INVESTIGATING THE IMPACT OF LANGUAGE STYLE AND VOCAL EXPRESSION ON SOCIAL ROLES OF PARTICIPANTS IN PROFESSIONAL MEETINGS

Ashtosh Sapru[1,2] and Herve Bourlard[1,2]

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

ABSTRACT

This paper investigates the influence of social roles on the language style and vocal expression patterns of participants in professional meeting recordings. Language style features are extracted from automatically generated speech transcripts and characterize word usage in terms of psychologically meaningful categories. Vocal expression patterns are generated by applying statistical functionals to low level prosodic and spectral features. The proposed recognition system combines information from both these feature streams to predict participant's social role. Experiments conducted on almost 12.5 hours of meeting data reveal that recognition system trained using language style features and acoustic features can reach a recognition accuracy of 64% and 68% respectively, in classifying four social roles. Moreover, recognition accuracy increases to 69% when information from both feature streams is taken into consideration.

***Index Terms*—** Social Role Labeling, Language style features, acoustic features.

1. INTRODUCTION

Analyzing spoken documents in terms of speaker role information is useful for enriching the content description of multimedia data. It can be used in applications like information retrieval, enhancing multimedia content browsing and allowing summarization of multimedia documents [1]. Speaker roles are stable behavioral patterns in an audio recording and the problem of role recognition consists in assigning a label, i.e., a role to each of the speakers. Automatic labeling of speaker roles has been widely studied in case of Broadcast News recordings. These roles are imposed from the news format and relate to the task each participant performs in the conversation like anchorman, journalists, interviewees, etc. In the last few years automatic role recognition has also been investigated for meeting recordings and broadcast conversations. Typical roles in these studies can vary with environment and applications such as Project Manager in AMI corpus [2], student, faculty member in ICSI corpus [3]. Common features used in these studies extract relevant information from conversation features, lexical features and prosody [4, 5, 6].

For the studies mentioned above participants role is formal and considered to remain constant over the duration of entire audio recording. Other role coding schemes have also been proposed in literature which put roles in a more dynamic setting, such as socio-emotional roles (here after referred to as social roles) [7, 8, 9, 10]. Social roles describe relation between conversation participants and their roles "*oriented towards functioning of group as a group*". Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants and can be generalized across any type of conversation. They are also related to phenomena studied in meetings like social dominance, engagement and also hot-spots [11]. Besides, these social roles can also provide cues about state of meeting. Meeting segments where participants take more active roles are likely to have richer information flow compared to segments where participants only take passive roles.

Previous studies on automatic social role recognition have mainly considered nonverbal information. Common format in many of these studies is to predict social role within a segment of recording, where the role of each participant is assumed to stay constant. One of the first studies investigated social roles in meetings recorded for problem solving sessions [7]. They used a support vector machine classifier to discriminate between social roles using features expressing participants activity from both audio and video. The study in [8] used the same corpus and similar features, however, a generative framework was used to represent the influence of other participants on the current speakers role. Other studies [12, 9, 10] have also investigated social role recognition in professional meetings (AMI corpus) using combination of prosodic cues and word ngrams.

In comparison to earlier studies, the focus of our work is to investigate the influence of social roles on speaking style of meeting participants. Moreover, our investigations are conducted on a larger database containing 128 different speakers for a total of nearly 12.5 hours of meeting data. To the best of our knowledge, this is the first study which extensively investigates speaking style features for automatic social role recognition. By speaking style we mean, *how participants talk* instead of *what they talk*. Our definition of speaking style includes both language style, as well as, acoustic analysis of vocal expression patterns.

Existing findings in psychology have linked language style with use of simple functional words - pronouns, prepositions, articles and a few other categories. Language style has been used to analyze personality traits [13]. 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 [14] and group cohesion [15].

