Understanding Social Signals in Multi-party Conversations: Automatic Recognition of Socio-Emotional Roles in the AMI Meeting Corpus

Alessandro Vinciarelli
University of Glasgow
Glasgow G12 8QQ
Scotland (UK)
vincia@dcs.gla.ac.uk

Fabio Valente, Sree Harsha Yella, Ashtosh Sapru
Idiap Research Institute
Rue Marconi, 19
CH-1920 Martigny (Switzerland)
{fabio.valente,sree.yella,ashtosh.sapru}@idiap.ch

***Abstract*—Any social interaction is characterized by roles, patterns of behavior recognized as such by the interacting participants and corresponding to shared expectations that people hold about their own behavior as well as the behavior of others. In this respect, social roles are a key aspect of social interaction because they are the basis for making reasonable guesses about human behavior. Recognizing roles is a crucial need towards understanding (possibly in an automatic way) any social exchange, whether this means to identify dominant individuals, detect conflict, assess engagement or spot conversation highlights. This work presents an investigation on language-independent automatic social role recognition in AMI meetings, spontaneous multi-party conversations, based solely on turn organization and prosodic features. At first turn-taking statistics and prosodic features are integrated into a single generative conversation model which achieves an accuracy of 59%. This model is then extended to explicitly account for dependencies (or influence) between speakers achieving an accuracy of 65%. The last contribution consists in investigating the statistical dependency between the formal and the social role that participants have; integrating the information related to the formal role in the recognition model achieves an accuracy of 68%. The paper is concluded highlighting some future directions.**

***Index Terms*—Social signals, AMI meetings Corpus, role recognition, social and formal roles, turn-taking patterns, non-verbal communication.**

I. Introduction

Several decades of research in conversation analysis [1] and role theory [2] have shown that human behavior, especially when it comes to social interaction, is not random, but follows principles and laws stable enough to produce observable effects. This applies in particular to roles, typically defined as "characteristic behavior patterns" [3] that can be identified and recognized as such by interaction participants (see the seminal work in [4]). However, even though roles tend to induce physical, possibly machine detectable behavioral evidences, statistical approaches have been used only recently to model, analyze and automatically extract this type of information from archives of interaction recordings.

Automatic role recognition based on statistical classifiers has been studied in meeting recordings like the CMU corpus [5], the AMI corpus [6], [7] and the ICSI corpus [8] as well as broadcast [9], [10] and telephone [11] conversation corpora. Typical features consist in turn-taking patterns, i.e., the way speakers take turns in the discussion, turns durations, overlaps between participants, stylistic and prosodic features as well as lexical features. However, the roles considered in those studies are typically scenario-specific (e.g. the *Anchorman* in news or the *Project Manager* in meetings) and cannot be used easily for data different from those used in each work for the experiments. Furthermore, the works above consider only data where the role of a person does not change during the entire recording, a significant simplification for automatic approaches.

For this reason, this work focuses on the Socio-Emotional roles [12]. These are inspired from Bales Interaction Process Analysis [4] and are "oriented toward the functioning of the group as a group" [15], independently of any particular scenario or corpus. The related coding scheme attributes to each participant in an interaction a role in between the following:

- *Protagonist*: a speaker that takes the floor, drives the conversation, asserts its authority and assumes a personal perspective.
- *Supporter*: a speaker that shows a cooperative attitude demonstrating attention and acceptance as well as providing technical and relational support;
- *Neutral*: a speaker that passively accepts others ideas without expressing others ideas;
- *Gatekeeper*: a speaker that acts as group moderator, mediates and encourages the communication;
- *Attacker*: a speaker who deflates the status of others, expresses disapproval and attacks other speakers.

It is intuitive that the same speaker can change social role over time but its social role will not change frequently within a short time window and at each time instant, a speaker has a single social role in the conversation.

Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants, and are related to phenomena like engagement in the discussion, hot-spots [13] (segments of engaged speech showing

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

amusement or disagreement) and also social dominance [14]. Previous works on automatic social role recognition have been mainly performed on corpora that study group decision making like the Mission Survival Corpus [12], where SVM classifiers trained on audio and video activity features extracted from a 10 seconds long window are used for this purpose [15]. Later in [16], the use of the influence model, coupled HMMs generatively trained on audio and video activity features, was shown superior to the SVM. In this case, features were extracted from one minute long window during which the role of each speaker is considered constant; each chain of the coupled HMMs represents features from a single speaker. The influence that each speaker has on other participants is modeled trough the chains coupling which can recognize joint activity of multiple speakers. Furthermore, studies like [15],[16],[17] have outlined how social roles appear strongly correlated with non-linguistic cues, typical of social signaling [18].

This work investigates the recognition of social roles in the AMI corpus [19], a collection of professional meetings. Previous studies on those data have mainly addressed the recognition of formal roles [6], [7], [8] like the Project Manager during a project brainstorming session. The paper provides three contributions:

1 at first a language-independent generative model that accounts for turn-taking patterns, turn duration and prosody is proposed towards the recognition of social roles; those features have been mainly considered in literature for the recognition of formal roles - this work investigates their application for recognizing another role coding scheme.
2 the model is modified to account for the "influence" that each role has on others, the rationale being that it could better capture group actions and dependencies between speakers. This is achieved introducing context-dependent role models aiming at capturing joint behavior between speakers.
3 As last contribution, the paper investigates the dependencies between the social and formal roles proposing the use of the formal role as auxiliary information for the social role recognition. The rationale behind this consists in the fact that, in professional meetings, social roles could be partially influenced by the status of each participant.

Let us now describe the data and their annotations.

II. Dataset and Annotations

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 act as members of a team expected to design a new remote control. Each participant plays one (and only one) of the following roles during an entire meeting: Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID). These roles will be referred to as formal roles. The meeting is supervised by the PM. The corpus is manually transcribed at different levels (roles, speaking time, words, dialog act).

Accurate annotations in terms of social roles were manually obtained for five of the scenario meetings above (ES2002d, ES2008b, ES2008d, ES2009d, IS1003d) for a total of 20 different speakers and 3 hours of recordings. In order to compare results with previous studies on other corpora [12], the same annotation guidelines and heuristics used to produce the role annotations in the Mission Survival Corpus [16] were applied. Annotators were provided with audio and video and could assign a mapping speaker-to-role at any time instant. In other words, given a set of participants $\{S\}$ and the role set $\{R\} = \{P, S, N, G, A\}$ (P = protagonist, S = supporter, N = neutral, G = gatekeeper, A = attacker), a mapping $\varphi(S) \rightarrow R$ speaker-to-role is produced for each time instant. Annotators could use the time resolution they preferred to assign the roles - ideally down to the video frame-rate. However it is intuitive that the same speaker can change role over time but roles do not change frequently within a small time window. Manual annotations are then post-processed as described in [16]; at a given time instant t, the role becomes the most frequent role that the speaker has in a one-minute long window centered around time t[1].

The resulting role distribution of the five meetings is depicted in Figure 1 (left): most of the time is attributed to the Protagonist/Supporter/Neutral roles and only 5% of the time is attributed to the Gatekeeper. No speaker is labeled as Attacker because of the collaborative nature of the professional meeting. Furthermore Figure 1 (right) plots the social role distribution conditioned to the formal role that each speaker has in the meeting. The Gatekeeper role, i.e., the moderator of the discussion, is consistently taken by the Program Manager which also take the Neutral role less frequently then other speakers. Participants different from the Program Manager rarely take the Gatekeeper role.

III. Feature extraction

The audio data are processed according to the following steps. The speech activity of the four speakers is obtained force-aligning the manual speech/non-speech segmentation with the system described in [20] to produce very precise speaker boundaries.

This segmentation is used to extract a sequence of speaker turns; although several definition of speaker turns have been given in literature, we consider here the simplified definition provided by [21] and [22], i.e., speech regions from a single speaker uninterrupted by pauses longer then 300 ms. To simplify the problem overlapping speech segments are ignored, i.e., the time in overlapping regions between speakers (including back-channels) is assigned to the speaker that currently holds the floor of the conversation.

