KTH Computer Science
and Communication

Modelling Engagement in Multi-Party Conversations

Data-Driven Approaches to Understanding Human-Human Communication Patterns for Use in Human-Robot Interactions

CATHARINE OERTEL

Doctoral Thesis

Stockholm, Sweden

TRITA-CSC-A-2017:05
ISSN-1653-5723
ISRN–KTH/CSC/ A-2017:05
ISBN: 978-91-7729-237-1

Cover by Kenneth Alberto Funes Mora

Tryckt av Universitetsservice US AB

KTH School of Computer Science and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framläggs till offentlig granskning för avläggande av filosofie doktorsexamen i tal- och musikkommunikation med inriktning på talkommunikation måndagen den 20 januari 2017 klockan 13.00 i F3, Kungl Tekniska högskolan, Lindstedtsvägen 26, Stockholm.

To my friends and family

Abstract

The aim of this thesis is to study human-human interaction in order to provide virtual agents and robots with the capability to engage into multi-party-conversations in a human-like-manner. The focus lies with the modelling of conversational dynamics and the appropriate realization of multi-modal feedback behaviour. For such an undertaking, it is important to understand how human-human communication unfolds in varying contexts and constellations over time. To this end, multi-modal human-human corpora are designed as well as annotation schemes to capture conversational dynamics are developed. Multi-modal analysis is carried out and models are built. Emphasis is put on not modelling speaker behaviour in general and on modelling listener behaviour in particular.

In this thesis, a bridge is built between multi-modal modelling of conversational dynamics on the one hand multi-modal generation of listener behaviour in virtual agents and robots on the other hand. In order to build this bridge, a unit-selection multi-modal synthesis is carried out as well as a statistical speech synthesis of feedback. The effect of a variation in prosody of feedback token on the perception of third-party observers is evaluated. Finally, the effect of a controlled variation of eye-gaze is evaluated, as is the perception of user feedback in human-robot interaction.

Acknowledgements

The last couple of years have been quite an adventure both personally and scientifically. I am very grateful for having had the opportunity to meet and collaborate with many great people along the way.

First and foremost I would like to thank my advisor Joakim Gustafson. For perfecting the balance of letting me grow academically without leaving me alone with the challenges of the PhD. For always having his door open and listen with an open mind and kind heart to even my craziest ideas. I could not have wished for a better advisor!

I would also like to thank Jens Edlund without whom I would probably never have joined TMH in the first place. His out-of-the-box thinking has inspired many of the corpora designs.

Thanks also belong to Jean-Marc Odobez for many discussions, very constructive feedback, great collaborations, and especially for making the KTH-Idiap project possible in the first place.

I would like to thank Alan Black for taking the time out of his super busy schedule and making the room for our weekly discussions over the course of the last 1.5 years. Collaborating with Alan made me think about my research in novel ways and for this I am very grateful.

Many thanks also go to Nick Campbell and Justine Cassell for sharing their knowledge in many inspiring discussions.

During the course of my PhD studies I had the honour and privilege to work in different research labs.

I would like to thank my friends and colleagues from Trinity: Brian, Emer, Cieran, Frank, John K., John D., Amelie and especially Céline for their support and encouragements. Without Céline, and her crash-course in statistics, programming and paper writing, my start into the PhD would have been much harder. I hope I will be able to do the same for the new PhD students coming now!

I also would like to thank my colleagues from CMU: Samuel, Alex, Yoishi, Rhan, Tanmay, Samantha, Brittany, Dave and Prasanna for being awesome and very inspirational colleagues and for introducing me into the world of HCI and Speech Synthesis.

Special thanks belong to Samuel, J.B. and Pei who have been my family over-seas and who made my stay so much more fun.

I would also like to thank the Perceptionists at Idiap: Samira, Alex, Kenneth, Nam, Wu Di, Rui, Gülcan, Yu, Weipeng for making me feel like one of your own.

Last but not least I would like to thank my past and present colleagues at TMH for always being supportive, taking part in so many recordings, test-runs and perception experiments and for creating a nice and friendly atmosphere which made me feel part of a greater team.

Special thanks go to Sofia, Jana, Simon, and Eva, my office mates who became friends, and who contributed so much my every-day well-being at the lab. Thanks for your awesomeness!

I would very much also like to thank my colleagues with whom I had the opportunity and privilege to work on various studies and without whom this thesis would not have been possible in its present form: Fred Cummins, Jens Edlund, Petra Wagner, Nick Campbell, Stefan Scherer, Céline de Looze, Stéphane Rauzy, Giampiero Salvi, Gabriel Skantze, Anna Hjalmarsson, Nigel Ward, Steven Werner, David Novick, Elizabeth Shriberg, L.P. Morency, Kenneth Funes Mora, Samira Sheikhi, Jean-Marc Odobez, Joakim Gustafson, Samer Al Moubayed, Jonas Beskow, Bajibabu Bollepalli, Ahmed Hussen-Abdelaziz, Martin Johansson, Maria Koutsombogera, José David Lopes, Jekaterina Novikova, Catharine Oertel, Gabriel Skantze, Kalin Stefanov, Gül Varol, Marcin Wlodarczak, Alexey Tarasov, Andreas Windmann, Fabio Tamburini, Denis Arnold, Jing Guang Han, John Dalton, Brian Vaughan, Ciaran Dougherty, Alan W. Black, Jana Götze, Mattias Heldner, Yu Yu, Uwe Altmann, Emer Gilmartin, Karl Weilhammer, Robin Siegemund, Ricardo Sá, Anton Batliner, Florian Hönig, Elmar Nöth

I want to very much thank my former supervisors from Bielefeld Thorsten Trippel, Dafydd Gibbon and Petra Wagner for providing the foundation and support which enabled me later to start my PhD studies.

Many thanks also belong to the people of ISCA-SAC: Maria, Na, Andi, Angel and Jeanin! ISCA-SAC has been a very important part of this journey for me. I am very grateful of having had the possibility to work with you and very happy that SAC is prospering so much now.
Many thanks also to Helen Meng and Lori Lamel for their great support in running SAC.

I would like to thank Bine, Joe, Cormac, Nike, Andreas, Meg, Samuel, J.B., Zofia, Marcin, Marc, Nina, José, Eva, Simon, Aga, Roger, Laurent and Noémie for your friendship, constant support, encouragements and for many great trips and adventures. A special thanks also belongs

to Nike for proofreading this thesis and Kenneth for designing the cover!

A million thanks go to Kenneth, for your always calm and happy attitude, your patience, your love and incredible support.

Finally, I want to thank my family. In particular Yvonne, Katharina, Anna-Marie, and Getrude. Without your love and support none of this would have been possible.

Contents

Publications and contributors

The studies included in this thesis have been made in collaboration with others. My individual contributions to these collaborations are specified below.

Study I Oertel, C., Cummins, F., Edlund, J., Wagner, P., & Campbell, N. (2013a). D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces, 7*(1-2), 19-28.

I carried out the gaze analysis and developed the annotation scheme for group involvement. I also developed the identification of acoustic cues to the individual contribution to group involvement. I collaborated with Céline on the analysis of mimicry and group involvement.

Study II Oertel, C., & Salvi, G. (2013). *A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue*. Paper presented at the ICMI 2013, Sydney, Australia.

I provided the research ideas and set the hypothesis. I also designed the corpus recordings and carried out the involvement as well as the gaze annotation. I wrote the vast majority of the paper. Giampiero carried out the machine learning and mathematical formulations.

StudyIII Oertel, C., Funes Mora, K. A., Gustafson, J., & Odobez, J.-M. (2015). *Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions*. Paper presented at the ICMI 2015, Seattle, USA.

I carried out the statistical analysis as well as the machine learning. I also wrote the majority of the paper and designed the study and hypothesis in collaboration with Jean-Marc and Joakim.

Study IV Oertel, C., Gustafson, J., & Black, A. W. (2016b). *Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances*. Paper presented at the Interspeech 2016, San Francisco, USA.

In designed the study in collaboration with Joakim and Alan. I did the perception test set up as well as the analysis. I carried out the machine learning and wrote the vast majority of paper

Study V Oertel, C., Lopes, J. D., Yu, Y., Mora, K. A. F., Gustafson, J., Black, A. W., & Odobez, J.-M. (2016c). *Towards building an attentive artificial listener: on the perception of attentiveness in audio-visual feedback tokens*. Paper presented at the ICMI 2016, Tokyo, Japan.

I designed the study in collaboration of Joakim, Alan and Jean-Marc.I generated the stimuli in collaboration with José.I carried out the perception test, conducted the analysis and machine learning and wrote the majority of the paper. I collaborated with Kenneth on the top-lesser token estimation. The headnod identification as well as the head nod feature extraction has been carried out by Yu Yu, Kenneth and Jean-Marc.

Study VI Oertel, Catharine, Joakim Gustafson, and Alan W Black. "On Data Driven Parametric Backchannel Synthesis for Expressing Attentiveness in Conversational Agents." In *Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*, 43-47. Tokyo, Japan, 2016.

I designed the study, carried out the analysis, provided the feature vectors for synthesis and wrote themajority of paper. In collaboration with Alan and Joakim I designed the resynthesis corpus. Alan carried out the synthesis.

Study VII Hjalmarsson, A., & Oertel, C. (2012). *Gaze direction as a back-channel inviting cue in dialogue*. Paper presented at the IVA 2012 workshop on Realtime Conversational Virtual Agents, Santa Cruz, USA.

Anna and I designed the study together and also worked very closely together in writing and analysis.

Study VIII Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication, 65*, 50-66.

I headed the uncertainty analysis and collaborated with the co-authors on the design of the study. I collaborated on the gaze annotations and co-authored all papers on which this journal paper is based on. I also designed the map stimuli

Study IX Oertel, C., Salvi, G., Götze, J., Edlund, J., Gustafson, J., & Heldner, M. (2013b). *The KTH Games Corpora: How to Catch a Werewolf*. Paper presented at the Multimodal Corpora 2013, Edingburgh, UK.

I headed the corpus recording as well as the design. I wrote the majority of the paper.

Study X Oertel, C., Mora, K. A. F., Sheikhi, S., Odobez, J.-M., & Gustafson, J. (2014). *Who Will Get the Grant?: A Multimodal Corpus for the Analysis of Conversational Behaviours in Group Interviews*. Paper presented at the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions.

I orginasied and carried out the recordings at KTH and initiated the collaboration with Idiap. Kenneth monitored the Kinect recordings remotely. I wrote the vast majority of the paper, gaze analysis was carried out in collaboration with Kenneth.

Non-included papers

The author has also written and co-authored other published papers, some of which are related to the contents of the present thesis. These are presented in the list below.

Related to the present thesis

Al Moubayed, S., Beskow, J., Bollepalli, B., Gustafson, J., Hussen-Abdelaziz, A., Johansson, M., . . . Varol, G. (2014). *Human-robot collaborative tutoring using multiparty multimodal spoken dialogue*. Paper presented at the 2014 ACM/IEEE international conference on Human-robot interaction, Bielefeld, Germany.