In this work, we have used a psychologically validated state-of-art text analysis module, Linguistic inquiry and word count (LIWC) [16], to extract the language style features of meeting participants. Furthermore, we also analyze relation between social roles and vocal expression patterns by applying large scale acoustic feature extraction. The acoustic feature vector is calculated from various low level prosodic and spectral features. The role recognition system is implemented by fusing information from both language style and acoustic features in a discriminative classifier. In the remainder of the paper, Section 2 describes the data and the role annotations, Section 3 and Section 4 provide details of the methodology used for extracting language style and acoustic features, the experiments and results are presented in Section 5. The paper is then concluded in Section 6.

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. A subset of 59 meetings from the scenario portion of AMI Meeting Corpus containing 128 different speakers (84 male and 44 female participants) is selected from the entire corpus. Subsequently each meeting was segmented into short clips (with a minimum duration of 20 seconds) based on presence of long pauses i.e. pauses longer than 1 second. Within each such meeting segment social role of the participant is assumed to remain constant. From each meeting, a total duration of approximately 12 minutes long audio/video data was selected. Meeting segments are resampled so as to cover the entire length of recording comprising various parts of meeting such as openings, presentation, discussion and conclusions.

Since social roles [7] are subjective labels and require human annotators, the annotation scheme was implemented as follows. Each annotator is asked to view and listen the entire video segment and tasked with assigning a speaker to role mapping based on a list of specified guidelines. These guidelines define a set of acts and behaviors that characterize 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 assumes a cooperative attitude demonstrating attention and acceptance and providing technical and relational support; *Neutral* - a speaker that passively accepts others ideas; *Gatekeeper* - a speaker that acts like group moderator, mediates and encourages the communication; *Attacker* - a speaker who deflates the status of others, expresses disapproval and attacks other speakers. At least 10 annotators were asked to label each video clip. A total of 1714 clips were annotated using this procedure.

Figure 1 shows the distribution of roles over all the meeting segments present in the data set.

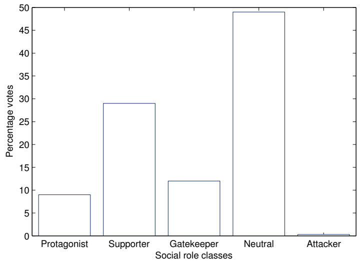


Fig. 1. Social role distribution in the annotated corpus. The vertical axis represents percentage votes for each class as labeled by multiple annotators.

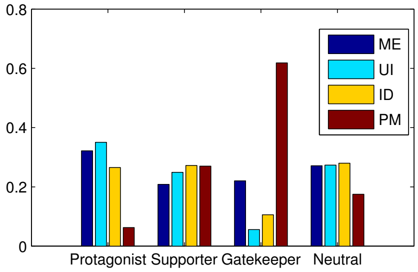
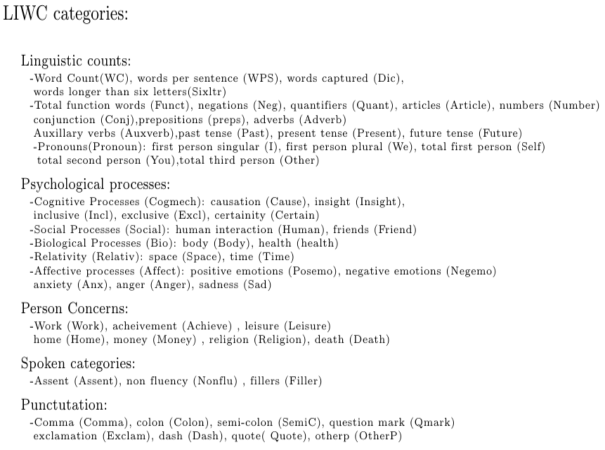


Fig. 2. Social role distribution conditioned on formal role that the speaker has in the meeting.