[1]This is also the window size typically used for recognizing hot-spots in ICSI meetings corpus.

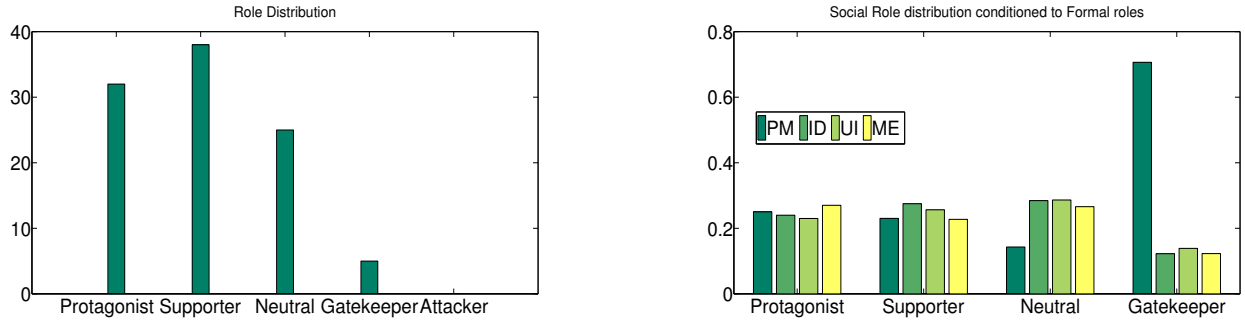

Fig. 1. (Left Plot) Role distribution on the 5 meetings annotated in terms of social roles. (Right Plot) Social role distribution conditioned to the formal role that each speaker has in the meeting.

Furthermore the following measures are extracted from the speech regions that compose each turn: F0 frequency mean, standard deviation, minimum, maximum and median for each turn, mean and standard deviation of energy for each turn and mean speech rate over the turn. Those measures are then concatenated to form a single feature vector of dimension nine which undergoes a speaker level z-normalization as in [13]. The resulting feature vector will be designated in the following as $\{X_t\}$.

In summary, each meeting is transformed into a sequence of speaker turns:

$$M = \{(t_1, d_1, X_1, s_1, r_1, f_1), ..., (t_N, d_N, X_N, s_N, r_N, f_N)\} \quad (1)$$

where:

- N is the total number of speaker turns.
- t_n is the turn start.
- d_n is the turn duration
- X_n is the vector of prosodic features.
- s_n designates the speaker.
- r_n designates its social role in $\{P, S, N, G\}$.
- f_n designates its formal role in $\{PM, UI, ID, ME\}$.

During the training, the social role r_n is known while the inference will consists in recognizing r_n when all the other elements in Eq. 1 are known.

IV. Statistical modeling

Let us statistically model the conversation as a sequence of elements that compose Eq. 1. The most simple model is a first-order Markov chain, represented using the Dynamic Bayesian Network formalism in figure 2 (Model 1) where variables r_n accounts for the social roles. Its probability can be written as:

$$p(M) = \prod_{n=2}^{N} P(X_n|r_n)P(d_n|r_n)P(r_n|r_{n-1}) \quad (2)$$

The term $P(r_n|r_{n-1})$ in Eq. 2 represents the turn-taking patterns, i.e, the way speakers take turn in the conversation, modeled as a simple bi-gram model. In other words, the role taken by a speaker at turn n depends by the role taken by the previous speaker at the turn $n-1$. Turn-taking patterns have been proven effective in recognizing formal roles in several datasets [6], [9], [21]. Bi-gram models are typically enough to capture most of the information compared to higher order n-grams and they can be estimated by counting.

The term $P(X_n|r_n)$ represents the probability of the prosodic feature vector modeled using a Gaussian Mixture Model (GMM) trained by standard EM on prosodic features labeled with role r. The number of components is empirically fixed to four. The term $p(d_n|r_n)$ represents the probability of the turn duration and is modeled using a Gamma distribution similarly to [11]. Its parameters are estimated by maximum likelihood estimation using the turns labeled with role r. Also turns durations in conversations are strongly related to social phenomena [11].