Al Moubayed, S., Beskow, J., Bollepalli, B., Hussen-Abdelaziz, A., Johansson, M., Koutsombogera, M., . . . Skantze, G. (2013). *Tutoring Robots*. Paper presented at the International Summer Workshop on Multimodal Interfaces, Bilbao, Spain.

Altmann, U., Oertel, C., & Campbell, N. (2012). Conversational involvement and synchronous nonverbal behaviour. In A. Esposito, A. Vinciarelli, R. Hoffmann, & V. Müller (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science* (pp. 343--352): Springer.

De Looze, C., Oertel, C., Rauzy, S., & Campbell, N. (2011). *Measuring dynamics of mimicry by means of prosodic cues in conversational speech*. Paper presented at the ICPhS 2011, Hong Kong, China.

Edlund, J., Oertel, C., & Gustafson, J. (2012). *Investigating negotiation for load-time in the GetHomeSafe project*. Paper presented at the Workshop on Innovation and Applications in Speech Technology (IAST), Dublin, Ireland.

Han, J. G., Dalton, J., Vaughan, B., Oertel, C., Dougherty, C., De Looze, C., & Campbell, N. (2011). *Collecting multi-modal data of human-robot interaction*. Paper presented at the Cognitive Infocommunications (CogInfoCom), Budapest, Hungary.

Koutsombogera, M., Al Moubayed, S., Bollepalli, B., Hussen, A., Oertel, C., Stefanov, K., & Varol, G. (2014). *The Tutorbot Corpus—A Corpus for Studying Tutoring Behaviour in Multiparty Face-to-Face Spoken Dialogue*. Paper presented at the LREC 2014.

Oertel, C. (2013). *Towards developing a model for group involvement and individual engagement*. Paper presented at the Doctoral Consortium at ICMI 2013, Syndney, Australia.

Oertel, C., Cummins, F., Campbell, N., Edlund, J., & Wagner, P. (2010). *D64: A Corpus of Richly Recorded Conversational Interaction*. Paper presented at the Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, Valetta, Malta.

Oertel, C., Looze, C. D., Vaughan, B., Gilmartin, E., & Wagner, P. (2011a). *Using Hotspots as a Novel Method for Accessing Key Events in a Large Multi-Modal Corpus*. Paper presented at the New Tools and Methods for Very-Large-Scale Phonetics Research, Pennsylvania, USA.

Oertel, C., Scherer, S., & Campbell, N. (2011b). *On the use of multimodal cues for the prediction of involvement in spontaneous conversation*. Paper presented at the Interspeech, Florence, Italy.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012a). *Gaze Patterns in Turn-Taking*. Paper presented at the Interspeech 2012, Portland, USA.

Oertel, C., Wlodarczak, M., Tarasov, A., Campbell, N., & Wagner, P. (2012b). *Context cues for classification of competitive and collaborative overlaps*. Paper presented at the Speech Prosody, Shanghai, China.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2013a). *Exploring the effects of gaze and pauses in situated human-robot interaction*. Paper presented at the 14th Annual Meeting of Special Interest Group on Doscourse and Dialogue-SIGDial, Metz, France.

Skantze, G., Oertel, C., & Hjalmarsson, A. (2013b). *User Feedback in human-robot interaction: Prosody, gaze and timing*. Paper presented at the Interspeech 2013, Lyon, France.

Skantze, G., Oertel, C., & Hjalmarsson, A. (2014). *User feedback in human-robot dialogue: Task progression and uncertainty*. Paper presented at the HRI Workshop on Timing in Human-Robot Interaction, Bielefeld, Germany.

Ward, N. G., Werner, S. D., Novick, D. G., Shriberg, E., Oertel, C., Morency, L.-P., & Kawahara, T. (2013). *The Similar Segments in Social Speech Task*. Paper presented at the MediaEval, Barcelona, Spain.

Weilhammer, K., Oertel, C., Siegemund, R., Sá, R., Batliner, A., Hönig, F., & Nöth, E. (2009). *A Spoken Dialog System for Learners of English*. Paper presented at the International Workshop on Speech and Language Technology in Education, University of Birmingham, UK.

Windmann, A., Wagner, P., Tamburini, F., Arnold, D., & Oertel, C. (2010). *Automatic prominence annotation of a German speech synthesis corpus: towards prominence-based prosody generation for unit selection synthesis*. Paper presented at the SSW 2010, Kyoto, Japan.

Part I. Background

1. Introduction

Human-human communication is characterised by our ability to read and respond to each other's opinions, attitudes, and interest level. In fact, learning to understand the dynamics of interpersonal communication is quite a daunting task, and yet most of us acquire it with ease while growing up. Our responses are to a large degree based on our interpretation of the verbal content of an utterance. However, nonverbal cues, such as smiles, head nods, and the direction of eye-gaze, are of great importance too. Many of these responses, particularly those that react to the nonverbal cues, happen rapidly and unconsciously. Caught up in a conversation, humans do not typically try to decipher prior to acting what it was exactly that triggered their instinctive reaction.

Once we have mastered this skill, it becomes quite an intrinsic part of our everyday conversational dances. Going against this acquired sensitivity would not only require a conscious effort, it would also be considered impolite and perhaps even upsetting by conversational partners. As listeners, it is our role to indicate that we are following the conversation and that we are paying attention to the speaker. If we feel bored or if we would like to leave the conversation, it is appropriate to indicate this with subtle cues rather than with speaking out, which would end a conversation abruptly. As a speaker, it is our role to pay attention to the listener in order not to make the conversation longer than the listener wishes it to be.

An everyday example of our capability to detect whether people are following the conversation may be demonstrated by using Figure 1, in which a group of people is sitting around a table. For us humans, it is relatively easy to guess that Person A is paying attention the speaker, Person B is distracted by focusing on his paper, Person C and D are engaged in a side conversation, and Person E is listening but is not directly involved. Person F might have been the last addressee, and Person G is the current speaker.

Figure 1: A group of people sitting around a table.

In recent years, there has been a growing interest in the analysis of dyadic and multi-party communication. The main goals of these analyses, were to design computational models that could recognize non-verbal behaviour, examples of which are the identification of roles within a group (Beyan et al., 2016; Sanchez-Cortes et al., 2013), the recognition of personality (Aran & Gatica-Perez, 2013; Cabrera-Quiros et al., 2016), the understanding of first impressions (Nguyen et al., 2014), and the understanding personal relationships (Zhao et al., 2016).

Most of these studies, including this thesis, can be classified under the header of *Social signal processing (Pantic et al., 2011; Vinciarelli et al., 2009). Social signal processing* is "the new research and technological domain that aims at providing computers with the ability to sense and understand human social signals" (Vinciarelli et al., 2009). *Social signal processing* defines several challenges for the machine analysis of social signals. Among these are: the recording of the scene, the automatic detection of people within a given scene, the extraction of audio-visual cues, their behavioural analysis, and their classification within a given context. These challenges will also be discussed in this thesis.

1.1 Motivation

The work described in this thesis is the result of my initial curiosity and puzzlement when trying to understand group dynamics. I asked myself what it is that makes people understand whether someone is willing to continue their participation in a conversation or not, especially when there are more than two people involved in the interaction. This fascination, paired with an interest in developing new technologies especially for the use in human-robot interactions, lead to a gradual shift from very open, in-the-wild data-collection scenarios to very meticulous and controlled perception experiments over the course of this thesis.

During the process of analysis, it became more and more apparent that, if aiming for modelling engagement, it was essential to focus on the listener. Even though inferring information about the listener is much harder than inferring information about the speaker, it is the listener who is most important in signalling whether a conversation is reaching an end or whether it continue.

Hence, the common thread running through this thesis is the perception of group involvement, and the participation of individuals within a group. The question is asked how differences in audio-visual cues can help explain differences in perception and how their generation in a virtual agent or robot can aid the development of more engagement-aware human-computer interaction.

1.2 Research questions

While research questions concerning human behaviour in groups are manifold, this thesis will focus on the concept of group involvement, individual engagement, and listener types. The exact definition of these terms is rather complex and will thus be discussed in detail in Chapter 2.1. This thesis will also consider the attention of people towards each other in a conversational context, and it will investigate how the models developed in human-human interaction can be used for human-robot interaction. Culture can be an important factor influencing group interactions; however, its explicit analysis lies outside the scope of this thesis. Similarly also the analysis of dominance (Charfuelan et al., 2010; Frauendorfer et al., 2014; Sanchez-Cortes et al., 2013) and rapport (Gratch et al., 2006; L. Huang et al., 2011; Zhao et al., 2014), while related, lies outside the scope of this thesis.

The work described in this thesis explores the following research questions:

A: ***What are the audio-visual cues to group involvement and individual engagement?***

1. In which way should a corpus recording be designed in order for it to capture as high variance in conversational dynamics?
2. Is group involvement a binary or scalar phenomenon?
3. Is it related to prosodic mimicry and is the fusion of audio-visual cues beneficial to its prediction?
4. Is it possible to define eye gaze features that describe individual engagement as well as group involvement? Are these features useful for automatic prediction of engagement and involvement?

B: ***How do human listeners signal their degree of engagement?***

1. Are there listener-type-specific differences in audio-visual behavioural patterns?
2. And if there are differences, are they perceived differently in terms of attentiveness, focusing solely on listener feedback tokens? Are there backchannels types that are perceived to be more attentive than others?

C: Can we model an artificial listener that can display different degrees of attention?

1. Is it possible to provide a speech synthesizer with an explicit control of the degree of attentiveness?
2. Can statistical parametric speech synthesis be used to validate prosodic models of perceived attentiveness?
3. Is it possible to affect the amount of listener feedback in human-robot interaction by controlling for the gaze direction of the robot?
4. How is uncertainty in human-robot interaction expressed in the feedback utterances of the listener?

The answer to these questions, gained through the studies included in this thesis, will be presented in the following chapters.

1.3 Thesis outline

1.3.1 Included papers

Study I

The studies summarised here are based on the D64 corpus. The D64 corpus is a richly recorded corpus of group interactions "in the wild". All the studies comprised in this journal paper share a concern for the dynamics of interpersonal communication. In particular, they are concerned the study of group involvement as well as mimicry and their non-verbal correlates. These non-verbal correlates include gaze, blinking, and prosodic cues such as pitch range and intensity. Prediction experiments are carried out in order to investigate which modality best predicts group involvement. Moreover, a methodology is proposed that is able to describe the dynamics of mimicry.

Study II.

This study is based on the KTH werewolf corpus. The aim of this study is to define those gaze features that make it possible to describe both group involvement as well as conversational engagement. The features that are proposed here are presence, entropy, symmetry, and MaxGaze. Individual engagement is assessed by the participants of the study themselves, whereas group involvement is assessed by third party observers. Descriptive statistics as well as machine-learning methods are employed in order to investigate the usability of the proposed gaze features.

Study III

This study is based on the KTH Idiap Group Interviewing Corpus. It aims at discovering patterns in non-verbal behaviour (e.g. gaze and, visual and acoustic backchannel) of different listener categories, such as attentive listeners, side participants, and bystanders. A further aim is to understand the relationship between individual engagement and the different listener categories as well as to build a classifier which tests how well the different listener categories may be classified automatically.

