Using Linguistic and Vocal Expressiveness in Social Role Recognition

Theresa Wilson
University of Edinburgh
School of Informatics
twilson@inf.ed.ac.uk

Gregor Hofer
University of Edinburgh
School of Informatics
ghofer@inf.ed.ac.uk

ABSTRACT
In this paper, we investigate two types of expressiveness, linguistic and vocal, and whether they are useful for recognising the social roles of participants in meetings. Our experiments show that combining expressiveness features with speech activity does improve social role recognition over speech activity features alone.

Author Keywords
role recognition; group interaction; analysis of user states

ACM Classification Keywords
I.2.7 Computing Methodologies: Artificial Intelligence—*Natural Language Processing*

INTRODUCTION
Researchers in social psychology have long been interested in in group interaction, including the roles that people take on when interacting with others [3, 1]. More recently, researchers have begun to explore the automatic recognition of roles [2, 10, 5, 6], and whether this information can be used to provide automatic coaching and feedback to meeting participants, in order to improve their interaction and effectiveness in meetings [11].

In this paper, we investigate the recognition of *social roles*. Pianesi et al. [10] describes social roles as "roles oriented toward the functioning of the team as a group." Drawing from early research by Benne and Sheats [3] and Bales [1], Pianesi et al. distinguish five social roles: *the Gatekeeper*, who moderates the discussion; *the Protagonist*, who takes the floor and drives the discussion with her ideas; *the Supporter*, who is attentive to the discussion and encouraging; *the Attacker*, who disapproves and challenges the ideas of others to the detriment of the group; and *the Neutral*, who presents a passive audience.

To date, speech activity features and motion features, such as fidgeting, have been applied to the task of recognising social roles. Given the nature of social roles, we hypothesise that features capturing speaker *expressiveness* would also be useful for recognising social roles. For example, we would expect the Protagonist to be more expressive than the Gatekeeper. In this paper, we investigate two types of speaker expressiveness for this task: linguistic and vocal. Our experiments show that combining expressiveness features with speech activity does improve social role recognition over speech activity features alone.

RELATED WORK
Researchers have investigated the automatic recognition of different types of roles in meetings, such as task or communicative roles [2, 10, 5] and roles that reflect a participant's function within a project or organisation [6]. Pianesi et al. [10] and Dong et al. [5] also investigate the recognition of social roles.

In [10], Pianesi et al. present the Mission Survival Corpus, which they annotate with both task and social roles. They also conduct experiments in recognising social and task roles using SVMs with speech activity and fidgeting features. Dong et al. [5] propose an influence model for recognising task and social roles.

DATA
The AMI Meeting Corpus [4] 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, Marketing Expert, User Interface Designer, and Industrial Designer) tasked with designing a new remote control. Each team participates in a series of four meetings, which represent different stages in the design process: project kick-off, functional design, conceptual design, and final design and evaluation. Meetings in the AMI Corpus have been transcribed, and a subset of the data has been richly annotated with everything from dialogue acts and topics to abstractive meeting summaries.

The existing annotations in the AMI Corpus include subjective content annotations [14], which capture when opinions, sentiments, (dis)agreements and other types of subjective content are being verbally expressed. These annotations are used to train a system for automatically classifying utterances as subjective or objective. This information is then used in our system for recognizing social roles.

Submitted for review to IUI 2010.

