# AUTOMATIC SPEAKER ROLE LABELING IN AMI MEETINGS: RECOGNITION OF FORMAL AND SOCIAL ROLES

Ashtosh Sapru[1,2], Fabio Valente[1]

[1] Idiap Research Institute, 1920, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
*ashtosh.sapru@idiap.ch,fabio.valente@idiap.ch*

## ABSTRACT

This work aims at investigating the automatic recognition of speaker role in meeting conversations from the AMI corpus. Two types of roles are considered: formal roles, fixed over the meeting duration and recognized at recording level, and social roles related to the way participants interact between themselves, recognized at speaker turn level. Various structural, lexical and prosodic features as well as Dialog Act tags are exhaustively investigated and combined for this purpose. Results reveal an accuracy of 74% in recognizing the speakers formal roles and an accuracy of 66% (percentage of time) in correctly labeling the social roles. Feature analysis reveals that lexical features provide the higher performances in formal/functional role recognition while prosodic features provide the higher performances in social role recognition. Furthermore results reveal that social role recognition in case of rare roles in the corpus can be improved through the use of lexical and Dialog Act information combined over short time windows.

***Index Terms***— Speaker Role Labeling, AMI Meetings, Formal and Social Roles, Structural, Lexical and Prosodic feature analysis.

## 1. INTRODUCTION

Automatic labeling of speaker roles has been widely studied in case of Broadcast News (BN) recordings finding applications into indexing, summarization and retrieval. Typical roles considered in BN audio are formal roles (also referred as functional roles), i.e., roles imposed from the news format and related to the task each speaker performs in the show like anchorman, journalists, interviewees or soundbites. Common features used to train statistical classifiers consist of lexical features [1] as well as structural features from the recording, prosodic features and Dialog Acts [2, 3, 4]. More recently, automatic role labeling has also been studied in spontaneous conversations including Broadcast Conversations (BC) [3, 5, 6] as well as meeting recordings [7, 8]. Formal roles studied in conversations, change depending on the data settings, for instance anchorman/guests in BC talk shows [3, 6, 5], the Project Manager in the AMI corpus [7] or the faculty members in the ICSI corpus [8].

Beside formal roles, other coding schemes have been proposed in literature with the purpose of generalizing across any type of conversations and settings, for instance, the Socio-Emotional roles [9]. Social roles are inspired from Bales work [10] and characterize the relationships between group members and their roles "oriented towards the functioning of the group as a group". This coding scheme attributes to each participant in the discussion a role in between Protagonist/Supporter/Neutral/Gatekeeper or Attacker at each time instant. Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants, they can generalize across any type of conversation and can be related to phenomena studied in meetings like engagement, hot-spots [11] and also social dominance. Previous works on social roles have mainly used non-lexical features [12, 13, 14] focusing on how participants interact over long time windows (up to one minute) in the conversation, with the exception of [15] where information capturing speaker expressiveness, derived from lexical and prosodic features, was used over short time windows.

Both formal and social roles find applications into analysis, indexing, summarization and question answering, however, their automatic recognition has been addressed using completely different approaches. The literature on the first has mainly focused on the use of lexical and structural features [3, 5, 6, 7] while the literature on the second has mainly made use of non-lexical information [12, 13, 14] (prosody and turn-taking statistics). This work aims at studying in exhaustive way and on the same dataset which of the various features proposed in literature are able to capture information on speaker formal/social roles. For this purpose the AMI corpus scenario meetings [16] are used where, the scenario imposes some constraints to the conversations, thus defining, a set of formal roles that each speaker takes in the discussion while participants spontaneously interact taking in turn different social roles over time. Features that will be studied consists of structural and conversational features, prosodic features, lexical features and Dialog Acts. In the remainder of the paper, Section 2 describes the data and the role annotations, section 3 describes the methods, the various features and results on both social and formal roles. The paper is then concluded in section 4

## 2. DATA AND ANNOTATION

The AMI Meeting 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. These roles will be referred as formal roles. The meeting is supervised by the Project Manager who follows an agenda with a number of items to be discussed with other speakers. The corpus is manually transcribed at different levels (roles, speaking time, words, dialog act). Formal roles do not change during the meeting.