Study IV

This study, too, is based on the KTH Idiap Group Interviewing Corpus. Its objective is to investigate backchannels in terms of their perceived attentiveness. Furthermore, it wishes to understand how backchannels that differ prosodically need to be realised for people to perceive the difference. In order to achieve this goal, a perception test is set up and the backchannels are analysed prosodically. Having identified the relevant prosodic features, we apply ordinal machine learning techniques in order to test whether attentiveness in backchannels could be predicted automatically.

Study V

This study is a follow-up of Study IV. While Study IV investigated acoustic backchannels in terms of their perceived attentiveness, the current study investigates the perceived attentiveness of head nods. It also examines how a fusion of both modalities is perceived by third-party observers. The head nods are generated based on visual information extracted from the KTH Idiap corpus recordings. They are, therefore, quite close to human head nods. In order to test third-party observer impressions, crowd-sourced perception tests are set up, and the head nods as well as the acoustic backchannels are realized in a virtual agent. Ordinal machine learning is carried out in order to test how well attentiveness can be predicted in head nods as well.

Study VI

This study is building on study IV in that it tries to use study IV's findings and use it for the synthesis of backchannels. The first step was the creation of a parametric speech synthesiser based on re-enacted recordings of the KTH Idiap group interviewing corpus. In a second step, a crowd-sourced perception test is set up in order to investigate whether synthesised backchannels are perceived in a similar fashion to the human produced backchannels. Also, a perception test is used to examine whether it is possible to generate backchannels that are perceived as more or less attentive than those from KTH Idiap corpus.

Study VII

This study investigates to what extend the gaze direction of a virtual agent can influence the feedback behaviour of its human interlocutor. For this investigation, a study is designed in which a virtual agent is given a cooking recipe and the human is asked to provide feedback. The gaze direction of the virtual agent during the narration is controlled for based on the hypothesis is that the human interlocutor will provide more feedback when the gaze of the virtual agent is directed towards him or her.

Study VIII

This study explores human-robot interaction in the context of a map-task scenario. A robot is providing direction to find a given landmark and the human interlocutor has to sign in the route. Most relevant to the present discussion is analysis of uncertainty in feedback realisations of the human interlocutor. The question is asked: is the lexical form of feedback tokens distributed differently in the tokens that are perceived to convey certainty and those that are perceived to convey uncertainty? Moreover, an acoustic analysis is carried out to identify the acoustic features that distinguish certain from uncertain feedback tokens. Finally, a classifier is trained in order to investigate the prediction performance of the features.

Study IX

This study describes the Stockholm Werewolf Corpus. The Stockholm Werewolf corpus is a multi-party interaction corpus which is situated in the context of the role-playing game *Die Wärwölfe von Düsterwald* ®. Particular focus is given to the positioning of the cameras, the seating of the participants, and the timing of the distribution of the questionnaires. This enables us to assess the group's involvement and the individual participants' engagement in the conversation. The corpus is annotated both for gaze and group involvement.

Study X

This paper describes the KTH Idiap Group Interviewing Corpus. The Corpus is designed to reduce manual annotation to a bare minimum, which was accomplished by making the automatic extraction of audio-visual cues as easy as possible. Furthermore, it is designed to maximize variation in conversational dynamics. In order to illustrate the potential of this corpus in terms of non-verbal analysis, some preliminary gaze transition patterns are also included.

2. Individual Engagement and Group Involvement

2.1 Definition

Quantifying a person's level of engagement in a conversation and estimating the involvement of the whole group of people is not a trivial task to accomplish. Since the mid 20th century, a growing number of researchers have started to investigate this topic. The first issue they had to tackle was the conceptualisation of the terms involvement and engagement themselves.

There are a lot of terms used to describe the same or a similar concept. Such terms include: "interest", "arousal", "attention", "conversational involvement", "face engagement", "group interest", "group engagement". The current section will provide an overview over definitions and how they coincide or differ from the use of *group involvement* and *individual engagement*, as used in this thesis.

Peters et al. (Peters et al., 2009) summarise that despite there being differences in the use of engagement related concepts there still appear to be two fundamentals which describe engagement; the first one is attentional- and the second one is emotional involvement. The following section attempts to provide an overview of the different uses of engagement and involvement.

Coker and Burgoon (Coker & Burgoon, 1987) attempted to form one of the first conceptualisations. They described conversational involvement as consisting of five distinct variables: "the degree of animation and dynamism"," the tendency to be interested in, attentive to, and adaptive to the other in a conversation", the "immediacy" in the behaviour of the interlocutors, and their degree of "social anxiety". Goffman (Goffman, 1966)defined face engagement as follows:

> "Face engagements comprise all those instances of two or more participants in a situation joining each other openly in maintaining a single focus of cognitive and visual attention – what is sensed as a single mutual activity, entailing preferential communication rights."

Another way to describe different levels of engagement in a group of people was proposed by Clark (Clark, 1996), who build on Goffman (Goffman, 1967). He refers to people within an interaction as belonging to different participation categories. To make this classification, he first distinguishes between participants and non-participants. The group of participants he considers to consist of "the speaker", "the current addressee" and "the sideparticipant". The group of non-participants includes the categories of "bystanders" and "overhearers". Sidner et al.(Candace L. Sidner et al., 2005), in contrast, define engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". Gatica-Perez, Mccowan et al. (Gatica-Perez et al., 2005) define group interest as "the perceived degree of interest or involvement of the majority of the group". Salam, Celiktutan et al. (Salam et al., 2016) defined "group engagement" as "the engagement state of two entities in the interaction together with another entity". Their definition is based on the assessment their annotators made for the following statements: "The person to my left/right is engaged/involved in the interaction with the robot", "is bored", "is interested by what the robot is saying", "likes the interaction". Oertel, Scherer et al. (Oertel et al., 2011b) define group involvement as "a group variable which is calculated as the average of the degree to which individual people in a group are engaged in spontaneous, non-task-directed conversations".

In their definitions, Goffman (Goffman, 1966) and Sidner (Candace L. Sidner et al., 2005) describe engagement as a process. They encompass the action of getting engaged as part of the construct of engagement itself. Our definition and treatment of the construct of engagement is different from these approaches in that we restrict engagement to the context of an already on-going conversation.

Goffman (Goffman, 1966), moreover, emphasises both the importance of the single cognitive as well as the importance of the visual focus of attention. In our work, we take into consideration more than a single focus of attention. In fact, we define four different gaze variables that are designed to summarise different aspects of engagement (cf. Study II).

Coker and Burgoon (Coker & Burgoon, 1987) and Poggi (Poggi, 2007) appear to stress the psychological aspect of engagement. Our work is probably closest to their work in this respect.

Finally, similar to Gatica-Perez, Mccowan et al. (Gatica-Perez et al., 2005), we see group involvement as being the average of the individual participant's engagement. It has to be noted here that this does not entail annotating each participant separately before calculating the average. Group involvement is annotated as a variable in its own right without explicitly taking the engagement of the individual participants into account. In this regard, our work is different from Salam, Celiktutan et al. (Salam et al., 2016) who did not annotate group involvement as a variable in its own right.

2.2 Multi-Modal Corpora for Group Interactions

Over the last decade, more and more corpora have been created, such as the ones described in, (Rehm & Nakano, 2008) (Lücking et al., 2013), (Traum et al., 2012), (Sun et al., 2011) (Chen et al., 2005) and (Mana et al., 2007). They were often created with very specific tasks in mind, and often contain scripted or task-directed speech.

The following section will review existing corpora that are most relevant to this thesis, with some comment on how the design choices might influence engagement modelling using audio-visual cues.

All of the following corpora are multi-party corpora. They vary, however, in their design and set-up. For example, the AMI (Carletta, 2007) meeting corpus and the CHIL corpus (Mostefa et al., 2007) are meeting corpora. In both corpus collections whiteboards present in the room, and participants were taking notes.

The Idiap Wolf Database (Hung & Chittaranjan, 2010) involved subjects that were playing a role-playing game called *Wärwölfe von Düsterwald* ®. In this recording the participants were seated in a half circle when playing the game. Moreover, as the corpus was not designed with the study of group involvement and individual engagement in mind no questionnaire was administered during the recordings. Such a questionnaire, could have captured participants' impressions of their own engagement as well as each other's engagement.

In (Kawahara et al., 2013), Kawahara et al. recorded a corpus to investigate the level of interest and comprehension in scientific poster presentations. They recorded data in which one researcher would present his or her research interest to two further people. Each presentation, which typically lasted between 20-30 minutes, was divided into four or eight separate components and was followed by an overall question/answer phase.

Figure 2 The D64 recording setup

The D64 corpus (Oertel et al., 2013a) is the corpus which forms the basis for Study I**.** It is a richly recorded corpus of approximately two days of interactions. Four to five participants met in an apartment in Dublin which was equipped with all kinds of sensors. They talked about any topic which sprang to mind over the course of these two days. While the D64 corpus allowed for insights into group dynamics due to the openness of the setting, it also had several disadvantages. Both the light setting and the seating arrangement of the participants were not ideal for the manual annotation of eye-gaze. They also made an automatic estimation of eye-gaze impossible. Additionally, the acoustic settings were not always optimal. In some instances, the dishwasher was running or the kettle was boiling, which made it impossible to extract and carry out prosodic analysis for certain sections of the corpus. One further disadvantage, at least for the study of individual engagement, was the difference in hierarchy between participants. The group consisted of several professors and two master's students, which might have influenced the conversational dynamics. Finally, the conversation was not designed to contain a rich set of changes in conversational dynamics over a short period of time. This was both a disadvantage and an advantage. It was a disadvantage since the corpus data needed more annotations than a shorter, more dynamic corpus in order to achieve similar results in terms of individual engagement and group involvement. It was an advantage, as the corpus displayed conversational dynamics as they unfold in every-day conversations.

The Tutorbot corpus (Koutsombogera et al., 2014) investigates multi-party conversations in a game scenario. Here, the focus lies with a tutoring scenario. It is the task of one of the participants to play the role of a supportive tutor or neutral tutor. Participants had to jointly sort cards that were laid out in front of them.

(Moubayed et al., 2013) base their corpus on their design of a three-party quiz-scenario. The setup used is a small round table with three Tobii gaze trackers, three web cameras and a Microcone microphone array with 6 directional microphones. The participants did not have any objects on the table that was part of the interactional task.

This thesis is concerned with the modelling of group involvement and individual engagement using eye-gaze. However, the presence of objects, such as whiteboard, cards, and sheets of paper, will alter people's gaze-pattern distributions, as was found for example by (Oertel et al., 2013a) and (Johansson et al., 2013) for head pose patterns in human –robot interaction. For example, it will be more difficult to assess whether averting the eye-gaze downwards is caused by disengagement or is simply the result of the participant gazing at a sheet of paper. While the inclusion of objects, such as whiteboards, cards, and posters can be very useful for many scenarios, it is a complicating factor for the purpose of this thesis.