It can be seen that the neutral role has been labeled most often by annotators. This is followed by supporter, gatekeeper and protagonist. Comparatively the attacker role has received the fewest labels from multiple annotators. A reason for this distribution may be due to collaborative nature of AMI meetings. The reliability of labeling scheme as measured through Fliess's kappa shows a value 0.5 which is considered to have moderate agreement according to Landis and Koch's criterion [7]. In terms of inter annotator agreement we find that neutral label is most reliable one as measured through category wise κ statistic with a value of 0.7. The intermediate level of agreement is present for supporter 0.36 and gatekeeper 0.38 labels. This is followed by the protagonist role which shows a fair level of agreement with a κ value of 0.29. One difference from the earlier studies [9, 10] is the higher percentage of gatekeeper role. A reason for this behavior can be explained from Figure 2 which reveals that annotators were more likely to associate the role of PM with gatekeeper compared to other formal roles.

3. LANGUAGE STYLE FEATURES

LIWC is a computerized text analysis program that quantifies the language style used by participants in a conversation. It counts 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 words) processes. The core part of LIWC program is a dictionary composed of almost 4500 words. 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. Figure 3 describes the the details of organization of various categories and subcategories.



LIWC categories:

Linguistic counts:
-Word Count(WC), words per sentence (WPS), words captured (Dic),
words longer than six letters(Sixltr)
-Total function words (Funct), negations (Neg), quantifiers (Quant), articles (Article), numbers (Number)
conjunction (Conj),prepositions (preps), adverbs (Adverb)
Auxillary verbs (Auxverb),past tense (Past), present tense (Present), future tense (Future)
-Pronouns(Pronoun): first person singular (I), first person plural (We), total first person (Self)
total second person (You),total third person (Other)

Psychological processes:
-Cognitive Processes (Cogmech): causation (Cause), insight (Insight),
inclusive (Incl), exclusive (Excl), certainity (Certain)
-Social Processes (Social): human interaction (Human), friends (Friend)
-Biological Processes (Bio): body (Body), health (health)
-Relativity (Relativ): space (Space), time (Time)
-Affective processes (Affect): positive emotions (Posemo), negative emotions (Negemo)
anxiety (Anx), anger (Anger), sadness (Sad)

Person Concerns:
-Work (Work), acheivement (Achieve) , leisure (Leisure)
home (Home), money (Money) , religion (Religion), death (Death)

Spoken categories:
-Assent (Assent), non fluency (Nonflu) , fillers (Filler)

Punctutation:
-Comma (Comma), colon (Colon), semi-colon (SemiC), question mark (Qmark)
exclamation (Exclam), dash (Dash), quote(Quote), otherp (OtherP)

Fig. 3. Language style categories defined in LIWC.

In the current study, all the speech transcripts were generated using output of AMI-ASR system [17] which has a word error rate of less than 25%. In our analysis, all the utterances of a participant within a meeting segment were processed using LIWC. The number of categories along which word usage can be measured in LIWC sums up to 80. The scores generated by LIWC were interpreted as percentage of number of occurrences of words belonging to a category, with the exception of word count (WC). Since the transcripts used in this work were generated from an ASR system, and as such were not marked with punctuation, the scores generated for punctuation categories were not considered. Moreover, some of categories like WPS (words per sentence) were considered as redundant, since we consider all the words spoken by the participant within a meeting segment as a single sentence. Also, preliminary experiments revealed that some of the categories like death, religion, home had very few to null occurrences. So out of a total of 80 categories, 24 were discarded and rest of the analysis was carried out on remaining 56 categories.

Table 1. 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, Logrithm of Harmonics to Noise ratio(HNR)

4. ACOUSTIC FEATURE EXTRACTION

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. We have been motivated in following this approach, as recent studies have revealed that systematically generated large acoustic features can capture complex phenomena, like leadership emergence in online speeches [18] and recognizing conflicts [19] 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 Independent headset microphones (IHM) is processed at frame rate to extract low level descriptors (LLDs), within every segment of meeting recording, where the social role of participant is constant. The next pass projects each participant's LLD contour to a fixed size feature vector using statistical functionals.