Social role annotation are obtained with the same guidelines as [9] (CHIL project) where annotators where provided with audio and video and could assign a mapping speaker-to-role at any time instant. The guidelines define a set of acts and behaviors that character-

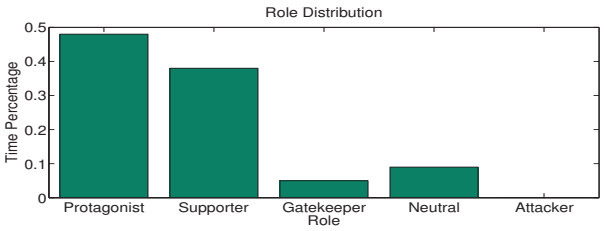


**Fig. 1**. Social role distribution in the five meetings in terms of time for which a speaker holds the conversation floor. In contrary to [14], statistics are obtained from manual annotations without applying any temporal smoothing on the social role annotations nor assuming temporal continuity.

ize each social role and is 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 shows a cooperative attitude demonstrating attention and acceptance providing technical and relational support; **Neutral** - a speaker that passively accepts others ideas; **Gatekeeper** - a speaker that acts like group moderator, mediates and encourage the communication; **Attacker** - a speaker who deflates the status of others, express disapproval and attacks other speakers. Accurate annotations in terms of social roles were manually obtained for five scenario meetings (ES2002d, ES2008b, ES2008d, ES2009d, IS1003d) for a total of 20 different speakers and 3 hours of recordings.

The coding scheme assumes that the same speaker can change social role over time (even during the same turn) but at each time instant, a speaker has a single social role. The amount of speaker's actual speaking time labeled according to the different roles is depicted in Figure 1 where it can be notice that the Protagonist/Supporter roles are the most common one, the Neutral and Gatekeeper are rare roles in those meetings while no instances of Attacker are found because of the collaborative nature of the meetings. As speaker roles continuously change over time, previous works [12, 13, 14] smoothed the manual annotations considering roles constant over long time-windows (up to one-minute) and assigning speakers the most common role they have in that window thus leveraging a role over several turns. Furthermore these works completely discarded information like words and Dialog Acts considering only prosodic and structural features from the conversation. All those assumptions will be relaxed here and the work will investigate how those roles can be recognized using shorter time windows like speaker turn.

## 3. METHOD AND EXPERIMENTS

In order to combine structural, prosodic, lexical and Dialog Act information in discriminative fashion, this work makes use of boosting algorithms. 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 version of Boosting algorithm used was multi-class Boosting defined in [17] and implemented using Boostexter. The weak learners are one-level decision trees. This algorithm provides a very simple and effective way to combine continuous features as well as discrete features.

### 3.1. Formal Role Recognition

The scenario meetings (138 recordings) from the AMI corpus are used for studying formal role recognition; the test set consists of 20 meetings while the remaining are used for training/development. Audio manually segmented from the Independent Headset Microphones (IHM) is force-aligned for obtaining precise speech/non-speech segmentation. This segmentation is used to extract a sequence of speaker turns; although several definitions of speaker turns have been given in literature, we consider here the definition used by [8, 18], i.e., speech regions from a single speaker uninterrupted by pauses longer then 300 ms. Each turn is then associated with one of the four formal roles (PM,ME,UI,ID). Based on this turn segmentation the following features, to be used by the booster, are extracted for each speaker:

**Structural features** : the total speech time and the total number of turns in the meeting per speaker as well as the speaker centrality, estimated as in [3, 7] based on Social Network Analysis, computed as the incoming, outgoing, and total number of links to nodes where nodes represent speakers, and the incoming and outgoing links are established by turn-taking patterns. Those features are used into the booster as continuous features.

**Prosodic features** : the fundamental frequency (F0) is computed from the headset microphones using 30ms long windows shifted by 10ms. After that speaker turn statistics like mean, maximum, minimum, median and the standard deviation are computed. A histogram-based speaker normalization is applied before discretizing each of them into 16 bins of equal area under the normal distribution. The discretized F0 statistics have already been successfully used for recognition Broadcast News roles in [19]. Beside those also speech rate over the turn and mean frame-level RMS are extracted over speaker turns, normalized and binned. The per-speaker discretized features are then used into the booster.

**Lexical Features** : N-gram of words (from force-aligned manual transcriptions) are used into the booster in order to capture role lexical information under the rationale that different roles will make use of different vocabulary. Word bi-grams are here used.