Given the review of existing corpora, we decided to set up a number of requirements in order to answer the research questions pursued in the current thesis work. These corpus requirements are listed below:

1) It should contain multi-party conversations of at least three people but preferably more
2) The corpus should comprise spontaneous, non-scripted conversations
3) It should be rich in conversational dynamics
4) It should be recorded with a rich set of sensors, as this will allow for more scalable and easier post processing
5) There should be no eye-gaze distraction by objects in the room or by unequal participant-seating distributions

We first designed the Stockholm Werewolf corpus (Oertel et al., 2013b). The Stockholm Werewolf corpus is a multi-party interaction corpus which is situated in the context of the role-playing game *Die Wärwölfe von Düsterwald* ®. Particular focus is given to the positioning of the cameras, the seating of the participants, and the timing of the distribution of the questionnaires. This enabled us to assess the group's involvement and the individual participants' engagement in the conversation. The corpus was annotated both for gaze and group involvement.

Figure 3: KTH Warewold corpus

In (Oertel et al., 2014), we decided to use a scenario that was somewhere in between a game and real-life: a group interview. The exact details of the corpus set-up can be found in Study X.

The scenario was close enough to a real life experience for the participants: it was something they could identify with. Moreover, we included four distinct phases. In the first one, the participants had to introduce themselves. We assumed that this was the least cognitively changing task. Subsequently, the participants had to give an elevator pitch about their respective PhD projects and explain the impact on society. Finally, they were asked to come up with a joint PhD proposal. In many of these phases, the participants had to listen to the other participants and had to contribute to their contributions. These phases

allowed for the later analysis of four distinct conversation dynamic sections. Moreover, we kept the internal evaluation paradigm, but we amended it: this time, the participants themselves did not do their own evaluations; rather, the moderator of the group interview carried out the evaluations after the interview had finished.

Finally, the KTH Idiap Group Interviewing Corpus was designed to reduce manual annotation to a bare minimum, which was accomplished by making the automatic extraction of audio-visual cues as easy as possible. Furthermore, it is designed to maximize variation in conversational dynamics.

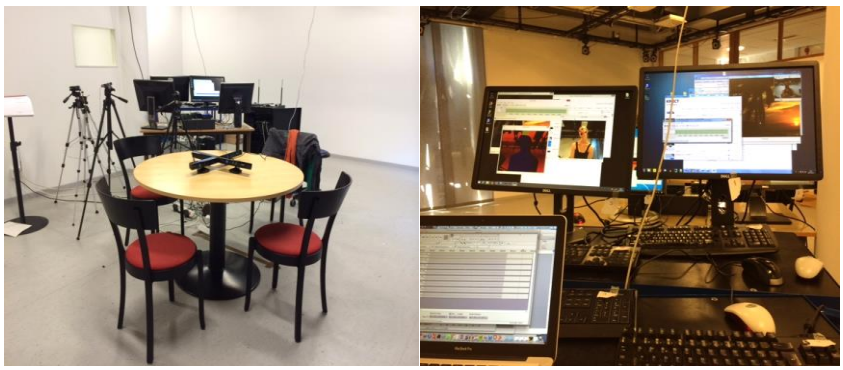

Figure 4: KTH-Idiap Group Interviewing Corpus Set-Up

After the Stockholm Wolfscorpus (Oertel et al., 2013b) and the KTH Idiap Corpus (Oertel et al., 2014) were recorded, several other corpora were created with the same scope as these two. Later examples of multi-party corpus collections include (Litman et al., 2016), who designed the Teams Corpus in order to investigate entrainment in groups of 3-4 people in a games scenario.

A final example is (Stefanov & Beskow, 2016) who, in their corpus, study the focus of attention of groups of participants. For this, they recorded groups of three participants while they were sorting cards. They also recorded groups of three participants, discussing their travel experiences without any objects present that might distract the visual focus of attention.

2.3 Data Annotation

The following section will provide an overview of the different approaches that are available to annotate engagement. There is quite a wide variety of ways in which engagement annotations are approached. For example, the instructions provided to annotators may already vary much per study. There are some that provide relatively detailed instructions for separating various contributing factors of engagement (Salam et al., 2016), while others depend more on the intuition of the annotators (Bonin et al., 2012). Most studies appear to rate involvement directly, but at least one study receives involvement annotations indirectly by using reactive tokens and question types(Kawahara et al., 2013). In the majority of cases, no explicit distinction is made between speakers and listeners. There are some exceptions, however, such as (Levitski et al., 2012), who explicitly investigate the role of the silent participant. Furthermore, (Yu et al., 2016), designed their annotation scheme on the basis of the users' completion of the whole turn, thus implicitly annotating speaker engagement. The window size for a given annotation also varies notably: it ranges from 5 seconds in (Kim et al., 2016), to turns in (Yu et al., 2016), and minutes in (Kawahara et al., 2013). The granularity in which engagement is annotated ranges from a binary distinction (Levitski et al., 2012) to a 10-point scale (Salam et al., 2016). Most studies, however, choose a scale between four to six points. With the exception of (Kim et al., 2016), all other annotation schemes annotate engagement as a continuous rather than an ordinal variable. The number of annotators for a given segment also varies widely, yet all studies use at least two annotators. Finally, with the exception of (Gatica-Perez et al., 2005),(Salam et al., 2016), and (Lai et al., 2013), all studies treat engagement as an individual variable rather than a group variable.

For the studies included in this thesis, it should be noted that different annotation schemes have been used. The studies included in Study I are based on the relatively detailed annotation scheme proposed in the same study (10 points with explicit instructions). In Study II, we used an annotation scheme that was quite different from the one proposed in Study I. This was decided for the following reasons: first of all, the time window of five seconds that was used in Study I seemed to have been quite small. This means that the involvement state did not change over quite a number of windows. It was therefore decided not to maintain the five-second annotation window for the next study. Secondly, the conversational dynamics in Study II were quite different from the ones observed in Study I, since the participants were assigned predefined roles dependent on the game itself. Consequently, it was decided to use to use the annotation scheme as proposed in Study II**.** In Study III, we combined the categorical approach used in Study II with the scalar approach used in Study I**.** In particular, we used the scalar approach for engagement annotation (through a seven-point scale), as we noticed in Study I that the 10 points were not used by the participants. We also used the categorical approach for the definition of listener categories.

In general, the advantage of event-based measurements is that they aim to point towards the exact change in social dynamics. The disadvantage of such an approach, however, is that it is difficult for people to agree on the exact point in time when such a change occurs, since people assign different weight to different phenomena. The advantage of an interval-based approach is that it is less prone to annotator discrepancies. As the boundaries of the segment are predefined, annotators do not need to agree on the exact moment of change in the dynamics: they only need to average the perception of individual engagement or involvement over the given interval. Averaging, of course, has the disadvantage that some details may be lost. This may be counteracted by choosing smaller intervals.

2.4 Concluding Remarks

The previous section provided an overview of the definitions, data, and annotations that form the basis of this thesis. Existing definitions of individual engagement and group involvement have been discussed, and their relevance for the studies comprised in this thesis has been explained. The previous section has also compared existing approaches to the annotations of engagement and involvement. The choices that were made with regard to these annotations have been justified. Moreover, corpora in the area of multi-party, multi-modal interactions have been compared and the need for new corpus recordings has been justified. The following chapter will discuss nonverbal cues and their relevance to the interpretation of the dynamics of human communication.

3. An Overview of Relevant Nonverbal Cues

This chapter will provide an overview of relevant nonverbal cues for the analysis of group involvement and individual engagement. Special focus will be given to results relating to eye-gaze, prosody, backchannel, and head nods. Backchannels and head nods will be of particular importance when analysing and generating listener behaviours for human-robot interaction.

3.1 Gaze

Paying attention eye gaze is very important when trying to decipher human behaviour. Humans are able to use eye-gaze on purpose to signal information to their conversational partners. One example of this is signalling the focus of attention in order to establish a common point of reference. Another example is the purposeful aversion of gaze from one's conversational partner in order to signal that one is no longer interested in continuing the conversation. There are, however, also patterns that are much more unconscious, for example, those related to turn taking (Kendon, 1967),(Argyle & Graham, 1976),(Vertegaal et al., 2001) , or to a person's engagement in the conversation.

This thesis is concerned with the modelling of the second category, the more unconscious patterns. Most research on eye-gaze in conversational settings has been carried out on dyadic conversations. The following section tries to provide an overview of the current state of the art. It is important to consider environmental

factors when trying to model eye-gaze under varying contexts, as it is otherwise impossible to make any universal judgements (Peters et al., 2010). While there are some patterns that seem to hold more globally, it has, for example, been found that many gaze patterns vary substantially from one speaker pair to the next (Cummins, 2012). It has also been found that, gaze coordination is related to factors such as established common ground and mutual knowledge (Shockley et al., 2009).

Figure 5: Human and Robot looking at each other

Another interesting finding is that speaker and listener are different with regards to their gaze patterns. While the listener gazes at the speaker for long periods of time, the speaker gazes at the listener in short but comparatively frequent periods (Argyle & Cook, 1976). Moreover, (Janet Beavin Bavelas & Gerwing, 2011) found that the listener is more likely to produce feedback when a mutual gaze is established between a speaker and a listener (gaze window). Furthermore, they found that the speaker would avert his gaze more frequently during listener feedbacks and, as a result, would end the period of mutual gaze.

With regards to turn taking and gaze patterns, (Oertel et al., 2012) found that, for overlaps with speaker change, the previous speaker's partner-oriented gaze increased suddenly following the onset of the incoming speaker's turn. Moreover, they found that the extent to which incoming speakers tend to look away appears to be greater for speaker changes than for turn continuations. It was also greater, for overlap than for out-of-overlap constellations. In addition, silences and overlaps that were shorter than 130ms were similar to the silences with speaker change. Finally, they found that changes in gaze patterns appear to be observable well beyond the boundaries of the discourse phenomenon in question. Another study discovered that, in face-to-face conversation, mutual gaze occurs significantly more often during visual backchannels (Truong et al., 2011).

Recently (Zhao et al., 2016) found that mutual gaze is often established during self-disclosure and that listeners subsequently avert their gaze. When reference is made to a shared experience, listeners are also more likely to avert their gaze. This happens, when they are violating a social norm as well.

For multi-party conversations, (Bednarik et al., 2012) found that a low degree of engagement in their participants lead to long gaze durations towards the same interlocutors, while higher degrees of engagement coincided with participants gazing at the same target for shorter durations and, therefore, also gazing at more interlocutors

For human-avatar interaction, (Edlund & Beskow, 2009) found that human interlocutors were more prone to initiate a turn when the avatar's head and gaze were directed straight ahead, as opposed to when the head was turned slightly to the side and the gaze was shifted away from the listener. Another study showed that a robot that uses intentional gaze aversions is able to effectively manage the floor and is also perceived to be more thoughtful (Andrist et al., 2014). It was demonstrated by (Andrist et al., 2012) that a virtual agent employing affiliative gaze elicited more positive feelings of connection in the participants, while a virtual agent employing referential gaze improved participants' learning.