The recognition step consists in finding the mapping $\varphi^*(S) \rightarrow R$ speakers-to-role such that the likelihood 2 is maximized i.e.:

$$\varphi^* = \arg\max_{\varphi(.)} \prod_{n=2}^{N} P(d_n|\varphi(s_n))P(X_n|\varphi(s_n))P(\varphi(s_n)|\varphi(s_{n-1}))$$

Drawing a parallel with Automatic Speech Recognition $P(r_n|r_{n-1})$ represents the "Language Model", i.e., the prior information of a role sequence, while $P(d_n|r_n)$ and $P(X_n|r_n)$ represent the acoustic model composed of two different feature streams (duration and prosody). The Language model is a probability value while the other two terms are pdf thus similarly to ASR systems, a scaling factor is introduced to bring them in comparable ranges. The scaled likelihoods in Equation 2 thus becomes:

$$p(M) = \prod_{n=2}^{N} P(X_n|r_n)^{\gamma 1} P(d_n|r_n)^{\gamma 2} P(r_n|r_{n-1}) \quad (3)$$

Where $\gamma 1$ and $\gamma 2$ are obtained on a development data set in order to bring the pdf to same range of values as the language model.

This simple model accounts for information on turn-taking patterns, turn durations and prosodic behavior of speakers; however the only term able to capture dependencies between speakers is $P(r_n|r_{n-1})$ while the emission probability

$p(d_n|r_n)$ and $P(X_n|r_n)$ only depends on the current role r_n neglecting the history in the sequence.

Social roles are indicative of group behaviors and the influence that a speaker has on others has been pointed as a central effect in determining those roles, see e.g. [16]. The influence is verified not only on the speech activity but also on the prosodic behavior and on the visual features (body movement, focus of attention; for instance a Protagonist would induce Supporters to look at him while speaking).

Thus the following modification is proposed: observation associated with the nth turn not only depends on the speaker role that generated the turn but also on the previous speaker role, i.e., $p(d_n|r_n, r_{n-1})$ and $p(x_n|r_n, r_{n-1})$. The rationale behind this consists in the fact that, for instance, a protagonist may have a different prosodic behavior in taking turn after a neutral speaker or after another protagonist. Drawing again a parallel with ASR, this can be seen as a left-context role model, where the four distributions $p(.|r_n)$ are replaced with the sixteen left-context dependent model $p(.|r_n^{r_{n-1}})$. The probability of a sequence becomes then:

$$p(M) = \prod_{n=2}^{N} P(d_n|r_n^{r_{n-1}})P(X_n|r_n^{r_{n-1}})P(r_n^{r_{n-1}}|r_{n-1}^{r_{n-2}}) \quad (4)$$

$P(d_n|r_n^{r_{n-1}})$ designates a gamma distribution whose parameters are estimated by maximum likelihood. $p(x_n|r_n^{r_{n-1}})$ designates a four-components GMM obtained performing MAP adaptation on means and weights corresponding to the $p(x_n|r_n$ GMM. Turn taking patterns are modeled as before, i.e., $P(r_n^{r_{n-1}}|r_{n-1}^{r_{n-2}}) = P(r_n|r_{n-1})$.

Figure 2 (Model 2) represents equation 4 using the same DBN formalism as before. The dashed extra edges that are introduced respect to Model 1 can be seen as a form of "influence" that the role of the speaker $n-1$ has on the speaker n both in terms of turn duration and in terms of prosody. The inference step, as before, consists in finding the mapping $\varphi^*(S) \rightarrow R$ speakers-to-role such that the likelihood 4 is maximized.