**Dialog Act Tags** : Dialog Acts (DA) aim at capturing the speaker's intention in the discussion. AMI corpus is annotated in terms of 14 broad 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, Elicite-Comment*) and also social acts (*Be-positive, Be-negative*). As potentially correlated with several phenomena in conversations (see [11]), the per-speaker DA counts (14 per each speaker) are included in the booster. Table 1 reports the performance of the various structural, prosodic, lexical features as well as the dialog act tags in recognizing the formal roles. As there are four speakers mapping into four roles for each of the recordings and speakers do not change formal role during the meeting, performances are reported in terms of role accuracy. It can be noticed that structural, prosodic features and DA are able to detect the meeting Project Manager with an accuracy of 70% while they often provide quite low performance on other roles. On the other hand, lexical features provide the highest performances on

**Table 1**. Per-role and total accuracy obtained using structural, prosodic, lexical features and Dialog Act tags for formal role recognition.

| | Per-role Accuracy | | | | Total |
|---|---|---|---|---|---|
| Features | PM | ME | UI | ID | accuracy |
| Structural/Convers. | 0.7 | 0.2 | 0.5 | 0.4 | 0.45 |
| Prosodic | 0.7 | 0.2 | 0.1 | 0.35 | 0.34 |
| Lexical | **0.9** | **0.75** | **0.8** | **0.6** | **0.77** |
| Dialog Acts | 0.7 | 0.4 | 0.35 | 0.45 | 0.48 |
| ALL | 1 | 0.7 | 0.65 | 0.6 | 0.74 |

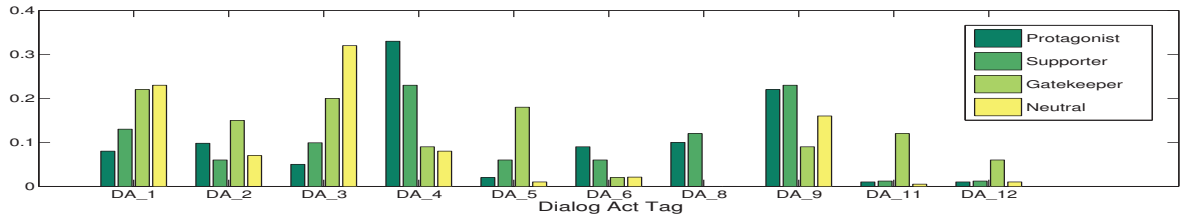


**Fig. 2**. Normalized DA tag distribution on data annotated for social roles for the most common DA tags. DA_1 (Backchannel), DA_2 (Stall), DA_3 (Fragment), DA_4 (Inform), DA_5 (Elicit-Inform), DA_6 (Suggest), DA_8 (Elicit-Offer), DA_9 (Assess), DA_10 (Elicit-Assessment), DA_11 (Comment).

all the formal roles achieving a total accuracy of 77% (90% for the PM). Word N-grams that are more discriminant between roles are those related to the agenda and its items discussed in the meeting. Whenever the various features are combined together, all Project Manager instances are recognized correctly while a degradation in recognizing the other roles in verified. This suggests, as already shown in [7], that the formal roles in the AMI corpus are mainly captured by lexical information.

### 3.2. Social Role Recognition

A set of five annotated meetings (3 hours of meeting data) comprised our setup for evaluating social roles. Cross validation is performed where the complete data is randomly partitioned into five disjoint sets of training and test data. Speakers do change their social role during the meeting and, frequently they change it also during a single turn. Previous works have investigated the use of long time-windows and mainly non-lexical features. As first investigation, let us consider the recognition of social role at *turn level*. The combination is based on the same boosting technique previously described although features and statistics are extracted from a single turn. The feature set consists of : **turn-based structural features**, i.e., the duration of the current turn as well as the durations of the previous and following turns, and the relative position of the turn in the meeting. **Prosodic features** are extracted and normalized similarly to what previously described at turn level and also **Lexical features/Dialog acts** from each speaker turn are included in the booster.

During training, the social role of a speaker is considered as the most common role the speaker has over the duration of the turn. In comparison to AMI formal roles, social roles are not equally distributed across recordings (see Figure 1), thus performances are reported in terms of F-measure/Precision/Recall. During testing, a social role in between Protagonist/Supporter/Gatekeeper/Neutral is assigned to each turn and the F-measure/Precision/Recall in term of correctly labeled time are then computed.

Table 2 reports the performance of various structural, prosodic, lexical features as well as the Dialog Act tags in recognizing the speakers social role on each turn as well as the total accuracy in terms of correctly labeled time. It can be noticed that the highest total accuracy (62%) is achieved by the prosodic features which outperform all other features in recognizing the Protagonist, Supporter and Neutral roles. Lexical features have the highest performance in terms of Gatekeeper recognition. The worst performance is obtained using the Dialog Act tags. Their combination achieves a total accuracy of 64% with an F-measure of 0.72 and 0.62 for the two most common roles (Protagonist and Supporter) and F-measure of 0.32 and 0.42 for the two rare roles (Gatekeeper and Neutral). The per-role F-measures reveal that, whenever role recognition is done at turn level, the only significant improvement over prosodic features comes for the Gatekeeper role.