3.2 Prosody

The current section will provide an introduction to prosody, and it will explain why prosody is relevant to the study of individual engagement and group involvement. Prosody may be roughly explained as speech melody. It is related to intonation, stress and rhythm. It can also be a very important cue to signalling whether one is finished with one's turn or not (Beattie et al., 1982).

The main auditory features of prosody are *pitch*, *loudness*, *quantity* and *voice quality*. These features are closely related to the acoustic features *fundamental frequency* (F0), *intensity*, *duration* and *spectral characteristics*. When analysing pitch, it should be converted from Hertz to Semitones, since pitch is perceived by the human ear on a logarithmic scale. In general, when examining these features, speaker differences should be accounted for by normalization.

The same auditory features that are related to prosody are also known to be related to more global paralinguistic patterns, such as excitement in the speaker. For the studies included in this thesis, it is important to keep the prosodic background in mind for the analysis of backchannels. It is also important remain aware of the fact that prosodic variations such as rising pitch do not carry any meaning on their own. They do, add meaning, however, to the verbal content depending on the linguistic context (syllable, utterance, discourse or social). It has been found, for example, that the speaker invites the listener to give backchannels in conversations by raising the pitch and loudness at the very end of an utterance (Gravano & Hirschberg, 2009).

3.3 Backchannels

The following section will deal with research on backchannels. Although backchannels are quite short in duration and unobtrusive in their realization, they may carry crucial information about the listener's reaction to the speaker's speech. It may, for instance, indicate the listener's attention, feelings as well as his/her understanding of the

conversation. In short backchannels support the conversation (Yngve, 1970). Using backchannels inappropriately may have negative consequences for the dialogue has been shown by (Janet B. Bavelas et al., 2000). They investigated the role and effect of listeners in storytelling. For this, they designed two conditions. In the first condition, they distracted the listeners while in the second condition, they did not. They found that, in the distracted condition, the number of responses was lower and the listeners made less specific responses. Furthermore, they found that third-party observers evaluated the storytellers as having performed a story less well when they were in the distracted condition.

In general, it may be stated that backchannels constitute a very important part of human-human communication. Therefore, their modelling is also very important for human-robot interactions. In fact, when designing conversational agents or robots, it is very important to both get the timing of the backchannel token right and to also choose the correct feedback realisation. A failure to do so may result in a disadvantageous evaluation of the conversational system. Additionally, it may even result in an abortion of the dialogue. The correct modelling is of particular importance in systems that focus on the role of the listener rather than the role of the speaker. For example, a tutoring robot that evaluates the certainty of the pupil's listening feedback in order decide if it has to elaborate the explanation in order to be fully understood.

There are also some recent studies, however, that have concentrated on the form. (Bailly et al., 2016), for example, investigated the backchannels produced by a professional interviewer during a neuropsychological test. They demonstrated that the choice of lexical items and, their frequencies, as well as the prosodic contour of the backchannels were related to the interviewees' performance in the test and to the neuropsychological interviewing protocol.

(Gustafson & Neiberg, 2010)studied prosodic cues to engagement in backchannels produced by a radio host in 73 calls to a call-in show. The found non-lexical listener tokens with rising pitch functions as continuation and signal high interest, while those with

falling pitch functions as acknowledgement and signal lesser interest. For bisyllabic tokens they found that it was the pitch slope of the loudest syllable that decided which of these two interest levels the feedback signals.

(Kawahara et al., 2016) investigated whether choosing forms of backchannels that fit the dialogue context influences the perception of naturalness, empathy, and understanding. For this, they evaluated their model, which considers the dialogue context as well as the form of backchannel, against two baselines. For the first baseline, it is always the same backchannel that is chosen. For the second baseline, a random backchannel is always chosen. Their model performs better than both baseline models.

In (Yamaguchi et al., 2016) the morphological patterns of backchannels were examined. The researchers were especially interested in the ways in which the linguistic features of the preceding utterance relate to the subsequent backchannel form. These linguistic features included the boundary of the end of the previous utterance as well as the linguistic complexity of the previous utterance. They built a model in order to predict the morphological pattern of backchannels based on the previous context. Their model performed better than the baseline models.

3.4 Attitudes and Functions in Backchannels

While the previous section summarised research on backchannels in general, the following section will provide an overview of research that has dealt with attitudes and functions in backchannels in particular.

In (Allwood et al., 1990, 1992), feedback and its functions are discussed. They proposed that feedback may have four functions namely, ”contact”, “perception”, “understanding”, and “attitudinal response”. For this thesis, and therefore also for this background overview, the fourth function, “attitudinal response”, is of most interest.

The ways in which uncertainty is realised in speech have been documented in several studies (Liscombe et al., 2006; Pon-Barry, 2008). (Lai, 2010) investigated, however, how different intonation

contours in feedback tokens such ("yeah", "right", "really") influence listeners' perception of uncertainty. They found rather than uncertainty, backchannels with rising pitch signal to the speaker that the question under discussion is unresolved and that it is the speaker that has to fix it.

(Malisz et al., 2012) analysed the prosodic characteristics of the German "ja", "m", and "mhm". They focused on investigating differences in pragmatic functions, as well as differences in prosodic realisation between listeners who were distracted with another task and listeners who were not distracted. They found that listeners who were not distracted tended to speak more loudly and tended to have a less variable energy level. They also discovered that and the variability of pitch was higher when the level of attentiveness was higher. They argued, therefore, that prosodic features may be strongly dependent on segmental structure, e.g. nasality vs. orality and syllabic structure vs. monosyllabic structure.

In another study on the relationship between the prosodic realisation of feedback token was related to the perceived function, (Neiberg et al., 2013) found that feedback token were often were multi–functional. The perceived functions included understanding, agreement, certainty, and negative surprise. They found that perceived interest and understanding in feedback token was related to a fast speaking rate and a moderate F0 rise.

(Pammi & Schröder, 2009) found that if you combine certain segmental forms and intonation the resulting backchannels gets un unpredicted meaning that is different that their individual meanings.

3.5 Visual Backchannels

The following, final section of this chapter will try to provide an overview of research into visual backchannels. Visual backchannels include tilting the head, nodding, and smiling.

In a similar fashion to the research mapping the form and function of acoustic backchannels, there have been endeavours to map visual backchannels. As is the case for acoustic backchannels, there is also an interest in whether the combination of several visual feedbacks

would change the perceived meaning. (Heylen et al., 2007) for example, found evidence that there are some visual feedback tokens that are convincing by themselves to describe a given function such as "nod"- or "agree". They also found that some visual cues are better used in combination describing a given function, such as the combination of "tilt" and "frown" for disbelieve.
In (Heylen, 2008), it was pointed out that the "level of excitement and the positive versus negative attitude (argumentative, defensive, etc.) towards the other are important variables to work with in future".

3.6 Concluding remarks

Backchannels may be realised in different modalities and may convey different meanings. (Duncan, 1974), building on (Yngve, 1970), has listed several possibilities including "head nods", "requests for clarification", "sentence completion", and "m-hm". According to Duncan this latter category stands for a whole group of "verbalized signals" containing items such as "yeah", "right" etc.

In this thesis, however, a differentiation is made between feedback token and backchannels. While the term feedback token comprises both backchannels and feedback tokens, the term backchannel is reserved specifically for only non-lexical realisations such as "mhm". Lexical realisations such as "yeah" and "okay" are classified here as belonging to the group of feedback tokens. This chapter has reviewed the literature for both feedback tokens and backchannels, the findings for feedback tokens are relevant for the included studies, but the studies themselves are only concerned with the category of backchannels.

4. Automatic Modelling

4.1 Applications for Individual Engagement and Group Involvement

The following chapter deals with an overview of work that has been carried out in the field of group involvement and individual engagement. The current section will introduce the reader into the topic by providing an overview over possible applications the automatic modelling engagement and involvement could have. While the detection of group involvement and individual engagement may in itself provide insights into behavioural patterns of humans during interactions, it may also have very practical implications for systems that are designed to improve communication between people.
Examples of such applications include systems that, on the one hand, provide feedback to the individual about his/her communication capabilities- and, on the other hand, provide more global feedback on the behavioural patterns of the group as a whole.

Feedback to the individual may include feedback on the impact a speaker has on his/her audience while speaking publicly. For example (Curtis et al., 2015) investigated the effect of public speaking on the engagement of the audience. Using for example their results on the relation to audio-visual could give insights into points in time in which the speaker managed to capture the interest of the audience as well as highlight periods in time in which the speaker lost the attention of the audience.

A system like this could be useful to the individual, both in an online and in an offline mode. In an online mode, it could be used as a training device. A person could practise his/her talk before delivering it in real life and could, in this way, get real-time feedback. This feedback may inform the speaker at which points in time he/she should perhaps modulate his/her voice or use more or less gestures. Additionally, a tool like this could potentially help to reduce fear of public speaking. Such a system has, for example, been introduced by (Chollet et al., 2016; Hincks & Edlund, 2009). A study on the communication capabilities of participants has been carried out by (Rasipuram et al., 2016) and a study focusing on the behaviour of teaching assistants and the effect their behaviour has on class participation has been carried out by (Gerritsen et al., 2015).

Feedback to a group as a whole might be useful in an organisational setting. Most people holding an office job, for example, attend meetings, sometimes up to times a day. While the purpose of these meetings is to increase productivity, this is not always the result. Sometimes communication between participants is simply not optimal. A system that is able to provide feedback during the meeting itself has the potential to improve the conversation while it is ongoing by, for instance, indicating to a participant that he is talking too much or that another person should be included more. In order to do this efficiently, it is important to consider through which modality the feedback should be provided, as for example has been investigated by (Damian et al., 2016). Moreover, if these meetings are recorded, a system could post-process the data as well in order to provide information about the engagement of the individual and the involvement of the group as a whole. Such post-processing would, in addition to short-term views, also provide long-term views. Research that has explored methods to investigate the productivity and happiness of employees in a company or organizational setting has been proposed by (Finnerty et al., 2014; Mashadi et al., 2016; Olguin et al., 2009). In these studies, it was suggested that participants' location may be tracked through the use of badges.

However, (Mashadi et al., 2016) have pointed out that the participants stressed their need and desire for privacy. All of the employees expressed that they would game the system if information about their interactions would be forwarded to their bosses or managers. A further potential application could be engagement detection for use in smart meeting rooms such as the one described in(Moore, 2002).
The following sections will provide an overview of the work that has been done on individual engagement and group involvement.

4.2 Group Involvement

The current section aims to provide an overview of research carried out on the automatic engagement, and/or involvement, estimation in groups. Here we provide an overview of studies that used different terms, but are referring to the same, or a very similar concept as we are in this thesis. Such terms are for example "group interest", "interest", "engagement", "individual engagement" and "group engagement". Most of the research has been carried out on human-human interactions: (Altmann et al., 2012; Gatica-Perez et al., 2005; Kawahara et al., 2013; Kim et al., 2016; Oertel & Salvi, 2013; Oertel et al., 2011a). (Salam et al., 2016) is the exception to this, since the perception of individual engagement and group engagement was studied here in interaction with a robot.

