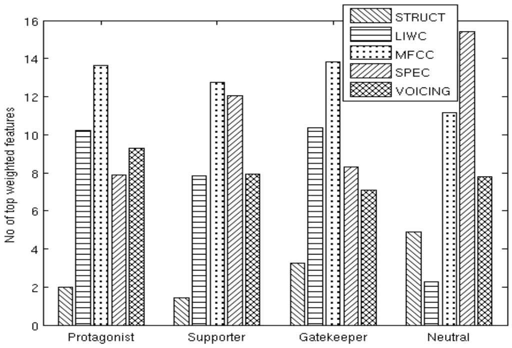

Fig. 10. Distribution of long term feature groups with largest parameter weights β_i used in predicting each social role.

The relation between social roles and long term features is shown in Figure 10. 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 [5].

D. Influence of rater agreement

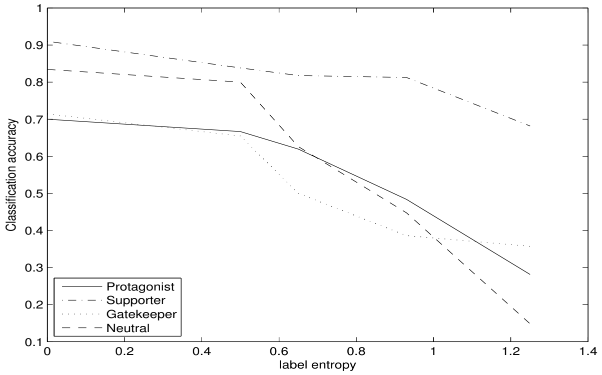

Fig. 11. Accuracy in recognizing individual role labels as a function of label entropy.

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 high, while high label entropy instances indicate substantial disagreement between raters. Figure 11 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.

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 12 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 suggets that proposed classification method is robust against the effect of label noise.

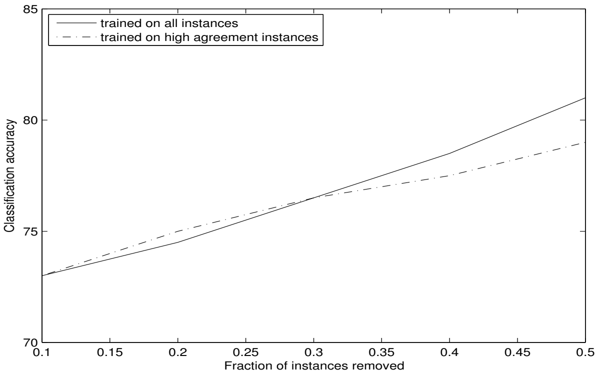

Fig. 12. Comparison in performance of proposed models when trained on all labeled instances and instances with lower label entropy. In both cases the models are evaluated on low entropy labels.

E. Evaluation on AMI 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 III. 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%).

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 13, we plot the average conversation floor entropy for various topics in natural meetings. The AMI

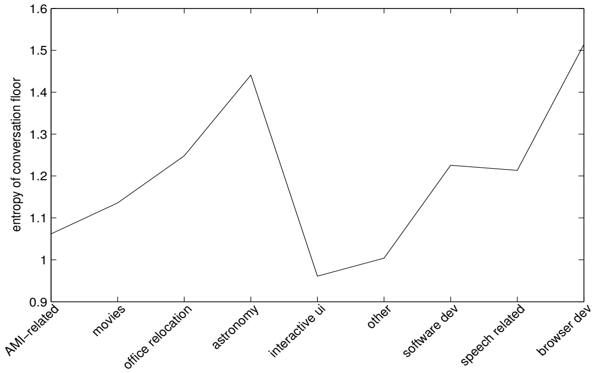

Fig. 13. Average conversation floor entropy for various scenarios in natural meetings.

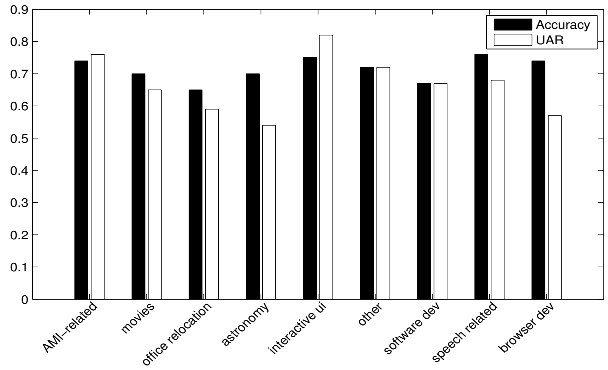

Fig. 14. Role recognition accuracy and UAR for various scenarios in natural meetings.

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 14. 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 protagonists and gatekeepers. Furthermore, these topics also have higher than average conversation floor entropy (see Figure 13). This suggests that active speakers in these meetings do not exhibit the dominant characteristics associated with these social roles.