Table 1 shows the extracted LLDs, including F0 and speech energy, which have been shown as being informative for social role recognition [10], voice quality features including jitter and shimmer that capture the perception of harshness in voice and spectral features including MFCC coefficients, which have been shown to be related to aspects of personality like openness and conscientiousness [20]. Statistical and regression functionals defined in Table 2 were used to obtain features vectors from the contours of LLDs and their first order derivatives. This procedure has the advantage that it yields fixed size feature vector for each participant

Table 2. Set of functionals applied to contours generated from lld descriptors.

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

Table 3. Correlation values between LIWC categories and social roles. All values are statistically significant at $p < 0.0001$

Process	Category	Examples	ρ
		Protagonist	
Linguistic	Word Count	-	0.32
Cognitive	Causation	because,effect	0.29
Cognitive	Inhibition	stop,constrain	0.28
Cognitive	Inclusive	and,include	0.27
Linguistic	Quantifiers	few,many	0.26
		Supporter	
Personal	Achieve	earn,win	-0.20
Personal	Work	job,project	-0.16
Relativity	Past	common verb	-0.16
Linguistic	Function	-	0.14
Spoken	Assent	okay,yes	0.12
		Gatekeeper	
Linguistic	We	us,our	0.27
Social	Social	they,talk	0.26
Cognitive	Inclusive	and,include	0.26
Linguistic	Prepositions	to,with	0.26
Spoken	Non fluency	few,many	0.25
		Neutral	
Linguistic	Function	-	-0.57
Cognitive	Cogmech	cause,know	-0.54
Linguistic	Pronouns	i,them	-0.50
Linguistic	Auxiliary verbs	am,have	-0.52
Linguistic	Prepositions	to,with	-0.49

within the meeting irrespective of the duration for which they are speaking and hence can be used directly for social role classification. All the acoustic features were extracted from open-source feature extractor openSMILE [21]. Additionally, features derived from turn duration and counts were also included.

5. EXPERIMENTS

To verify the relevance of linguistic style on social role of meeting participants, we performed a correlation analysis. Table 3 shows the top correlated categories with each of the social roles. All the reported, Pearson's correlation coefficient ρ values are statistically significant. Protagonists have highest correlation with WC (word count), suggesting that for this role, participants hold the maximum conversation floor. Also, protagonists have higher correlation with cognitive process like causation, inhibition etc. These dimensions are expected to represent complex cognitive thinking, with higher usage of causation in language showing higher levels of thinking. Supporters show lower correlation with linguistic categories, the top positively correlated variables is assent, which is associated with use of words like okay, agree, yeah, yes etc. The analysis of *We* words suggests that they are more likely to be used by participants taking the gatekeeper role. This category of linguistic process is in general associated with feeling of commitment towards the group, as well as, maintenance of group longevity [15]. Neutral speakers are negatively correlated with linguistic features. They tend to use fewer function words and show lower cognitive thinking, keeping with their role of being mostly passive speakers.

Table 4. Correlation values between prosodic and spectral features and social roles. All values are statistically significant at $p < 0.0001$

LLD	Functional	ρ
	Protagonist	
Δ MFCC 8	range	0.32
Δ MFCC 8	max	0.31
LogHNR	Quadratic regression error	0.27
	Supporter	
Log Energy	skewness	0.2
Δ MFCC 0	kurtosis	0.19
	Gatekeeper	
MFCC 4	range	0.32
Δ MFCC 7	minimum	-0.30
	Neutral	
MFCC 1	minimum	0.61
Log Energy	range	-0.55

For acoustic feature analysis, we again performed a correlation based study. Table 4 shows the various acoustic features which have higher correlation with different social roles. While earlier studies in social role recognition have mainly focused on voicing related features [9, 10], Table 4 reveals the importance of spectral features. MFCC derived features are more correlated with social roles in comparison to voicing related features, i.e., features derived directly from F0 contour and voice quality features like jitter and shimmer. For both protagonists and gatekeepers, we observe higher correlation with range functionals, indicating a higher variation in articulation. Supporters in general show much lower correlations with acoustic features, the most correlated features are related to higher order moments, skewness and kurtosis. Participants acting as neutral speakers are more likely to be passive speakers, as indicated by negative correlation with variation in energy.

