Automatic Clustering of Conversational Patterns from Speech and Motion Data

S. Feese[1], B. Arnrich[1], G. Tröster[1], B. Meyer[2], K. Jonas[2]

[1]*Wearable Computing Lab., ETH Zurich, Zurich, Switzerland.* [1]*sfeese@ethz.ch*
[2]*Social and Business Psychology, University of Zurich, Zurich, Switzerland*

Abstract

Behavioral observation of teams is critical not only for team research, but also for leadership trainings. However, most current behavior observation methods used in team research often rely on manual video annotation which is time consuming and thus costly. In our work, we follow a sensor-based approach to automatically extract important pattern that capture conversational structure. Relying on automatic speaker detection from multiple microphones, we first employ a rule-based clustering to measure the proportions of typical conversational regimes taking place in a conversation and extend our method to learn more complex regimes from data using finite mixture models. Furthermore, we investigate the use of motion activity cues extracted from motion sensors. We apply our methods to a recent study on leadership in small groups and show how team meetings can be characterized.

Introduction

Conversation regimes characterize who speaks with whom in the course of a conversation. Typical pattern that can occur in a three person meeting are monolog, dialog and chat. In monologs, only one person speaks while other persons listen, whereas in dialogs two persons speak in an alternating manner. In the Chat regime all three persons participate in the part of the conversation and talk equally much. For a three person meeting all possible combinations are illustrated in the top line of Figure 2.

In this work we will focus on clustering such conversation patterns from audio recordings in a meeting scenario including three participants. We will first employ a rule-based clustering to illustrate the benefit of clustering. Second, we will use Gaussian Mixture Models to cluster conversational slices that are described by speech cues in order to capture more of the conversational dynamics. Additional to the speech modality, we apply the clustering method to motion data.

Related Work

Previous work on automatic analysis of social interactions in small groups dealt with automatic inference of conversation structure, analysis of social attention and the detection of personality traits and roles. A review on the on the topic can be found in [1]. These works mostly relied on computer vision based techniques and speech related cues such as speaking length, speaker turns and number of interruptions. Conversational patterns have been discovered using topic models in [2] and conversational scenes have been modeled with Dynamic Bayesian Networks in [3]. Wearable sensors in form of sociometric badges have been employed by Pentland and collaborators to measure “honest signals” in daily life [4]. In addition to speech activity cues as in [2], we investigate motion activity cues extracted from data measured with on-body motion sensors.

Methods

The complete processing chain is illustrated in Figure 1. Speaker diarization was performed by employing a threshold based approach. On a sliding window (length: 25~ms; step size: 10~ms) the signal energy was calculated for each group member. Speech was detected if the energy difference between the group members’ energy value and the mean value of the other group members was greater than an empirically set threshold. Speech activity segments shorter than 30~ms were then removed and segments of the same speaker within 1000~ms were merged.

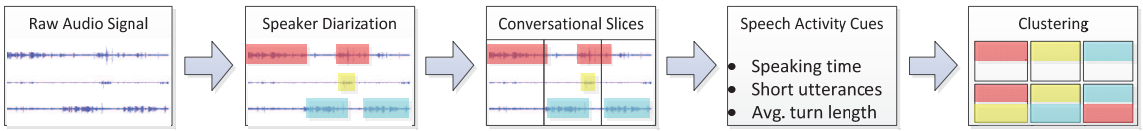

Figure 1. Processing steps.

To summarize the speech status of each group member over a two minute period conversational slices are extracted. Based on the speech activity cues proposed in [2], we calculate for each slice the following speech activity cues: total speaking time, number of short utterances as well as median and inter-quartile-range of the turn duration. We choose two minute slices because it resembles a rough minimum of a dialog length if one assumes that one speech act is about 30 seconds in length and that it requires a minimum of two alternating turns of each person.

For the rule-based clustering, we only use the total speaking time of each group member per conversational slice. We define seven clusters that correspond to the conversational regimes monolog, dialog and chat as follows: In a monolog one group member speaks more as twice as much as each other group member, in a dialog two group members speak more as twice as the third group member each. The chat regime captures all other cases.

In order to incorporate more of the discussion dynamics, we also learn Gaussian Mixture Models with k (k between 1 and 10) mixture components using for each group member the above mentioned speech activity cues as observed data. The emission probabilities are modeled as single 12 dimensional Gaussians with diagonal covariance and all model parameters are learned with the Expectation Maximization Algorithm (see [5] for more on mixture models).

One important parameter to estimate is the number of mixture components, which we assume to be equal to the number of clusters in the data. As we are interested in clusterings that are most reliable against noise in the data, we follow a resampling approach to evaluate cluster stability for a different number of components. We employ the cluster stability criterion presented in [6] which measures the similarity of clustering solutions that were obtained on random subsets of the data to the solution that was obtained on the complete data set.