The third type of information here investigated is related to the correlation between formal and social roles. As shown in Figure 1, in the AMI data the two schemes do not appear independent. This information can be modeled simply computing probabilities $p(r_n|r_{n-1}, f_n)$, i.e., the probability that the speaker at turn n takes the social role r_n knowing that his/her formal role is f_n and the previous speaker has role r_{n-1}. Note that f_n, the formal role of speaker taking turn n, is assumed known and it is constant over the entire meeting. The new model is referred as Model 3 and its likelihood can be written as follows:

$$p(M) = \prod_{n=2}^{N} P(d_n|r_n^{r_{n-1}})P(X_n|r_n^{r_{n-1}})P(r_n|r_{n-1}, f_n) \quad (5)$$

When probabilities $p(r_n|r_{n-1}, f_n)$ are estimated, smoothing is applied to leverage the effect of the small dataset.

	Accuracy
Duration (influence)	0.50
Prosody (influence)	0.53
Model 2	0.65

TABLE III
ACCURACY OF CONTEXT DEPENDENT MODELS FOR TURN DURATION AND PROSODIC FEATURES. WHEN COMBINED TOGETHER (MODEL 2) THEY ACHIEVE A 65% ACCURACY.

V. EXPERIMENTS

Experiments are run on the five annotated meetings using a leave-one-out approach where the training/tuning is done on four meetings and the test is done on the remaining one. The procedure is repeated such that each meeting is used for testing; the test set thus does not contain any speaker from the training set. During the training, role labels are used to infer the model parameters used then for testing on the left out meeting. Scaling factors are obtained on the training data set, and then applied in the test meeting.

The test is done following the same procedure described in [16], i.e., using a one minute long window centered around a given time instant where the reference speaker role is the most frequent role that the participant had in the window. Thus the social role of each speaker is assumed constant over the window length of one minute. The center of the window is then progressively shifted by 20 seconds and the procedure is repeated till the end of the meeting. As the speakers social role is considered constant in the one-minute window, φ^* is obtained exhaustively searching the space of possible φ (four speakers and four roles for a total of $4^4 = 256$ possible mappings) and selecting the one that maximize the likelihood. Performances are reported in terms of accuracy and are obtained averaging the results on the left out meetings.

Table I reports the performances of the turn-taking patterns, the duration features and the prosodic features used individually and combined together using Model 1. It can be noticed that bigram turn-taking patterns achieve the highest accuracy, compared to duration and prosody features. Model statistics reveal that, on average, the protagonist produces longer turns compared to Supporters and Neutral, the most common bigram is the *[Protagonist Supporter]* bigram and Neutral turns are characterized by low energy/speech rate. The three different types of informations combined achieves an accuracy of 59%.

Let us now consider the left-context modeling (Model 2) as well as the use of information given by formal roles (Model 3). Table II reports their performances. Explicit influence modeling increases the accuracy from 0.59 to 0.65 for Model 2 compared to Model 1. Furthermore Model 2 appears largely superior to Model 1 in recognition of the Protagonist and the Neutral roles, i.e., the most and the least engaged roles in the conversation.

Table III reports the performances of context-dependent models for duration and prosody features only, i.e., without combining them with turns statistics. Comparing tables III and I it can be noticed that context dependencies increase

	Random	Turns (Unigram)	Turns (Bigram)	Duration	Prosody	Model 1
Accuracy	0.26	0.35	0.49	0.43	0.41	**0.59**

TABLE I
ACCURACY OF MODEL 1 AND ITS COMPONENTS (TURN-TAKING PATTERNS, TURN DURATION AND PROSODIC MODEL) IN RECOGNIZING THE FOUR SOCIAL ROLES.

	Total	Protagonist	Supporter	Neutral	Gatekeeper
Model 1	0.59	0.61	0.62	0.68	0
Model 2	0.65	0.70	0.63	0.79	0
Model 3	**0.68**	**0.72**	**0.65**	**0.80**	**0.15**

TABLE II
TOTAL AND PER-ROLE ACCURACY OBTAINED BY MODEL 1, 2 AND 3.