The social role recognition experiments in this work use support vector machine (SVM) as a supervised classifier. Each feature vector in the algorithm is considered as data point in a multidimensional feature space and the algorithm works by constructing a separating hyperplane between two classes. SVM with a linear kernel has been selected for classification in this study as they are well suited for classification with high dimensional acoustic and linguistic feature sets, due to their robustness against overfitting. For the multiclass

Table 5. Per role F-measure, Precision and Recalls obtained in recognizing social roles for chance, language style, acoustic and combination model. Significance of accuracy w.r.t. chance (*: $p < 0.01$). Significance of classwise recall for fusion model w.r.t. unimodal (acoustic) model ($\dagger : p < 0.01$).

	Per-role F-measure (Recall/Precision)				Overall Accuracy
Model	Protagonist	Supporter	Gatekeeper	Neutral	
Baseline (chance)	0.17 (0.17/0.17)	0.47 (0.48/0.46)	0.23 (0.22/0.25)	0.22 (0.23/0.22)	0.32
Language style	0.50 (0.46/0.54)	0.72 (0.80/0.66)	0.50 (0.45/0.57)	0.70 (0.66/0.75)	0.64*
Acoustic	0.48 (0.47/0.48)	$\mathbf{0.79}(\mathbf{0.86}^{\dagger}/\mathbf{0.74})$	0.50 (0.46/0.55)	0.74 (0.68/0.81)	0.68*
Fusion	$\mathbf{0.53}(\mathbf{0.55}^{\dagger}/\mathbf{0.52})$	0.78 (0.83/0.75)	**0.52 (0.48/0.57)**	$\mathbf{0.75}(\mathbf{0.71}^{\dagger}/\mathbf{0.80})$	$\mathbf{0.69}$*

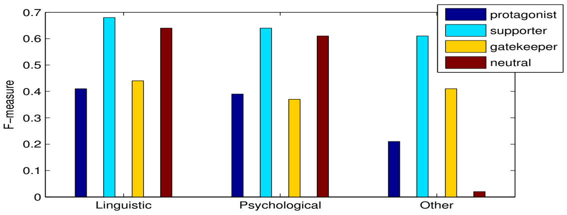


Fig. 4. Performance of various LIWC categories for social role recognition. Other represents a merged category(spoken and personal concerns).

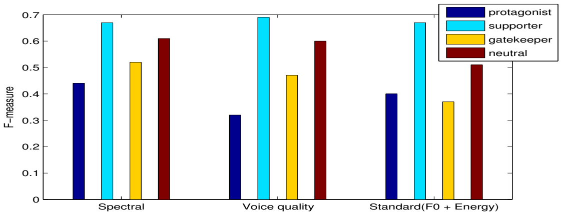


Fig. 5. Performance of three acoustic feature sets which characterize the influence of voice quality, spectral features and standard features (F0 and energy).

classfication a one on one strategy was used and each binary classifier was trained using libsvm [22], with cost parameter C set to 1. A normalized representation of features was used in all SVM experiments.

For evaluation of the proposed method, experiments were conducted using repeated cross-validation, wherein a set of meetings was kept for training/tuning the model parameters while a distinct set of meetings was used for evaluation. The partition of meetings was done keeping in view, that participant with same speaker identity does not appear in both training and test set. The ground truth for participant role labels was derived by majority voting. An initial filtering was done to consider only those meeting segments where a participant is active, also a few meeting segments, where majority voting resulted in participant having an attacker role label were not considered(see Figure 1). All the models were evaluated on a separated test set and performance measured in terms of recognition accuracy and F-measure/Precision/Recall.

