Language-Independent Socio-Emotional Role Recognition in the AMI Meetings Corpus

Fabio Valente[1], Alessandro Vinciarelli[2]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] University of Glasgow, G12 8QQ Glasgow,UK
fabio.valente@idiap.ch, Alessandro.Vinciarelli@dcs.gla.ac.uk

Abstract

Social roles are a coding scheme that characterizes the relationships between group members during a discussion and their roles "oriented toward the functioning of the group as a group". This work presents an investigation on language-independent automatic social role recognition in AMI meetings based on turns statistics and prosodic features. At first, turn-taking statistics and prosodic features are integrated into a single generative conversation model which achieves a role recognition 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 dependencies between the formal and the social role that participants have; integrating the information related to the formal role in the model, the recognition achieves an accuracy of 68%.

Index Terms: AMI meetings Corpus, role recognition, social and formal roles, turn-taking patterns, social signals.

1. Introduction

Conversation analysis and role recognition have been an active research fields for long time [1], [2]. Only recently statistical approaches have been used to model, analyze and automatically extract this type of information from archives of spoken conversations aiming at providing richer informations as compared to to the one extracted from purely transcribed speech. In between those, a lot of attention has been devoted to the recognition of roles. Speaker roles are stable behavioral patterns [2] that speakers exhibit during a conversations. Automatic role recognition based on statistical classifiers has been studied in meeting recordings like the CMU corpus [3], the AMI corpus [4], [5] and the ICSI corpus [6] as well as Broadcast [7] and telephone [8] 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. The roles considered in those studies are mainly formal roles constant over the entire duration of the conversation, e.g., the Project Manager during a professional meeting or the anchorman during a broadcast conversation. Several other coding schemes that characterize the speaker roles in conversations with respect to the dynamic of the discussion have been proposed. In between those, the Socio-Emotional roles [9], inspired from Bales work [10], characterize the relationships between group members and their roles "oriented toward the functioning of the group as a group". This coding scheme attributes to each participant in the discussion a role in between 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. Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants and can be related to phenomena like engagement, hot-spots [11] and also social dominance widely studied in meetings. It is intuitive that the same speaker can change social role over time but its 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. Furthermore they can provide another level of understanding on the meeting dynamics which can be used for indexing, retrieval or summarization purposes. For instance, a meeting (or a meeting chunk) where participants are neutral most of the time will not be as informative as a meeting where speakers take on turn the Protagonist role.

Previous works on automatic social role recognition have been mainly performed on corpora that study group decision making like the Mission Survival Corpus [12], [9] where SVM classifiers trained on audio and video activity features extracted from a 10 seconds long windows are used for this purpose. Later in [13], 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 a single speaker. The influence that each speaker has on other participants is modeled through the chains coupling which can recognize joint activity of multiple speakers. Furthermore, studies like [12],[13] have outlined how social roles appear strongly correlated with non-linguistic cues.

This work investigates the recognition of social roles in the AMI corpus, a collection of professional meetings. Previous studies on those data have mainly addressed the recognition of static (formal) roles [4], [5], [6]. The paper provides three contributions: at first a language-independent generative model that accounts for turn-taking patterns, turn duration and prosody is proposed; those features have been mainly considered in literature for the recognition of formal roles. After that, 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. Finally the paper investigates the dependencies between social and formal roles, proposing the use of the formal role as auxiliary information

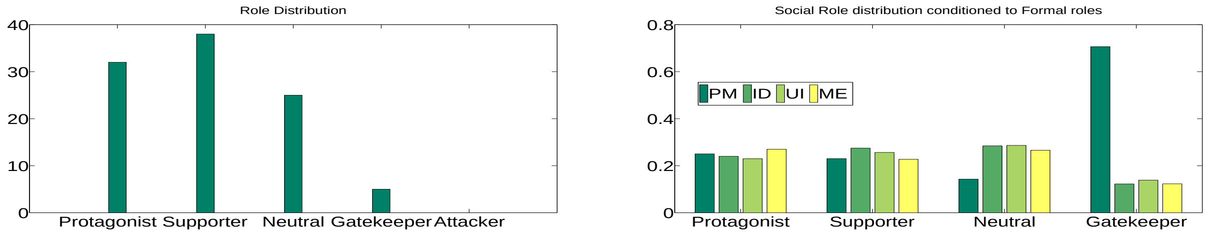

Figure 1: (Left Plot) Role distribution (percentage of total time) on the 5 meetings annotated in terms of social roles. (Right Plot) Social role distribution conditioned to the formal role that each speaker has in the meeting.

for the social role recognition. Let us now describe the data and their annotations.

2. 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 play the role of a design team (Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial Designer (ID)) tasked with designing a new remote control. Those roles will be referred as formal roles. The meeting is supervised by the project manager. 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 scenario meetings (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, the same annotation guidelines and heuristics used to produce the role annotations in the Mission Survival Corpus [9] (CHIL project) 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 available for each time instant. Manual annotations are then post-processed as described in [13]; 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 (percentage of total time) 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.

3. Feature extraction

The audio data from the headset microphones 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 [14] to produce very precise speaker boundaries. 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 [15] and [16], 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. Furthermore the following prosodic measures, related to the engagement in the discussion [11], are extracted from the speech regions that compose each turn: F0 frequency (mean, standard deviation, minimum, maximum and median for each turn), energy (mean and standard deviation 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 [11]. 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 associated with roles:

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

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, $r_n \in \{P, S, N, G\}$ designates its social role, $f_n \in \{PM, ID, UI, ME\}$ designates its formal role. During the training, the social role r_n is known while the recognition consists in inferring r_n when all the other elements in Eq. 1 are known.