The principle of clustering conversational slices can be generalized to other modalities. As an example, we show how the method can be used to discover conversational clusters from motion data. Analogous to above, each conversational slice can be represented through a set of motion activity cues which simply capture the activity of each body limb in terms of movement time and the number of short movements.

Data Set

We apply the clustering methods to a recent study on leadership behavior [7] in which groups of three participants were led by two different types of leaders. Fifty-five groups were asked to work on a hidden-profile decision making task to rank four fictive candidates with regard to their suitability for an open job position. Under the guidance of the group leader, the group had to discuss the suitability of each candidate and agree on a rank order which served as a measure of group performance. Groups with an individually considerate leader who encouraged all team members to contribute to the solution achieved a higher performance as opposed to groups that were lead authoritarian.

In the experiment each group member was equipped with a separate clip-on lapel microphone to record the speech of all group members (sampling frequency 44.1 kHz). Additionally, the upper body motion of each group member was captured with six inertial measurement units (XSens MTx) which were located on the lower arms, the back and the head (sampling frequency 32 Hz). In total, the data set contains 44 group discussions (16 groups were led authoritarian and 18 with individual consideration) with three participants each. Our data set totals to over 15 hours of discussion time.

Results and Discussion

In Figure 2 the results of the rule-based clustering solution are shown in terms of the mean speech distribution and the corresponding mean head orientation. The different types of conversation regimes are identified as expected. In case of monologs, one group member dominates in terms of speaking length. It appears natural that the speaker is looked at, which we can infer from the mean head orientation. In case of dialogs, speech is distributed across two speakers and considering the mean head orientation we can confirm that the dialog partners are orientated towards each other. To illustrate how the clustering solutions can discover some of the reasons why groups of individually considerate leaders achieved higher group performances, we present the cluster distributions across all groups. The bottom line of Figure 2 displays the mean and standard deviation of the relative amount of time that each cluster is present in a group discussion. In the experiment the leader was always person P1. As one can see, groups that were led by individually considerate leaders stayed longer in the regimes in which persons P2 and P3 speak with each other. This suggests that individually considerate leaders gave their followers more room to talk to each other which in turn led to more information being shared and thus higher group performance. The results of the model based clustering of speech activity cues are presented in the top of Figure 3. Not considering the trivial clustering solution of one cluster, we find from the stability analysis (top left) that models with four components are most stable in our data. The found clusters can be described by their corresponding cluster means which are illustrated in the top right of Figure 3 (different colors encode different group members). Cluster 1 represents conversational slices in which person P2 speaks mostly, while person P1 contributes with varying turn lengths. Person P3 speaks most in Cluster 2 and P1 acknowledges with short utterances. In Cluster 3 person P1 speaks most with long turns and person P2 and P3 acknowledge with short utterances. Finally, Cluster 4 represents conversational slices in which the group members talk with short utterances.

The bottom of Figure 3 presents the results of the cluster analysis in which we used motion activity cues to capture the movement of the lower arms, the back and the head during the conversational slices. The stability analysis suggests three clusters, which as we can see from the cluster means belong to slices in which one of the group member was active while the others did only shortly move their upper body limbs.