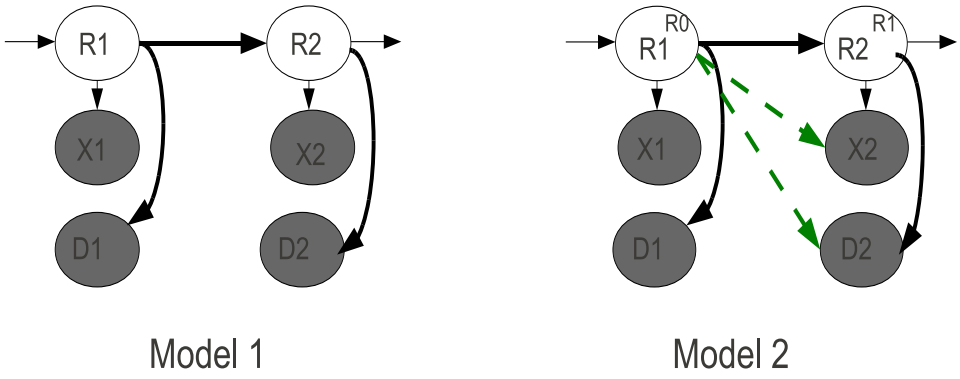

Fig. 2. Proposed DBN models: Model 1 is a multi-stream Markov process, Model 2 aims at explicitly modeling influence between speakers through left-context role models or equivalently assuming that the previous role has an influence on the observations of the current role.

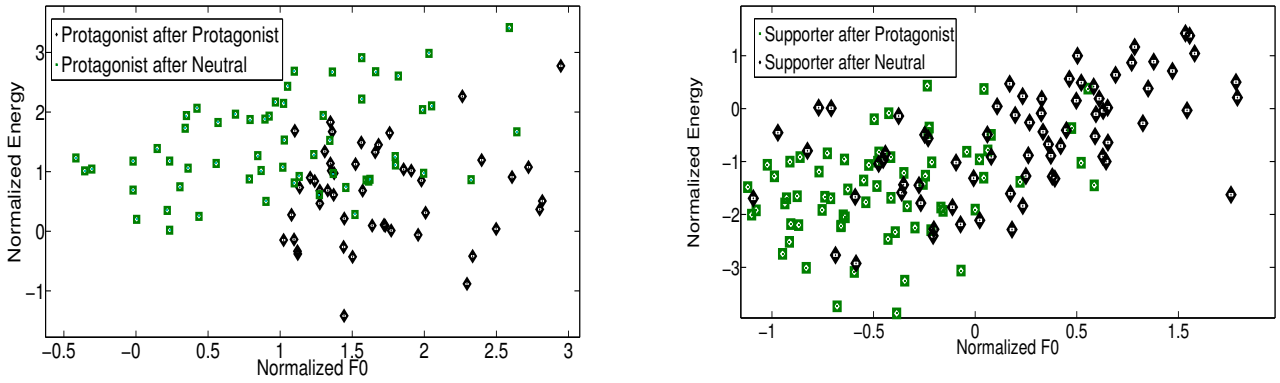

Fig. 3. Scatter of normalized f0 and normalized energy (meeting ES2002d) in case of turns generated by a Protagonist speaker after another Protagonist or a Neutral speaker (Left plot) and in case of turns generated by a Supporter after a Neutral speaker or a Protagonist (Right Plot).

the performance by 10% relative compared to the context independent models.

Figure 3 plots the scatter of normalized mean f0 and normalized energy averaged over turns generated by a Protagonist and a Supporter for a particular meeting (ES2002d). The scatter plots also those statistics in case the turns is generated after a Neutral or another Protagonist. It can be notice that, those features have different statistics if the turn is produced after another Protagonist or after a Neutral speaker. Similar differences are observed for Supporters taking turns after a Protagonist or after a Neutral speaker. In other words, the prosodic behavior of a speakers is conditioned by their previous speaker social role; this phenomenon is actually captured with the left-context role modeling is able to better model acoustic influences of a speaker on others.