For our first experiment, we investigate the relevance of various word categories in LIWC dictionary for social role recognition. Figure 4 shows the F-measures of social role labels for word groupings belonging to linguistic, psychological (cognitive and social categories were merged) and other (spoken categories and personal concerns where merged) categories. The figure reveals that for both linguistic and psychological word categories, the classification of all the social role labels is better than chance (shown in Table 5). However, protagonists and neutrals have very low recognition for other, where only gatekeeper and supporter roles are recognized. A reason for this behavior can be due to correlation values in Table 3, where gatekeepers and supporters are associated with all the three categories, in comparison to protagonists and supporters.

For the acoustic feature set, we performed a comparison study for the relevance of spectral, voice quality features against the standard feature set based on F0 and RMS energy. Figure 5 shows the results of our study when all the features in each feature set were used. We can note that for all the acoustic feature sets, the recognition of social roles as revealed by their F-measures is better than chance(shown in Table 5). However, spectral features in general perform better in recognizing most social roles, in comparison to voice quality features (derived from LLDs like jitter, shimmer, log HNR) and standard features. This was also seen in Table 4, where higher correlations with MFCC derived features for these roles where observed. Another reason for their improved performance may be due to higher number of feature dimensions which are obtained by using spectral features in comparison to standard features.

Our final investigation, compares the performance of multistream approach, obtained from decision level fusion of acoustic and language style feature streams against individual feature sets. A feature selection stage using information gain criterion, was used to reduce the dimensionality of acoustic feature set. Table 5 shows the social role recognition performance for fusion, language style and acoustic feature streams. For each feature stream a SVM model was trained which, in turn, was used to generate the decision scores for each instance of data. Fusion SVM were trained by combining the normalized scores for all the feature streams. Similar cross-validation strategy was followed while training and testing for all the individual feature streams, as well as, model fusion. Table 5 numbers reveal that while model fusion marginally improves the overall accuracy over acoustic model, there is a statistically significant improvement in recognition performance for protagonist and neutral roles. On the other hand, acoustic features show better performance in recognizing supporters.

6. CONCLUSION

In this work, we presented the influence of social roles on speaking style of participants in professional meetings. Our investigations revealed that automatically extracted language style features and acoustic features are correlated with social roles. The proposed automatic role recognition system, was able to perform non trivial classification of four social roles, reaching a recognition accuracy of 64% and 68% for language style and acoustic feature streams respectively. Furthermore, a decision level fusion of language style and acoustic features, improves the model's performance to 69%, and shows statistically significant improvement for some social roles. In summary, proposed approach leads us to conclude that both language style and acoustic features are relevant for automatic social roles recognition. We further plan to extend our study on other conversation environments.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] Gabriel Murray, Pei-Yun Hsueh, Simon Tucker, Jonathan Kilgour, Jean Carletta, Johanna Moore, and Steve Renals, "Automatic Segmentation and Summarization of Meeting Speech," *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2007.

[2] 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.

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

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

[5] 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.

[6] Marcela Charfuelan and Marc Schroder, "Investigating the prosody and voice quality of social signals in scenario meetings," in *Proceedings of the 4th international conference on Affective computing and intelligent interaction*, 2011.

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

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

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

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

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

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

[13] 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.

[14] Dairazalia Sanchez-Cortes, Petr Motlicek, and Daniel 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.

[15] Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker, "Language style matching as a predictor of social dynamics in small groups," *Communication Research*, 2010.

[16] 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.

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

[18] Felix Weninger, Jarek Krajewski, Anton Batliner, and Björn Schuller, "The voice of leadership: models and performances of automatic analysis in online speeches," *IEEE Transactions on Affective Computing*, 2012.

[19] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[20] Tim Polzehl, Sebastian Moller, and Florian Metze, "Automatically assessing personality from speech," in *Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing*, 2010.

[21] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, 2010, MM '10.

[22] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.