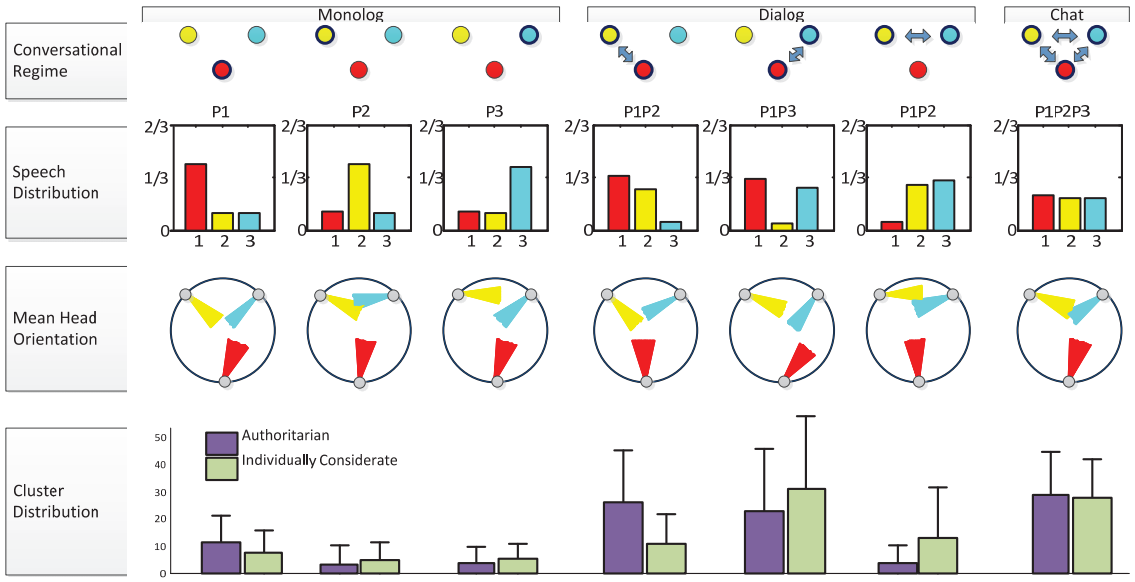

Figure 2. Rule-based conversational regimes with their corresponding speech distribution and mean head orientation. Only the speaking length of each group member per conversational slice was used for the clustering. Button row: Cluster distribution across discussions with two different leadership types.

Proceedings of Measuring Behavior 2012 (Utrecht, The Netherlands, August 28-31, 2012)
Eds. A.J. Spink, F. Grieco, O.E. Krips, L.W.S. Loijens, L.P.J.J. Noldus, and P.H. Zimmerman

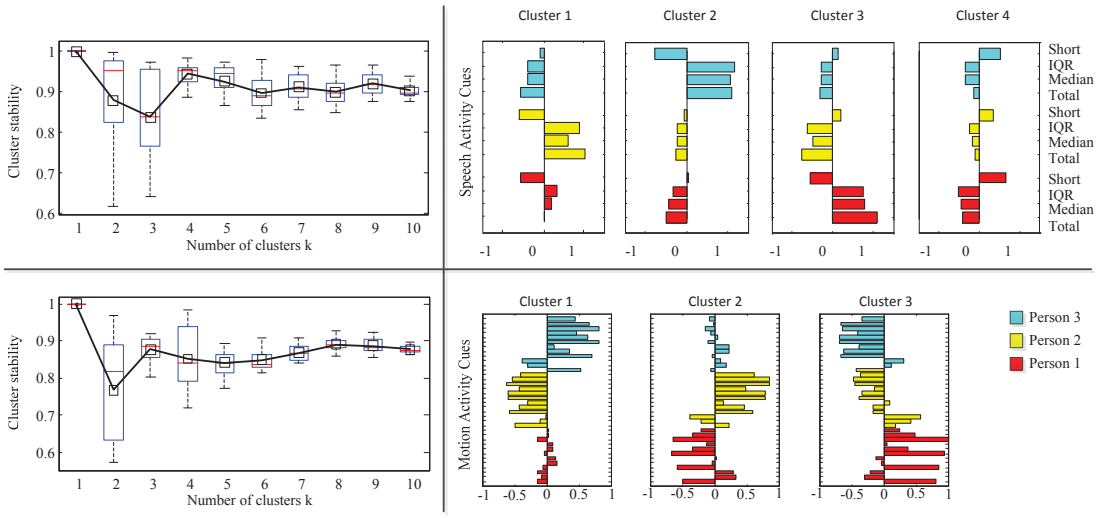

Figure 3. Top left: Cluster stability of speech activity cue based mixture modeling for different number of mixture components. Top right: Cluster means of the clustering solution for k = 4. Bottom left: Cluster stability using motion activity cues. Bottom right: Cluster means of the clustering solution for k = 3.

Conclusion

We have presented a rule-based method to cluster conversational slices into meaningful regimes and demonstrated their usefulness to discover possible reasons for higher group performance in a recent study on leadership. Furthermore, we have shown how speech and motion activity cues can be used to cluster conversational slices. While the clustering solutions based on speech activity appear to be stable and meaningful, the clusterings of simple motion activity cues appear to be less stable, suggesting that simple motion activity cues do not capture enough of the nonverbal communication expressed by body motion. In future, we will investigate other clustering algorithms and more meaningful motion cues such as head nodding, gesticulating and body postures.

References

1. Gatica-Perez, D. (2009). Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review. *Image and Vision Computing* **27**(12), 1775-1787.
2. Jayagopi, D.B., Gatica-Perez, D. (2010). Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models. *IEEE Transactions on Multimedia* **12**(8), 790-802.
3. Otsuka, K. (2011). Conversation Scene Analysis. *IEEE Signal Processing Magazine* **28**(4), 127-131.
4. Pentland, A. (2008). Honest Signals, Bradford Books.
5. Frühwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models, Springer.
6. Levine, E., Domany, E. (2001). Resampling Method For Unsupervised Extimation of Cluster Validity. *Neural Computation* **13**(11), 2573-2593.
7. Meyer, B., Jonas, K., Feese, S., Arnrich, B., Tröster, G., Schermuly, C.C. (2012). A social-signal processing view on leadership: Specific behaviors characterize considerate leaders. Paper accepted for presentation at the Academy of Management Annual Meeting 2012, Boston, MA.