4. 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 account for the social roles. Its probability can be written as:

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

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 on 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 [4], [7], [15]. Bi-gram models are typically sufficient to capture most of the information 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 vectors belonging to the role r. The number of components is empirically fixed to

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

four. The term $p(d_n|r_n)$ represents the turn duration probability and is modeled using a Gamma distribution similarly to [8]. 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 [8]. The recognition step consists in finding the mapping $\varphi^*(S) \to R$ speakers-to-role such that the likelihood 2 is maximized i.e.:

$$\varphi^* = \arg\max_{\varphi(.)} \prod_{n=1}^{N} P(d_n|\varphi(s_n))P(X_n|\varphi(s_n))P(\varphi(s_n)|\varphi(s_{n-1})) \quad (3)$$

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.

This simple model accounts for information on turn-taking patterns, turn durations and prosody; 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. [13]. The influence is verified not only on the speech activity but also on the prosodic behavior, body movement and focus of attention (for instance a Protagonist would induce Supporters to look at him while speaking). Thus the following modification is proposed: the observations associated with the nth turn not only depend 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 designated with $p(.|r_n^{r_{n-1}})$. The probability of a sequence becomes then:

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

$P(d_n|r_n^{r_{n-1}})$ designates a gamma distribution whose parameters are estimated by maximum likelihood from turn labeled r_n in left context r_{n-1}. $p(x_n|r_n^{r_{n-1}})$ 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 $\varphi^*(S) \to R$ 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:

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

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

5. 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; thus the test set 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 [13], 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 one-minute long window. 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 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 maximizes the likelihood. Performances are reported in terms of accuracy and are obtained averaging the results on the left out meetings.

Table 1 reports the performance 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. The 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 together achieve 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 2 reports their performances. Explicit influence modeling increases the accuracy from 59% (Model 1) to 65% (Model 2). Furthermore Model 2 appears largely superior to Model 1 in recognition of the Protagonist and the Neutral roles. Analysis reveal that Protagonists turns have different prosodic and durations statistics when they are produced for instance 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 left-context role modeling is able to better capture acoustic influences from a role to others. The social role which is recognized the worst is the Gatekeeper as it is a rare role in the dataset (less then 5% of total time). Nor model 1 or model 2 are able to recognize instances of Gatekeeper. Whenever the formal role information is added (Model 3) performance reaches 68% and

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

Table 1: 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 2: Total and per-role accuracy obtained by Model 1, 2 and 3.

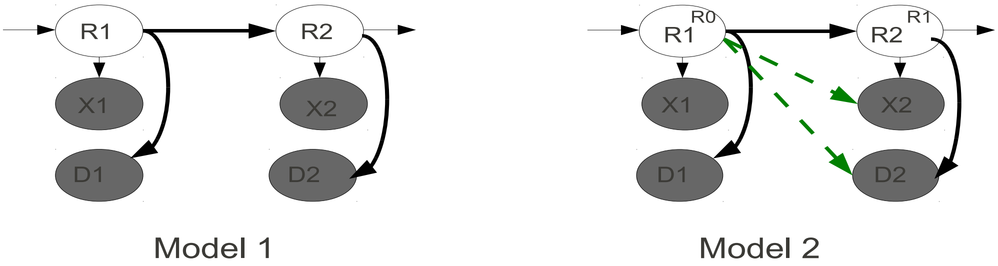

Figure 2: Proposed DBN models: Model 1 is a multi-stream HMM, Model 2 aims at explicitly modeling influence between speakers through left-context role models or equivalently assuming that the previous role has an influence on the observations of the current role.

few instances of the Gatekeeper role are recognized as consistently taken by the Project Manager.

6. Discussion and Conclusions

Social roles characterize the relationships between group members and they can be related to several phenomena studied in conversations like engagement, hot-spots and dominance thus providing richer information for accessing and summarizing those data. Furthermore they characterize the contribution of each speaker to the conversation. Automatic role recognition in meeting recordings like the AMI corpus have mainly addressed static (formal) roles that do not change during the recording. 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/static coding schemes.

The use of turn-taking patterns, turn duration and prosodic features integrated into a single generative conversation model recognize social roles with an accuracy of 59%. This model is then extended to account for joint speaker/role dependencies at the acoustic level (or according to the interpretation of [13], 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 these models. The last contribution consists in investigating the statistical dependency between the formal and the social role. Integrating the formal role information in the conversation model, increase the recognition rate to 68% permitting the recognition of Gatekeeper instances. The total 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.[2]

7. 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] 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 ICSLP*, 2004.

[4] Salamin H. et al., “Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction,” *IEEE Transactions on Multimedia*, vol. 11, November 2009.

[5] 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.

[6] 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.

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

[8] Grothendieck J et al., “Social correlates of turn-taking behavior.,” *Proceedings of ICASSP 2010*.

[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] Dines J. et al., “The segmentation of multi-channel meeting recordings for automatic speech recognition,” *Proceedings of ICSLP 2006*.

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

[16] 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.

[2] Authors thank the University of Edinburgh for providing the AMI corpus social role annotations. This work was funded by the Swiss Science Foundation through IM2 grant, by the EU through SSPnet grant and by the Hasler foundation through the SESAME grant