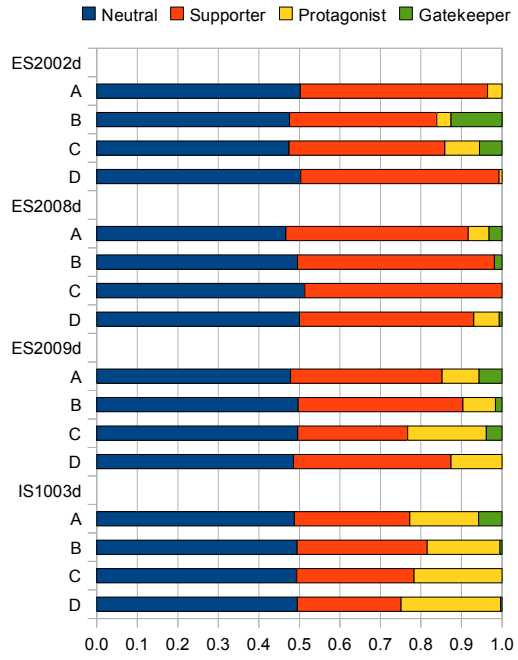

Figure 1. Distribution of social roles by meeting participant

For this work, we added a layer of social role annotations to four scenario meetings[1] using the annotation guidelines that were used to produce the role annotations in the Mission Survival Corpus [10]. Although we generally followed the existing guidelines, we did not annotate task roles, as was done for the Mission survival Corpus. We felt that much of the task role information already was being captured by the existing dialogue act annotations. We also allowed our annotators to mark more fine-grained role spans (i.e., roles that last for fewer than 5 seconds). Each meeting is approximately 40 minutes with four participants.

Figure 1 shows the distribution of the social role annotations in the four meetings, broken down for each participant. Note that in this data no participants take on the Attacker role. Although the amount of Neutral role taking is about the same across all the meetings and participants, the difference in the percentage of time the participants spend in the Supporter, Protagonist and Gatekeeper roles reveals clear differences in the character of the four meetings. Meeting IS1003d, for example shows all of the participants spending a fair amount of time in the Protagonist role, which indicates that everyone was participating in the discussion and contributing ideas. Meeting ES2008d, on the other hand, obviously has a much different tone, with the majority of the participants spending their time being supportive and not putting forth their own ideas.

FEATURES

We experiment with two types of features for recognising social roles: speech activity features and features that capture linguistic and vocal *expressiveness*. Speaker activity features have been used previously for role recognition, but the features we use for capturing expressiveness are new. To capture linguistic expressiveness, we use a feature that represents the subjectivity of the language being used. For vocal expressiveness, we use a feature that indicates the expressiveness of the speech based on prosodic cues. All features are produced automatically.

Speaker Activity Features

We use two speech activity features: *overlap* and *speakercount*. The overlap feature is a binary feature that indicates whether a participant other than the current speaker is also speaking. The speakercount feature is a count of the number of participants currently speaking.

Subjectivity Feature

The *subjectivity* feature is a binary feature that indicates whether or not the speaker is currently saying something subjective. To calculate this feature, we first train an automatic system to classify utterances as subjective or objective. The output of this system then is used to determine the value of the subjectivity feature for a given instance.

The subjectivity classifier is trained using the output of the AMI ASR system [7], spurt segments, and the subjective content annotations. The spurt segments are identified automatically by breaking the speech at pauses of at least 0.4 seconds. The gold subjective/objective labels for the spurts are determined based on their overlap with the subjective content annotations.

Our subjectivity classifier is an ensemble classifier that combines the output of three classifiers, each trained using different n-gram information. The first classifier is trained using n-grams of words from the ASR system. The second classifier is trained using n-grams of characters. The third classifier is trained using n-grams of phonemes, also produced by the ASR system. Raaijmakers et al. [12] showed that very shallow linguistic features, such as n-grams of characters and phonemes, are competetive or better than word n-grams for recognizing subjective utterances, and that combining all three types of n-gram information yields the best performance.

To train and evaluate the subjectivity classifier, we use 13 meetings from the AMI Corpus with subjective content annotations[2] and perform 13-fold cross validation. Following Raaijmakers et al. [12], each n-gram classifier is trained using BoosTexter AdaBoost.HM [13]. The parameters for the learning algorithm for each fold (number of rounds of boosting, n-gram length, and type of n-gram) are determined individually for each classifier and fold by separating out a small set from the training data for that fold to use for optimization.

