Catharine Oertel

KTH Royal Institute of Technology
Lindstedtsvägen 24
10044 Stockholm
Sweden

catha@kth.se
http://www.speech.kth.se/~cogb

1 Research Interests

My main research interests are in the area of **multimodal multiparty interactions**, and **socially-adaptive dialogue systems** which exploit models derived from such interactions. More specifically, I am interested in measuring the **engagement** of a person in a conversation as well as how this reflect on the conversational dynamics of the whole group of people. I am interested in using **data-driven** approaches which use this knowledge in **human-robot interactions** to increase the collaboration between **children** and improve the perception of attentiveness in **conversational agents**.

1.1 Multiparty Dialogue Systems

In the last years more and more research has focused on the development of multiparty dialogue systems. For such systems it is important to identify whom of several users the dialogue system should address or whether the users are addressing each other and not the dialogue system. It is also important for such a system to know when it should be interrupting a conversation and when it should stay in the background and convey the impression that it is following the conversation, and to what degree etc. These questions are becoming particularly pressing to answer when the purpose of designing the dialogue system is for it to be used in a pedagogical settings or for longterm relationships with its user. In order to be able to answer all these questions, it is necessary to investigate human-human multiparty conversations in terms of their conversational dynamics. In the following paragraphs, I am going to summarize several studies we carried out, and which focused on understanding group dynamics. I will then detail how I am planning to use these findings in a multimodal multiparty dialogue system.

1.2 Conversational Involvement in Multiparty Conversations

For a dialogue system, designed with the purpose of fostering collaboration, it is important to know whether all participants are actually involved in the conversation, and to what degree they are interested in continuing with the conversation. If one of the participants is loosing interest, it can be advantageous for the system to interrupt the current speaker or to change the topic of the conversation. To this end we investigated whether it is possible to both classify the engagement of an individual person as well as the group involvement of the whole group of people (Oertel and Salvi, 2013). In order to make such an estimation as robust as possible for a potential later implementation in an online system, we used "eye–gaze" as a feature. We proposed a number of features (presence, entropy, symmetry and maxgaze) that summarise different aspects of eye-gaze patterns and that allowed us to describe individual as well as group behaviour over time. We used these features to define similarities between the subjects and we compared this information with the engagement rankings, the subjects expressed at the end of each interactions, about themselves and the other participants. We analyzed how these features relate to four classes of group involvement and we build a classifier that is able to distinguish between those classes with 71% of accuracy.

While these results appeared promising, the study did not distinguish between speaker and listener behaviours. In order to fill this gap, we focused on investigating listener categories separately. We distinguished between an "attentive listener", a "side participant" and a "bystander". We then devised a thin-sliced perception test where subjects were asked to assess listener roles and engagement levels in 15-second video-clips taken from a corpus of group interviews. Results showed that humans are usually able to assess silent participant roles. We also found that the frequency of audio backchannel as well as headnods are higher in a "attentive listener" than a "side–participant" and higher in a "side–participant" than a bystander. We also found that mutual–gaze as well as gaze from the speaker are significantly different between the various listener categories (Oertel et al., 2015). In a final study, we then wanted to investigate whether only the frequencies or also the prosodic realizations of the backchannel tokens were different. We therefore used the same corpus to sample backchannels produced under varying conversational dynamics. Amongst other things we wanted to understand i) which prosodic cues are relevant for the perception of varying degrees of attentiveness. We found that duration, intensity and f0 slope are important cues to distinguish between more and less at-

tentive backchannels.

1.3 Socially Aware Dialogue System

Currently I am working on combining the results of both studies by building a "multiparty listener category module" for the dialog system framework IRISTK (Skantze and Al Moubayed, 2012).To also be able to generate the subtleties in audio-backchannels we are currently working on extending the work described in (Oertel et al., 2016), to not only the ranking of different levels of attentiveness in feedback token but also the synthesis. This conversational speech synthesiser will moreover also be able to generate other conversational phenomena such as hesitations, false–starts and self–corrections. We are also working on including convincingly natural synthesis of headnods and smiles.

2 Future of Spoken Dialog Research

I think that in the next 5 to 10 years the field of dialogue research will move even more towards developing multiparty systems as well as systems which can lead a non-task oriented conversation. More and more corpora are becoming available and sensing technology is evolving and becoming more affordable which will foster the possibilities of building such systems. I think that more research will also be devoted to developing dialogue systems designed for longterm interactions. In order to be able to build such systems it will be necessary to investigate how to use information from previous conversation in the current one. For more targeted applications, such as for example pedagogical agents, it will also become more and more important to provide dialogue systems with the capabilities to express empathy and build a rapport.

3 Suggestions for Discussion

- How can we build dialogue systems for longterm interactions ?
- What are the biggest weaknesses for current state-of-the-art dialogue systems?
- How useful can the internet-of-things be for the advancement of dialogue systems ?
- How to make current dialogue systems better suited for interactions with groups of children?

References

Catharine Oertel and Giampiero Salvi. A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue. 2013. *Proceedings of the 2013 ACM on International Conference on Multimodal Interaction.* 99-106.

Catharine Oertel, Joakim Gustafson and Alan W. Black. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in feedback utterances. 2016. *Proceedings of Interspeech 2016.* accepted.

Catharine Oertel, Kenneth A. Funes Mora, Joakim Gustafson and Jean-Marc Odobez. Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions. 2015. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.* 107–114.

Gabriel Skantze and Samer Al Moubayed. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. 2012. *Proceedings of the 2012 ACM on International Conference on Multimodal Interaction.*

Biographical Sketch

Catharine Oertel is a final-year PhD student at the Department of Speech, Music and Hearing at the Royal Institute of Technology in Stockholm, Sweden. She is working on the modeling of multiparty human-human dynamics for building more socially aware dialogue systems within the Horizon 2020 "Baby Robot" Project. She is involved in ongoing research collaboration with both Idiap Research Institute as well as Carnegie Mellon University. She received her M.A. in "Linguistics: Communication, Cognition and Speech Technology"' in 2010, from Bielefeld University, Germany.