In human-human studies, a further division can be made between those studies that focus on interactions amongst adults (Gatica-Perez et al., 2005; Kawahara et al., 2013; Oertel et al., 2013a; Oertel & Salvi, 2013) and those that focus on the interaction amongst children (Kim et al., 2016). Additionally, there are different approaches used to investigate engagement in terms of group interaction. While (Gatica-Perez et al., 2005), (Oertel et al., 2013a; Oertel & Salvi, 2013) and (Salam et al., 2016) examined group interest/engagement, (Kawahara et al., 2013) and (Kim et al., 2016) investigated the individual within the group but did not consider the group as a whole separate variable. In addition, the studies also differ in whether they consider engagement a binary or a multi-level phenomenon. While (Gatica-

Perez et al., 2005; Kawahara et al., 2013), for example, treated it for prediction purposes as a binary phenomenon, (Oertel et al., 2013a; Oertel & Salvi, 2013; Salam et al., 2016) investigated it as a multilevel phenomenon. The features that were used seemed versatile and differed per study. They can be grouped, however, into several categories: prosodic features, such as energy, F0, jitter and shimmer (Gatica-Perez et al., 2005; Kim et al., 2016; Oertel et al., 2013a; Oertel & Salvi, 2013); turn-taking features, such as speaker change or overlaps (Kim et al., 2016; Lai et al., 2013; Laskowski, 2008); motion features, such as global person motion or body posture (Gatica-Perez et al., 2005; Oertel et al., 2013a; Salam et al., 2016); and eye-gaze(Kawahara et al., 2013; Oertel et al., 2013a; Oertel & Salvi, 2013). In one case, lexical features were used too (Kawahara et al., 2013). Furthermore, the machine-learning approaches which were used were also quite versatile: they ranged from HMMs (Gatica-Perez et al., 2005) to ordinal learning approaches (Kim et al., 2016).
(Laskowski, 2008) used low-level vocal activity features in order to automatically detect *hotspots,* the (parts of a conversation where participants are more involved). He found that laughter is a very valuable feature for the prediction of hotspots. His system achieved an accuracy of 84%.

This thesis has brought the following contributions to the study of group involvement and audio-visual features. First of all, up to the date of submission of the papers included in study I, no one had looked at involvement as being a scalar rather than a binary phenomenon. Secondly, there had not been any investigation examining whether the fusion of modalities (eye-gaze + blinking with prosodic features) would increase the prediction accuracy of involvement. Thirdly, no one, to our knowledge, had quantified the contribution of eye-gaze to group involvement before. Up to the submission of Study II's no one, to our knowledge, had defined and quantified gaze variables as well as investigated individual engagement with respect to the role of eye-gaze in group involvement.

4.3 Individual Engagement

The current section focuses on studies that are concerned with the individual engagement of a participant both in a dyadic context and in a group. It has to be noted that in some instances the term "interest" is used in a similar fashion to the notion of engagement as defined in this thesis. Therefore, these studies have been included too.

(Schuller et al., 2009), for instance, investigated the notion of spontaneous interest in a dyadic conversational context. In their study, the subjects were explicitly instructed not to worry about being polite. The participants were also instructed to feign interest. Schuller, Müller et al. distinguished between 5 levels of interest and tested several machine-learning techniques. In the end, they reached an F1-measure of 76.0%.

(Qvarfordt & Zhai, 2005) investigated the concept interest in a Wizard-of-Oz study. They described gaze patterns that are typical of interested and disinterested participants. These patterns were observed during a task that made participants explore different landmarks on a map. Qvarfordt and Zhai observed that when a participant looked at an object with high intensity as well as with a long accumulated duration, this was an indication of interest. They made use of the observed patterns when implementing interest. The resulting system was evaluated positively by the participants.

(Y. Huang et al., 2016) recorded a corpus of 8 dyadic casual conversations in order to test the performance of several recognition algorithms for the level of engagement. They found that *convolutional neural network analysis* performed best.

There are comparatively few studies that have a particular focus on the listener. This may have been caused by that fact that, in general, it is much harder to infer the state of the listener than that of the speaker. The information that is available is expressed through either listener vocalisations, such as "mhm" and "yeah", or through nonverbal expressions, such as facial expressions (frowning, smiling), eye-gaze (looking at the speaker vs. looking away), and head nodding. In short, nonverbal expressions play a particularly important role when focusing on the listener.

One study that has investigated the listeners is (Levitski et al., 2012). They investigated the engagement of one silent participant in a group interaction with three participants. Here, engagement was analysed as a binary distinction between "engaged" and "passive".

Another study that has focused on listeners is (Oertel et al., 2015). In this study, the aim was to discover patterns in the nonverbal behaviour (e.g. gaze as well as visual and acoustic backchannel) of different listener categories (e.g. attentive listener, sideparticipant, and bystander). A further aim was to understand the relationship between individual engagement and the different listener categories. With this information, a classifier could be built that was able to test how well the different listener categories could be classified automatically.

4.4 Concluding remarks

This chapter has provided and overview on possible applications for individual engagement and group involvement. Additionally, it summarized research carried out on both group involvement and individual engagement. This background section is of particular importance for Studies I-III.

5. Attentive Artificial Listeners

After having discussed the analysis of non-verbal cues in human-human interaction, this chapter will now discuss how to use these findings in the development of artificial listeners. The chapter will begin with short overview of robots with social applications in mind. It will continue by describing how virtual agents and social robots can make use of engagement and listener behaviours. Finally, we provide an overview on how to synthesise listener behaviours. For the purpose of this thesis, no explicit differentiation is made between robots and Artificial Agents. The focus here is on their (potential) use rather than on their embodiment.

5.1 Applications of Social Robots in Society

The increasing relevance of robotics becomes clear when opening a newspaper: robots receive more and more coverage and pictures are printed of robot taking care of the elderly or of the entire household. Many societies across the globe are dealing with the consequences of a sharp rise in the ageing population. One these consequences is a rise in the number of people that are suffering from health problems. While 20 years ago, perhaps, one person would stay at home and look after the children and/or look the elderly, this is not the case anymore. This means, that the elderly are often dependent on outside support to help them in their daily life. The nurses providing this outside help for the elderly, however, often do not have much time to spend with them, aside from taking care of the most urgent issues. This may result in social isolation, as these senior citizens are confined to their homes due to their ailments and might not speak to anyone else for days. A

robot that could have a conversation with them might make them feel less alone- even though robots, of course, could never substitute a real human being. In addition, robots and social agents may take on the role of an alarm and remind people to take their medication. For younger people, too, virtual agents and robots may be useful. A virtual agent could, for example, take on the role of a personal assistant or a tutor or tutee and help children with their homework.

The main purpose of socially interactive robotics cf.(Fong et al., 2003) is to interact. In fact, a larger part of social robotics underlies the assumption that humans prefer to interact with robots in the same way they interact with one another. However, there are of course still differences between social robots; not just in their appearance, but perhaps more importantly, also in their design. There are, for example, robots that are solely designed to engage in conversation, whereas others are designed to adhere to social norms but still to fulfil a function (Fong et al., 2003). A further difference may be made between those robots that have been inspired by social psychology, to further our understanding of the workings of the mind, and those that have been designed to model human behaviour (Fong et al., 2003).

Figure 6: Examples of two Robots from the EU-funded project IURO.

(Feil-Seifer & Mataric, 2005), on the other hand, review the uses of socially assistive robots. They mention many different applications for socially assistive robots, including: "rehabilitation robots, wheelchair robots and other mobility, aides, companion robots, manipulator arms for the physically disabled and educational robots". The robots that are of particular importance for this thesis would fall into the category of companion robots.

5.2 The Use of Engagement and Listener Behaviours in Virtual Agents and Robots

In the area of virtual agents, the largest efforts in building artificial listeners have been made within the EU-funded project SEMAINE. This project developed a Sensitive Artificial Listener (SAL), that could interact with humans and react appropriately to their non-verbal behaviour (Douglas-Cowie et al., 2008; Schroder et al., 2012) Maatman et al. (Maatman et al., 2005) describe a listening agent that they later developed into a rapport building agent (Gratch et al., 2006; L. Huang et al., 2011). Embodied virtual agent that can display listening behaviour have also been used in speech-enabled computer games. The EU-funded project NICE developed a computer game with HC Andersen's fairy-tale characters (Gustafson et al., 2005). In this game users could interact with animated characters in a 3D-world using a combination of speech and gestures. The characters were equipped with non-verbal cues that allowed them to display emotions and state-of-mind, as well as verbal backchannels and non-verbal gestures to be used for displaying listening behaviour and to regulate turn-taking.

Emotional display

neutral surprise anger happiness sadness

State-of-mind gestures

idle listening thinking not understanding not hearing

Turn regulation feedback gestures

attention (while walking) attention -- yielding turn taking turn

Figure 7: Pictures from the NICE system (Gustafson et al, 2005)

There are also a number of research efforts on Social robots that can display listening behaviour. Some examples include, the iCat robot (van Breemen et al., 2005); the Kaspar robot (Dautenhahn et al., 2009), which was designed with special focus on research aimed at children on the autistic spectrum; and the Furhat robot(Moubayed et al., 2012), which allows for a very detailed control of eye-gaze as well as facial features. Other widely used robots include the research robot platforms iCub (Metta et al., 2010) and Aldebaran's Nao robot and RoboKind's Zeno. Recently robots with conversational abilities are coming to the commercial market such as Pepper, Jibo, Kirobo, and Robohon.

While there are many things that are important to model in a social robot, such as for example empathy (Castellano et al., 2013; Paiva et al., 2004), the focus of this thesis is engagement and listener behaviours. Therefore, the next section will review current literature on artificial agents and robots with this particular focus in mind.

Modelling engagement in human-robot interactions may be carried out from two perspectives. The first perspective is concerned with the modelling of engagement in the robot. The second perspective is concerned with the recognition of engagement in the human interlocutors by the robot and its appropriate response to it.

Most of the previous research has focused on the second perspective. However, there are also some papers on the first perspective. For example, (Sanghvi et al., 2011) investigated whether it is possible to automatically predict a child's engagement in a chess game with the iCat. In order to achieve this aim, they extracted several body features, such as Body Lean Angle, Slouch Factor, Quantity of Motion as Contraction Index, and some Meta-features. They were able to demonstrate that they exceeded the recognition accuracy of the human baseline with their computational model. Additionally, (Nakano & Ishii, 2010) implemented and evaluated an engagement module that is supposed to estimate a user's engagement in a conversation with a virtual agent. A robot that had been trained on gaze 3-grams from WoZ data collections was found to be perceived as more engaged and it improved the human-robot interaction.

Figure 8: Example of Human-Robot Interaction Scenario

The following studies are all concerned with the second perspective. Bohus and Horvitz (Bohus & Horvitz, 2014), for example, investigated whether it is more advantageous to forecast the disengagement of a human from a conversation than to use a baseline system that will only act once it detects disengagement. Furthermore, they investigated whether it is advantageous to use discourse markers such as "um" in order to bridge a period of time in which the system is unsure whether to disengage. They found that both the use of discourse markers to signal hesitations, were helpful.