Raaijmakers et al. uses a weighted linear combination of the output of the n-gram classifiers to create an ensemble classifier. For this work, we experimented with two additional

[1] ES2002d, ES2008d, ES009d, and IS1003d

[2] The four meetings with social role annotations are a subset of these meetings.

	Rec	Prec	F
Baseline	1.000	0.507	0.670
Words	0.584	0.700	0.619
Linear Combination	0.608	0.715	0.657
Majority Vote	0.614	0.727	0.666
Any Subjective Vote	0.732	0.642	0.684

Table 1. Results for subjectivity recognition

methods for combining the output of the n-gram classifiers: 1) majority vote and 2) any subjective vote (i.e., if any n-gram classifier predicts subjective, then classify as subjective). Results for these three ensemble methods are given in Table 1. We also show results for the majority-class baseline and the word n-gram classifier[3].

The classifier with the highest F-measure is the classifier that predicts subjective if any of the n-gram classifiers makes a subjective prediction. While this classifier has a lower precision than the others, including the word baseline, it is a clear winner in terms of recall. To generate features for the role recognition experiments, we use the output of this classifier.

Expressive Prosodic Feature

In speech-driven animation, the goal is to derive from the speech signal appropriate movements of the head and facial features. One task in this area is identifying when speech is expressive and synthesising appropriate head movements to accompany the speech. Our expressive prosodic features are produced by a system from Hofer and Shimodaira [8] trained to do this task.

Hofer and Shimodaira developed a system for predicting four types of head motion from the speech signal: *pause*, where the speaker is at rest, *default* for neutral activation, and *shift* and *shake*, which represent moderate and high levels of head activity. This system is trained on speech features (MFCC, Energy) from three speakers. It predicts the four types of head motion with a 70% accuracy.

For this work, we interpret the head motions predicted by Hofer and Shimodaira's system as different levels of expressiveness. We segment the speech of each speaker in the AMI meetings using their system. The result is a sequence of segments labeled according to the four expressiveness types outlined above. As an illustration, Fig 2 shows a speech signal and the resulting predicted sequence of expressive segments. These labels are used to calculate the expressive prosodic features for our experiments.

EXPERIMENTS AND RESULTS

For our role recognition experiments, we compare five different configurations of features:

ALL - all features are used
SPACT - only speech activity features
EXPR - only the expressiveness features
SPACT+PROSEXPR - speech activity plus expressive prosodic feature
SPACT+SUBJEXPR - speech activity plus subjectivity feature

In our experiments, we train conditional random fields (CRFs)[4] [9] using 4-fold cross validation, holding the participants in each meeting out in turn for evaluation. Because we are interested in the effect of our linguistic and vocal expressiveness features for role recognition, we limit our experiments to the portions of the data where the participants are speaking.

It is not clear what the best unit of classification is for recognizing participant roles. The systems that produce our expressiveness features use different levels of granularity. Pianesi et al. [10] used 330 ms segments, with sliding windows that varied from 0 to 14 seconds. For this work, we decided on seconds as our unit of classification.

Our CRF models are trained using the features for seconds n_0 to n_{-4} as well as the output of the previous second. Table 2 shows the results of our experiments. The F-measure for recognizing each role is given. This value is the average over the participants.

We hypothesized that the expressiveness features would be useful for recognising participant role, and our experiments generally support this hypothesis. The classifier that uses all features outperforms the classifier using only the speech activity features for all roles, although only the results for the Neutral class trend toward significant (paired t-test, $p = 0.096$). Adding the expressive prosodic features to the speech activity features yields improvements of 5.7% and 19.4%, respectively, for recognizing the Supporter and Protagonist roles, although the results are not significant due to the high variance over over the individual speakers. Nevertheless, given that our expressiveness features are based entirely on the output of automatic systems, one of which was trained on very different data, we find these results very encouraging.

CONCLUSIONS AND FUTURE WORK

In this paper, we investigate whether automatically derived linguistic subjectivity and expressive prosodic features can be used to improve social role recognition of participants in meetings. We found that combining these expressiveness features with speech activity features improves social role recognition over speech activity alone.