VII. Conclusion

In this work, we presented an approach for automatic recognition of social roles that emerge in small group meetings. 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 annotated meetings. We investigated various short term and long term features for recognition of social roles. The short term features are computed over short time windows and represent the influence of social roles on turn taking patterns. The long term features are computed over an entire meeting slice and capture the linguistic style and vocal expression of speakers. The proposed role recognition system is modeled by extending the framework of CRFs and integrates feature information at multiple time scales.

Our investigations revealed that automatically extracted speaker interaction features and long term features are useful cues for predicting social roles. The model trained with 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. Experiment results also reveal that the accuracy of proposed approach (74%) is significantly better than accuracy of SVM (70%). This suggests that combining feature information at multiple time scales in a single model increases the predictive capabilities of the automatic recognition system. We evaluated the generalized performance of the proposed approach on various scenarios of multiparty interaction. Experiments show that proposed model reaches a recognition accuracy of 72% on out of domain data, which is slightly lower than in domain accuracy of 74%. Although, further research on other corpora are needed to reach definite conclusions, our results suggest that the proposed approach is able to model the influence of social roles on behavioral patterns of speakers in small group interaction.

References

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

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

[3] R. F. Bales, "A Set of Categories for the Analysis of Small Group Interaction," *American Sociological Review*, vol. 15, no. 2, 1950.

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

[5] 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.

[6] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, Part 1, pp. 696–735, December 1974.

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

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

[9] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vision Comput.*, vol. 27, no. 12, pp. 1775–1787, Nov. 2009.

[10] 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*, New York, NY, USA, 2007, MULTIMEDIA ’07, pp. 261–264, ACM.

[11] A. Vinciarelli, “Sociometry based multiparty audio recordings summarization.,” in *ICPR (2)*. 2006, pp. 1154–1157, IEEE Computer Society.

[12] 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.

[13] 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.

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

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

[16] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, “Modeling functional roles dynamics in small group interactions.,” *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 83–95, 2013.

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

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

[19] P. E. Slater, “Role Differentiation in Small Groups,” *American Sociological Review*, vol. 20, no. 3, 1955.

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

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

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

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

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

[25] H. Salamin, S. Favre, and A. Vinciarelli, “Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction.,” *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1373–1380, 2009.

[26] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, “Rolenet: Movie analysis from the perspective of social networks,” *Multimedia, IEEE Transactions on*, vol. 11, no. 2, pp. 256–271, Feb. 2009.

[27] 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.

[28] 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.

[29] T. Wilson and G. Hofer, “Using linguistic and vocal expressiveness in social role recognition.,” in *IUI*, Pearl Pu, Michael J. Pazzani, Elisabeth Andr, and Doug Riecken, Eds. 2011, pp. 419–422, ACM.

[30] 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*, Sept. 2013, p. 6.

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

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

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

[34] N. Ambady and R. Rosenthal, “Thin Slices of Expressive behavior as Predictors of Interpersonal Consequences : a Meta-Analysis ,” *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[35] J. Carletta, “Assessing agreement on classification tasks: The kappa statistic,” *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, June 1996.

[36] J.L Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[37] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, “The ICSI Meeting Corpus,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03). 2003 IEEE International Conference on*, 2003, vol. 1, pp. I–364–I–367 vol.1.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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*, 2010, MM ’10.

[48] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, “Hidden conditional random fields for phone classification.,” in *INTERSPEECH*. 2005, pp. 1117–1120, ISCA.

[49] A. Quattoni, M. Collins, and T. Darrell, “Conditional random fields for object recognition.,” in *NIPS*, 2004.

[50] D. C. Liu and J. Nocedal, “On the limited memory BFGS method for large scale optimization,” *Math. Programming*, vol. 45, no. 3, (Ser. B), pp. 503–528, 1989.

[51] 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*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

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

[53] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Ashtosh Sapru Ashtosh Sapru received the BSc. degree in electrical engineering from NIT Jamshedpur, India, in 2004 and the ME. degree in signal processing from Indian Institute of Science, Bangalore in 2006. From 2006 to 2010 he worked as a senior engineer at Aricent India. Since 2011, he is a research assistant at the Idiap Research Institute, Switzerland and a doctoral student at the the École polytechnique fédérale de Lausanne (EPFL), Switzerland. His research interests include human behavior modeling, applied machine learning and signal processing.

Hervé Bourlard Hervé Bourlard received the Electrical and Computer Science Engineering degree and the Ph.D. degree in applied sciences both from Faculté Polytechnique de Mons, Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the École polytechnique fédérale de Lausanne EPFL, and (Founding) Director of a Swiss NSF National Centre of Competence in Research on Interactive Multimodal Information Management. Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now a member of the ICSI Board of Trustees.

His research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modelling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 6 books and over 300 reviewed papers (including one IEEE paper award) and book chapters. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of Speech Communication). He is the recipient of several scientific and entrepreneurship awards.