The social role which is recognized the worst in both cases is the Gatekeeper as it is a rare role (less then 5% of total time) in the dataset. Nor model 1 or model 2 are able to recognize instances of Gatekeeper. Whenever the formal role information is considered (Model 3) performance reaches 68% and few

instances of the Gatekeeper role are recognized.

VI. Discussion and Conclusions

Social roles characterize the relationships between group members and they can account for several phenomena studied in conversations like engagement, hot-spots and dominance. Furthermore they explicitly account for the contribution of each individual speaker to the group discussion. Methods for automatically indexing, retrieving and summarizing archives of spontaneous conversations would largely benefit from this type of information.

Automatic role recognition in meeting recordings like the AMI corpus have mainly addressed formal roles. This work presents an investigation on language-independent social role recognition in meetings using the same methodology and the same non-linguistic features proposed in the context of formal roles. Those features are typical of social signaling [18] in human interactions.

The use of turn-taking patterns, turns duration and prosodic features integrated into a single generative conversation model achieves an accuracy of 59%. This model is then extended to account for joint speaker/roles dependencies at the acoustic level (or according to the interpretation in [16], the influence) achieving an accuracy of 65%. The protagonist, the supporter and the neutral role are recognized well above the chance, while the gatekeeper which is a rare role in the corpus, is completely missed by this model.

The last contribution consists in investigating the statistical dependency between the formal and the social role. In fact the meeting supervisor, i.e., the project manager, appears to take the gatekeeper role consistently more then others. Integrating the formal role information in the conversation model, increase the recognition rate to 68% permitting the recognition of Gatekeeper instances. This recognition rate is comparable to what reported in other corpora like the Mission Survival Corpus.

Several other language-independent features will be investigated in future works like speaker overlaps/interruptions, disfulencies and the use of non-verbal vocalizations (laughter, hesitations, etc.) as well as longer and more complex dependencies between speakers. Furthermore annotation of several other AMI meetings recordings is currently ongoing and future works will study how those findings scale on larger datasets.

Acknowledgment

This work has been supported by the Swiss National Science Foundation under the NCCR IM2 grant and by the EU Network of Excellence SSPNet and by the Hasler Fundation under SESAME grant. The authors would like to thank the University of Edinburgh for providing the social role annotations.

References

[1] Sacks H., Schegloff D., and Jefferson G., "A simple systematic for the organization of turn-taking for conversation," *Language*, , no. 5, 1974.

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

[3] B.J. Biddle, "Recent developments in role theory," *Annual Review of Sociology*, vol. 12, pp. 67–92, 1986.

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

[5] Banerjee S. and Rudnick A., "Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants.," *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, 2004.

[6] Salamin H., Favre S., and Vinciarelli A., "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, November 2009.

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

[10] Vinciarelli A., "Capturing order in social interactions," *IEEE Signal Processing Magazine*, September 2009.

[11] Grothendieck J et al., "Social correlates of turn-taking behavior.," *Proceedings of the International Conference on Audio Speech and Signal Processing (ICASSP) 2010.*

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

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

[14] Rienks R. and Heylen D., "Dominance detection in meetings using easily detectable features," *Proceedings of MLMI*, 2005.

[15] Zancaro M. et al., "Automatic detection of group functional roles in face to face interactions," *Proceedings of the International Conference on Multimodal Interface (ICMI)*, 2006.

[16] Dong W. et al., "Using the influence model to recognize functional roles in meetings," *Proceedings of the International Conference on Multimodal Interface (ICMI)*, 2007.

[17] Lepri B., "Multimodal recognition of social behaviors and personality traits in small group interaction.," *PhD Thesis, Univerisity of Trento (Italy)*, 2009.

[18] Vinciarelli A., Pantic M., and Bourlard H., "Social signal processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, 2009.

[19] 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.

[20] Dines J. et al., "The segmentation of multi-channel meeting recordings for automatic speech recognition," *Proceedings of the International Conference on Speech and Language Procesing (ICSLP) 2006*.

[21] Laskowski K., "Modeling norms of turn-taking in multi-party conversation," in *In proceedings of ACL (Association for Computational Linguistics)*, 2010.

[22] 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.