(Corrigan et al., 2015) examined the relationship between task and social engagement. They found that, especially for social engagement, it is the users' perception of the robot's characteristics, such as the degree of friendliness, helpfulness, and attentiveness, that resulted in sustained engagement with both the task and robot.

(Szafir & Mutlu, 2012) designed a robot's behaviour in such a way that the robot was able to observe the attention of participants by means of electroencephalography (EEG). They implemented verbal and non-verbal behaviours that were designed to regain the participants' attention. They found that using these cues when participants were losing attention affected their recall abilities in a memory task by 43% over the baseline. These findings were regardless of the gender of the students. They also found that these interventions significantly improved women's motivation as well as their rapport. However, they could not find the same effect for men.

(Eichner et al., 2007) used online eye-gaze analysis in order to infer user attention in objects. They adapted the virtual agents' behaviour accordingly in an object presentation scenario. They found that adaptations to the user's attention supported the successful grounding of deictic agent gestures as well as natural gaze behaviour.

(Matsuyama et al., 2015) implemented a robot designed to keep participants in a four-party conversation engaged by monitoring their participation. If the robot noticed an imbalance in participation, it was able to regulate it.

Other studies had a stronger focus on modelling listener behaviour in general and on the appropriate timing of backchannels in particular. For example, (Morency et al., 2010) used multi-modal cues for the prediction of backchannel opportunities while (Cathcart et al., 2003; Nishimura et al., 2007; N. Ward & Tsukahara, 2000) used mono-modal cues. Moreover, (L. Huang et al., 2011)used backchannels as a strategy to establish rapport between a human and a virtual human.

With regards to visual backchannel, (Candace L Sidner et al., 2006) found that informing their subjects that the robot was able to recognize their head nods was an effective strategy in triggering further nods in the subjects. Additionally, the strategy was further supported by having the robot provide gestural feedback of its nod recognition.

(Poppe et al., 2013) made a comparison between the use of head nods and backchannels in human-robot interaction. They found that the best strategy in terms of third-party observer ratings was to copy the timings of the original listeners' backchannels. The researchers

suggest that this finding might in part be explained by the fact that a higher number of backchannels increases the perceived naturalness (with a ceiling effect (6-12 per minute)). Additionally, it may be explained by the fact that nods are rated as inappropriate less often than audio backchannels, regardless of their timing.

The first part of this section has focused on how to make robots and artificial agents more engagement aware. The second part of this section subsequently focused on how listener behaviours are perceived. It is important to note that the second part is of particular importance to the reader as it provides a frame of reference for Studies IV-VIII. The goal of these studies was to approach a social agent by first concentrating on the modelling of listener behaviours.

However, another important contributing factor has not been covered yet in this literature review: synthesis. Only when access is gained to speech synthesis that is designed for conversational as well, and not only read speech, will it be possible to move towards reaching a complete social agent.

The following section will provide an overview over the current state of the art in synthesis, which is of particular importance for Study VI.

5.3 Concluding remarks

The current chapter has provided an overview of potential future applications of social robots. It has also provided examples of today's conversational robot systems, and it has listed some examples of virtual agents and robots that already use engagement and listener behaviours. The next chapter will describe how audio backchannels can be realised for these systems.

6. Speech Synthesis for Artificial Listeners

The current section will provide an overview of studies on synthesis in general and of studies on conversational synthesis with a specific focus on backchannel tokens in particular.

Currently, there are two approaches that are being used in parallel. The first one is *Unit Selection Synthesis* and the second one is *Statistical Synthesis*. *Unit Selection Synthesis* is generated by recombining units of recorded human speech in different ways (Conkie, 1999; Hunt & Black, 1996). *Statistical synthesis,* in contrast, is based on the use of a statistical model, usually either Hidden Markov Models (HMM) (Raitio et al., 2011; Tokuda et al., 2000; Zen et al., 2009) or Deep Neural Networks (DNN) (Kang et al., 2013; Ling et al., 2013; Zen et al., 2013). While some of the latest synthesisers have the capabilities to produce speech that is, in parts at least, impossible to distinguish from human speech, their speech output often lacks expressiveness and a conversational character.

Previous work on expressive speech synthesis has typically focused on evaluating a distinct set of emotional states, such as happiness, fear, anger, and sadness etc. (Schröder, 2001). However, as Schröder (Schröder, 2004) points out:

> *"In a dialogue, an emotional state may build up rather gradually, and may change over time as the interaction moves on. Consequently, a speech synthesis system should be able to gradually modify the voice in a series of steps towards an emotional state. In addition, it seems reasonable to assume that most human-machine dialogues will require the machine to express only mild, non-extreme emotional states." p211*

One way of achieving this gradual effect is to record different voice qualities and use them directly during concatenative synthesis. This approach has been used in diphone synthesis (Schröder & Grice, 2003) as well as in unit selection synthesis. The technique has also been applied to emotional speech in, for example, (Hofer et al., 2005) and (Iida & Campbell, 2003). Finally, Aylett and Pidcock (Aylett & Pidcock, 2007)use a combination of DSP techniques and unit selection which resulted in the creation of synthesis voices with a more varied and subtle set of speech styles. Currently, HMM and DNN have become the most promising speech synthesis methods.

6.1 Conversational Synthesis

The aim of conversational synthesis is to make a synthesiser better equipped to synthesise conversational speech rather than to "just" read out text. While (Campbell, 2005) argues that the expression of affect by means of speaking style and voice quality is the main factor that distinguishes conversational speech from laboratory speech, other researchers have focused on modelling disfluencies, false starts, self-repetitions, filled pauses, and backchannels as intrinsic parts of conversational synthesis. In the following section, focus will be given to the synthesis of backchannels in particular; a short overview of the other aspects of conversational synthesis will be given below.

As yet, not much work has been carried out in the area of conversational speech synthesis. Work that focused on the synthesis of disfluencies was carried out by (Dall et al., 2014), who investigated the effect of filled pauses and speaking rate on speech comprehension in natural, vocoded, and synthetic speech. They found that the reaction times of participants to a synthesised filled pause were slower than to a silent pause. Dall et al. interpreted this as a short-coming of current synthesis techniques. This finding may, in part, be supported by the results of (Adell et al., 2007), who discovered that the insertion of filled pauses when using unit selection techniques increased the perceived naturalness of the synthesised speech.

Moving away from the synthesis of disfluencies to more general research on conversational synthesis, (Andersson et al., 2012) found that it is advantageous to use conversational speech as input for conversational HMM synthesis. Moreover, some synthesis techniques were developed that attempted to make a voice more expressive using the HMM paradigm, since voice quality is such an important factor for the perception of emotion in speech (Gobl & Ni Chasaide, 2003), and since conversational speech consists of subtle changes in expressiveness than of big changes between different emotions. Techniques were developed by morphing a standard synthesis voice towards the speaking style of a small corpus of expressive speech (Tachibana et al., 2005). Regarding DNN synthesis, Watts et al. (Watts et al., 2015) present a synthesis framework that allows for prosodic sentence level control.

There have also been endeavours to approach conversational synthesis through a consideration of the incremental aspect of conversational speech.(Buschmeier et al., 2012), for example, show that is advantageous for the perception of naturalness to use incremental language generation in conjunction with incremental speech synthesis.

It may be summarised that achieving subtle changes in expressiveness is quite challenging when using unit selection or statistical synthesis. Explicitly modifying prosody through the use of unit selection synthesis typically results in negative effects regarding

the overall quality. Additionally, achieving subtle changes using statistical synthesis is not trivial. Finally, in comparison to the modification of pitch or duration, the modification of voice quality is more difficult while trying to keep the perceived naturalness of the synthesised speech as high as possible.

6.2 Backchannel Synthesis

Like the area of conversational speech synthesis, the field of synthesis of backchannels remains largely unexplored. This section provides an overview of the few existing studies.

The influence of prosodic differences was examined by (Stocksmeier et al., 2007). They focused on investigating prosodic differences in the emotional and pragmatic perception of German "ja" in third-party observers. In order to do so, they synthesised twelve variants of the German "ja". Their results show that prosody is an important factor in the perception of emotions such as happiness, hesitation, and anxiety. in this particular feedback token.

(Pammi et al., 2010) imposed target intonation contours, using the prosody modification techniques MLSA vocoding, FD-PSOLA, and HNM. They did not only vary the intonation contours, they also varied the segmental form of the feedback token. They found an expected drop in naturalness, but they also found unexpected interactions between segmental form and prosody: the perceived meaning of the newly generated token could not have been derived separately from prosody and segmental form.

Campbell's approach (Campbell, 2007), in contrast, consisted of retrieving situation appropriate backchannels from a large data-base of tokens for unit selection synthesis.

6.3 The Perception of Backchannel Synthesis

While the previous section was concerned with the generation and synthesis of backchannel token, the current section is concerned with experiments relating to the perception of backchannel tokens.

(Schröder et al., 2006), for example, investigated the differences in the perception of emotional listener feedback tokens between German and Dutch speakers. They used feedback tokens such as "boah", in the context of a carrier sentence. They were also interested to see whether the perception of a given emotion changed when the feedback tokens were presented in isolation rather than in a carrier sentence. They found that the recognition rates of affect bursts in a carrier sentence were slightly lower than when they were presented in isolation. Also, Dutch participants achieved lower recognition rates for a given emotion when the feedback token was presented in isolation than German speakers. Moreover, the researchers found that there appears to be a general pattern in which negative outbursts were labelled as less acceptable by the participants than more positive outbursts.

In a different study, (N. G. Ward & Escalante-Ruiz, 2009) implemented a Wizard-of-Oz system in which a tutor was interacting with a student in a quiz scenario. The aim of the authors was to investigate whether the students would react more positively when the tutor's acknowledgment matched their certainty about their last quiz answer. In order to study this effect, the authors used various prosodic realisations of the acknowledgement "Good Job". Variations were achieved by taking a neutral "Good Job" token from the corpus and adding modifications to it, such as, elongation or creakiness. They compared the students' perception of the system with a baseline system. The authors found that the subjects' perception of naturalness was significantly higher in the system containing the prosodic modifications.

Similarly, (Forbes-Riley & Litman, 2011) adjusted the content of the output in a tutoring system based on the students' level of uncertainty.

6.4 Concluding remarks

This chapter has provided an overview of synthesis in general with a special focus on conversational aspects such as backchannels. This background section is of particular importance for Studies IV-VI.

.

7. Summary of findings

The next section will provide an overview of the results of the studies included in this thesis. In general, the findings of this thesis can be divided into three topics. The first topic is the modelling of group involvement and individual engagement. The second topic specialises on modelling the engagement of the silent participant (listener) in a conversation. Here, special attention is given to the modelling of listener behaviours. The third and final part of this thesis is concerned with the understanding and usage of listener behaviours in human-robot interactions.

7.1 Group Involvement and Individual Engagement

Both Study I and Study II have contributed to the investigation and modelling of group involvement and individual engagement. Study I summarised an array of findings that were based on the D64 corpus, which was concerned particularly with the study of group involvement. Study II was based on the KTH werewolf corpus and was designed in order to define gaze features. Defining gaze features would make it possible to describe both group involvement and individual engagement.