A feature wise analysis shows that the turns with higher durations are assigned more often to Protagonist role, whereas turns with small durations are assigned to Supporters/Neutral. For the prosodic features, boosting gives higher weights to F0 standard deviation, slope and intensity of the speech as discriminant between Protagonists, Supporters and Neutrals. The DA tag performance (see Table 1) is quite low compared to other features especially for the Protagonist and Gatekeeper roles. Figure 2 plots the per-role normalized distribution of the most common Dialog acts tags. It is possible to notice to notice that some DA correspond to particular roles more frequently then others, e.g., backchannels (DA_3) occurs more frequently in case of Neutral speakers, DA_4 (Inform) occurs more frequently in case of Protagonist speakers while DA_5 and DA_11( Elicit-Inform and Comment) occur more frequently in case of Gatekeepers. Further analysis shows that, especially during long turns, the speakers social role changes, e.g., a speaker can be labeled as Gatekeeper at the begin of the turn (e.g., while introducing the agenda) for becoming Protagonist (e.g., while starting the discussion on the agenda item) at the end of the same turn. Counting different DA tags over long turns produces poor performances especially in case of Protagonist which is thus confused with Supporter/Gatekeper roles.

To verify this hypothesis, the same recognition experiments are repeated trying to label social role using as recognition unit the DA start and end time. In other words, instead of considering the role constant over each turn, the role is considered constant over each DA duration. The various structural,prosodic and lexical features are extracted in the DA start and end time boundaries while the booster sees a single DA tag. As consequence, long speaker turn are broken in smaller units. Performances are reported in Table 3. Structural, prosodic and lexical features hold their overall performances while the DA performances increase from 39% (turn based classification) to 51%. The total accuracy achieves 66% with gain in recognizing the rare roles (Gatekeeper and Neutral speakers). In summary, while prosodic features hold the highest performances in recognizing social roles, improvements in labeling rare roles can be obtained combining also lexical and DA information when this combination happens over short time units.

## 4. DISCUSSIONS AND CONCLUSIONS

Speaker role recognition in conversations has been an active research field in last years. Two types of roles have been investigated in different literatures: formal/functional roles, defined from the conversation type, e.g., anchorman/guests in Broadcast data [3, 6, 5] or professional roles in meetings [7, 8], and social roles [12, 13, 14, 15] related to how speakers interact between them. The literature on the first has mainly focused on the use of lexical and structural features while the literature on the second has mainly made use of non-lexical information (prosody and turn-taking statistics). This work

**Table 2**. Per role F-measure, Precision and Recalls (percentage of time) obtained in recognizing social roles using speaker turn as recognition unit. Also the total amount of correctly labeled speaker time is reported.

| | Per-role F-measure (Precision/Recall) | | | | Accuracy per |
|---|---|---|---|---|---|
| Features | Protagonist | Supporter | Gatekeeper | Neutral | labeled time |
| Structural | 0.61 (0.58/0.64) | 0.55 (0.69/0.46) | 0.00 (0/0) | 0.02 (0.01/0.27) | 0.54 |
| Prosodic | **0.71 (0.70/0.72)** | **0.61 (0.69/0.55)** | 0.12 (0.08/0.23) | **0.38 (0.29/0.53)** | **0.62** |
| Lexical | 0.62 (0.58/0.68) | 0.55 (0.69/0.46) | **0.21 (0.14/0.42)** | 0.06 (0.03/0.22) | 0.55 |
| Dialog Acts | 0.23 (0.14/0.61) | 0.52 (0.86/0.38) | 0.12 (0.09/0.19) | 0.12 (0.09/0.20) | 0.39 |
| ALL | 0.72 (0.71/0.73) | 0.62 (0.69/0.56) | 0.32 (0.22/0.56) | 0.42 (0.33/0.57) | 0.64 |

**Table 3**. Per role F-measure, Precision and Recalls (percentage of time) obtained in recognizing social roles using DA start/end times as recognition unit. Also the total amount of correctly labeled speaker time is reported.