In future work, we will continue to explore the use of expressiveness for social role recognition. One important question is what is the upperbound on performance that we can expect from subjective information. To answer this, we will investigate how well the subjective content annotations themselves do for recognizing roles. Also, it may be that discriminating certain types of subjective content (e.g., agreements, positive sentiment) may be useful for this task. As for the expressive prosodic features, they were produced by a system that was trained on data very different from the AMI data used for

[3]Results for the character and phoneme n-gram classifiers are very similar.

[4]CRF++: http://crfpp.sourceforge.net/

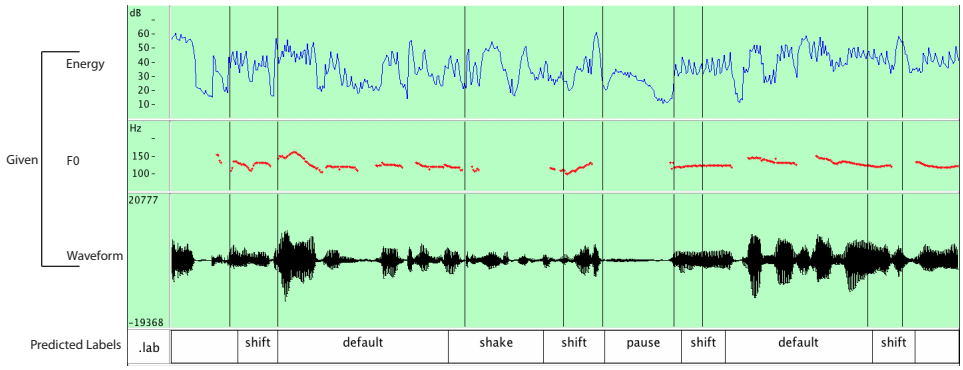

Figure 2. The top shows a times series of speech features that are used to predict a sequence of expressive segments.

	Neutral	Supporter	Protagonist	Gatekeeper
ALL	0.152	0.570	0.596	0.013
SPACT	0.102	0.560	0.516	0.000
EXPR	0.102	0.539	0.605	0.007
SPACT+PROSEXPR	0.109	0.592	0.616	0.000
SPACT+SUBJEXPR	0.133	0.562	0.515	0.011

Table 2. Results (F-measure) for Role Recognition

these experiments. We plan to retrain this system on AMI data, which should also improve performance. It will also be important to investigate how our single expressive prosodic feature compares to more detailed prosodic information. Finally, we plan to investigate the use of visual features that capture the amount of participant motion, as these types of features have proved useful in the past for role recognition.

REFERENCES

1. R. F. Bales. Task roles and social roles in problem-solving groups. In T. M. Newcomb and E. L. Hartley, editors, *Readings in Social Psychology*, pages 437–447. Hold, Rinhart and Winston, 1958.

2. S. Banerjee and A. I. Rudnicky. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proc. of Interspeech-ICSLP*, 2004.

3. K. D. Benne and P. Sheats. Functional roles of group members. *Journal of Social Issues*, 4:41–49, 1948.

4. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus. In *Proc. of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*, 2005.

5. W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proc. of ICMI*, 2007.

6. S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proc. of ICMI*, 2008.

7. T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The AMI system for the transcripton of meetings. In *Proc. of ICASSP*, 2007.

8. G. Hofer and H. Shimodaira. Automatic head motion prediction from speech data. In *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

9. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 2001.

10. F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language, Resources and Evaluation*, 41:409–429, 2007.

11. F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri. Multimodal support to group dynamics. *Personal and Ubiquitous Computing*, 12:181–195, 2008.

12. S. Raaijmakers, K. Truong, and T. Wilson. Multimodal subjectivity analysis of multiparty conversation. In *Proc. of EMNLP*, 2008.

13. R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

14. T. Wilson. Annotating subjective content in meetings. In *Proc. of LREC*, 2008.