| | Per-role F-measure (Precision/Recall) | | | | Accuracy per |
|---|---|---|---|---|---|
| Features | Protagonist | Supporter | Gatekeeper | Neutral | labeled time |
| Structural | 0.68 (0.81/0.59) | 0.46 (0.43/0.50) | 0 (0/0) | 0 (0/0) | 0.55 |
| Prosodic | **0.72 (0.76/0.68)** | **0.58 (0.59/0.56)** | **0.24 (0.16/0.48)** | **0.43 (0.34/0.60)** | **0.62** |
| Lexical | 0.65 (0.69/0.62) | 0.48 (0.48/0.53) | 0.12 (0.12/0.07) | 0.07 (0.04/0.36) | 0.54 |
| Dialog Acts | 0.63 (0.64/0.61) | 0.47 (0.52/0.43) | 0.11 (0.07/0.23) | 0.14 (0.09/0.35) | 0.51 |
| ALL | 0.74 (0.78/0.71) | 0.62 (0.62/0.62) | 0.37 (0.27/0.57) | 0.46 (0.38/0.60) | 0.66 |

**Table 4**. Summary of Social Role recognition results when turn and DA are used as recognition units.

| Recognition | Features | F-measure | | | | Total |
|---|---|---|---|---|---|---|
| Unit | | Pr. | Su | Ga | Nu | Accuracy |
| Turn | Prosody | 0.71 | 0.61 | 0.12 | 0.38 | 0.62 |
| Turn | ALL | 0.72 | **0.62** | 0.32 | 0.42 | 0.64 |
| DA | ALL | **0.74** | **0.62** | **0.37** | **0.46** | **0.66** |

extensively investigates and analyzes, on the same dataset, how the various features perform in the task of labeling those roles. As per authors best knowledge, the only dataset labeled according to both schemes is the AMI corpus, where the scenario imposes constraints on the participant formal roles during a professional meeting while speakers spontaneously interact taking in turn different social roles.

Speaker do not change their formal role during a meeting thus statistics extracted from the entire recording can be used for labeling those roles. Results reveal that lexical features are the one that provide the highest performances on all the four formal roles (77% of correctly labeled roles). Structural, prosodic features and DA tags are able to recognize the meeting chairperson (the Project Manger) but provide considerably lower performances on other roles. Consistently with [7] as well as other studies on broadcast conversations [1, 3], those results reveal that most of the formal role information can be captured by word/lexical information.

On the other other hand, speakers change their social role during the conversation and previous work on social role labeling [12, 14] made use of long time windows where the role is considered constant and obtained averaging over several turns. Results reveal (see the summary Table 4) that prosodic features produce the highest recognition (62% correctly labeled time) while the use of structural, lexical and DA information improves the performance up to 64%. The gain comes from rare roles like the Gatekeeper and the Neutral. As the speaker social role can also change during a turn, this work also investigates whether Dialog acts start/end times are better suited recognition units instead of speakers turns. Results show that the performances over rare roles like the Gatekeeper and the Neutral (see table 4) improve if DA start/end times are used as classification units, achieving a 66% accuracy. In summary, prosodic features are the most informative features for social role recognition (consistently with [12, 14, 15]) while the lexical and DA information can help in recognizing less frequent roles (Gatekeeper and Neutral) whenever combined with the prosodic information[1] [2].

## 5. REFERENCES


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

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

[3] Wang W., Yaman S., Precoda P., and Richey C., "Automatic Identification of Speaker Role and Agreement/Disagreement in B roadcast Conversation.," in *Proceedings of ICASSP*, 2011.

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

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

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

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

[8] Laskowski K. et al., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," *Proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, 2008.

[9] Pianesi et al., "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation, 41 (3)*, 2007.

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

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

[12] Zancaro M. et al., "Automatic detection of group functional roles in face to face interactions," *Proceedings of ICMI*, 2006.

[13] Dong W. et al., "Using the influence model to recognize functional roles in meetings," *Proceedings of ICMI*, 2007.

[14] Valente F. and Vinciarelli A., "Language-independent Socio-Emotional Role Recognition in the AMI corpus," in *Proceedings of Interspeech*, 2011.

[15] Wilson T. et al., "Using linguistic and vocal expressiveness in social role recognition," *Proceedings of the Conference on Intelligent User Interfaces(IUI)*, 2011.

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

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

[18] Shriberg E. et al., "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *in Proceedings of Eurospeech 2001*, 2001, pp. 1359–1362.

[19] Bigot B. et al., "Looking for relevant features for speaker role recognition," *Proceedings of Interspeech*, 2010.



[1] Authors would like to thank the University of Edinburgh for providing the social role annotations.

[2] This work was funded by the Hasler Stiftung under SESAME grant, the EU NoE SSPNet, and the Swiss National Science Fundation NCCR IM2.