One of the aims of these studies was to model group conversation as it occurs outside a lab environment. Another aim was to understand what cues humans use in order to decipher group dynamics. We wanted to identify the cues that were low level, since these were more robust for potential later use in dialogue systems.

In study I, we found that F0-range, F0-median, and intensity are related to the perception of group involvement as is the movement of the participants. Moreover, we investigated the use of eye-gaze and blinking. In a direct comparison between visual and acoustic cues, we found that we received better prediction results with the visual cues alone than with the audio cues alone. The strength, and at the same time the limitation, of this study was that the data was based on "recordings in the wild". This means that, besides third-party observer impressions, we did not have any control over the dynamics of the interaction or over the aim, the state of mind, or the objective of the participants.

This limitation led us to amend the follow-up study. First of all, we wanted to have more control over the dynamics of the conversation: we wanted to have not only third-party observer expressions but also first-hand impressions from the participants. These first-hand impressions concerned their own engagement as well as the engagement of the other participants. Since thinking about everyone's engagement in a conversation would undoubtedly influence the conversational dynamics, the impressions had to be timed in such a way that they did not interrupt the interaction while it unfolded.

Another way in which we amended the experiment in Study II was by adding more control for conversational dynamics. Since we had found in Study I that gaze appeared to be a promising cue to detect eye-gaze, we wanted to investigate this further. In order to do so, we situated the participants in the context of a game. The details of the exact corpus set-up can be found in Study IX. Within the game, participants had different objectives, of which we were keeping track, and which allowed us to analyse different aspects of individual behavioural differences within a group context. Moreover, since the game had built-in pauses and breaks, the unfolding of the group dynamics was not hindered by the evaluations of the participants.
From a qualitative analysis, it became clear that the perception of group dynamics seemed to be influenced by the number of people looking at any of the potential speakers. It also appeared to be

influenced by the number of people looking down or looking around the room. Additionally, whether a participant looked at the same person that all other participants were looking at, influenced the perception as well. Using these simple heuristics, we hoped to summarise aspects of group behaviour that together form the whole picture of group involvement. In order to reach this goal, we defined several variables: presence, entropy, symmetry, and MaxGaze. However, we wanted to go one step further: we did not only want to understand group behaviour, we also wanted to understand the changes in the behaviour of individuals within the group. For this reason, we applied the methodology both to the group as a whole and to the individual persons.

This led us to observe several phenomena. First of all, we observed that the clusters we received from the individual rankings of the participants were more fine-grained than the ones that used only gaze information. Yet, gaze information by itself could provide us with new insights about which speaker held a specific role in the conversation. From the individual ranking clusters, we received more than just two clusters. This lent further support to the claim that involvement is not simply a binary phenomenon but rather a scalar one.

While Study II focused on the individual within a group as well as on the group as a whole, it did not distinguish between speaker and listener behaviour. Listener behaviour is particularly difficult to model since the audio channel, which has the potential to bear a lot of information, is very limited. Therefore, the second part of this thesis focused on the modelling of listener behaviour.

7.2 Analysis of Listener Behaviours

In this part of the thesis we wanted to focus on the behaviour of the listener in particular. In Study III, we therefore investigated different categories of listener behaviour and tried to focus on their eye-gaze patterns. We also tried to focus on the frequency of feedback behaviours. We wanted to vary the conversational dynamics as much as possible to allow for the possibility that the same person would

change between listener categories. Due to scalability issues of eye-gaze annotations, we decided to carry out new recordings. Additionally, we used this opportunity to amend the scenario slightly. Instead of using a game, we decided to use a scenario that was somewhere in between a game and real-life: a group interview. The exact details of the corpus set-up can be found in Study X.

The scenario was close enough to a real life experience for the participants: it was something they could identify with. It was also easy enough to structure it for us, so that we could control for different cognitively challenging sections as well as for distinct conversation dynamic sections. We kept the internal evaluation paradigm, but we amended it: this time, the participants themselves did not do their own evaluations; rather, the moderator of the group interview carried out the evaluations after the interview had finished. Moreover, we included four distinct phases. In the first one, the participants had to introduce themselves. We assumed that this was the least cognitively changing task. Subsequently, the participants had to give an elevator pitch about their respective PhD projects and explain the impact on society. Finally, they were asked to come up with a joint PhD proposal. In many of these phases, the participants had to listen to the other participants and had to contribute to their contributions.

The challenge in making predictions about listeners in a conversation is, of course, that the number of channels for communication is limited. Therefore, it is much harder to make judgements about their particular state than about that of speakers. Yet, in Study III we wanted to use eye-gaze information as well as information on head nods and backchannel in order to describe and quantify listener categories. Subsequently, we wanted to use machine learning in order to test whether it would be possible to predict different listener categories.

We found, for example, that an attentive listener receives more eye-gaze from the speaker than a sideparticipant or a bystander. Also, an attentive listener looks downwards less than a sideparticipant or a bystander. Moreover, the highest amount of mutual gaze occurs between the speaker and the attentive listener, followed by the

sideparticipant and the bystander. When relating this to Studies I and II, we also found that the attentive listener is perceived to be more engaged in the conversation than the other two listener categories. The attentive listener is also expected to talk sooner in the future than any of the other listener categories.

Between the different listener categories, feedback behaviour differed significantly. It could, for example, be observed that the attentive listener produced more feedback than the sideparticipant and the bystander. Additionally, it could be observed overall that visual feedback was more frequent than audio feedback.

While we investigated the frequencies of feedback behaviours across different listener categories, we did not investigate the actual prosodic realisations of these feedback tokens, nor did we investigate how they were perceived by third-party observers without their visual context. Therefore, in Study IV we set-up a crowdsourcing experiment in which we sampled backchannel token from the recordings and asked third-party observers to rate them in terms of their perceived attentiveness. From this experiment we learned which prosodic cues are relevant for the perception of attentiveness in backchannel tokens. Study V subsequently investigated how different backchannels and head nods were perceived in a virtual agent. For this, we used the information on head nods and backchannel we had collected from the Kinect, and we copy synthesise it in a virtual agent. In order to produce comparable head nods and backchannels, we embedded them in the same carrier sentence. This ensured that annotators would rate them within the same interactional environment. We found that head nods were perceived to exert more attentiveness than backchannels. We also found that the combination of both channels was perceived to exert the highest level of attentiveness.

These studies analysed audio-visual listener behaviour in human-human interactions and tested them through perception tests. However, they did not test the patterns found in online human-robot interactions. The third and final part of this thesis therefore deals with this aspect of listener behaviour: its identification and generation in human-robot interactions.

7.3 Listener Behaviour in Human Robot Interactions

The final part this thesis is concerned with the generation, understanding, and use of listener behaviour in human-robot interactions.

The generation of listener behaviour was discussed in Study VI. In particular, we wanted to investigate whether we could design parametric synthesis in such a way that we had explicit control over the percentage of people that would perceive a given backchannel as more attentive than the one produced by the synthesiser. We were able to synthesise two to three different classes of attentiveness.

In Studies VII and VIII, we wanted to test the reaction of the user while he was interacting with the agent/robot in real time. This contrasted with the approach taken in the second part of this thesis, where we carried out perception experiments.

In Study VII, we therefore used the scenario of a narrator (the agent), who provided cooking recipes. The user had to provide backchannel whenever he/she thought it was the right moment to do so. This scenario was created in order to test whether the participant would be prone to provide more backchannel when the agent was looking directly at him/her than when the agent was looking to the side.

In Study VIII, we wanted to test whether it was possible to detect uncertainty in backchannel token when a robot is giving directions to a human. In order to make the description of the route more ambiguous we decided to add conflicting landmarks. For example, we added two pictures of a tower on the map: one depicting the "Tower of Pisa" and the other one depicting "The Eifeltower".

7.4 Research questions revisited

In the introductory chapter, five general research questions were stated.

A: *What are the audio-visual cues to group involvement and individual engagement?*

1. **In which way should a corpus recording be designed in order for it to capture a high variance in conversational dynamics?**

   Studies IX and X present two different settings for recording corpora which attempt to have a high variance in conversational dynamics. The first setting is a game scenario in which participants are very engaged, but in which the conversational dynamics are achieved by means of the different roles in the game. Study X, on the other hand, achieves variation in conversational dynamics by inserting four different phases into the corpus recordings. These phases are: the *introduction*, the *elevator pitch*, the *impact on society*, and the *collaborative project*. While the participants are mainly having a monologue during the *introduction* as well as during the *elevator pitch* and *impact on society*, they are engaging in a mainly collaborative discussion during the design of the *collaborative project*.

2. **Is group involvement a binary or scalar phenomenon?**

   Group involvement is a scalar phenomenon. The findings of Study I confirm that the prediction accuracies are better, in terms of reduction in error rate, with three levels of group involvement, than with simply a binary distinction.

3. **Is it related to prosodic mimicry and is the fusion of audio-visual cues beneficial to its prediction?**

   We show that prosodic mimicry is correlated with group involvement. Comparing the prediction accuracies of visual cues with those of acoustic cues as well as with those of a fusion of both modalities, the results of Study I confirm that the fusion of audio-visual cues is beneficial to the prediction of group involvement.

4. **Is it possible to define eye gaze features that describe individual engagement as well as group involvement? Are these features useful for automatic prediction of engagement and involvement?**

   Yes, the findings of Study II confirm that it is possible to define such eye-gaze features. For example, the features *entropy* and *presence* enable us to explain changes in the roles of the participants within the game. Moreover, *presence* is the feature that was able to best distinguish between low and high group involvement. The *entropy* feature, on the other hand, was able to distinguish best between *lead* and any other feature. *Symmetry,* or the amount of mutual gaze between participants, increases when the perception of group involvement increases. While similar findings have been made for dyadic conversation, they have not been shown before for multi-party conversations and group involvement. *MaxGaze,* similar to *presence,* is able to distinguish best between high and low group involvement. The results of Study II also confirm that these features are useful for automatic prediction. An average accuracy of 71.3% is achieved for the prediction of the 4 classes of group involvement.

B: *How do human listeners signal their degree of engagement?*

1. **Are there listener-type-specific differences in audio-visual behavioural patterns?**

   Yes, Study III shows that there are listener-type-specific differences in audio-visual behavioural patterns. In the study, a differentiation is made between three listener categories: *attentive listener*, *side-participant,* and *bystander*. The investigation demonstrates that the attentive listener produces more feedback than the side-participant and the bystander. This holds for both the visual feedback (head nods) and for the acoustic feedback (backchannels and feedback tokens). Moreover, the categories also differ in terms of eye-gaze patterns. Not only does the attentive speaker look down less than the side-participant or the attentive speaker, he also receives more gaze from the speaker. Consequently, he also has a higher amount of mutual gaze with the speaker. Interestingly, there is no significant difference in the amount of gaze directed towards the speaker between any of the aforementioned listener categories.

2. **And if there are differences, are they perceived differently in terms of attentiveness, focusing solely on listener feedback tokens? Are there backchannels types that are perceived to be more attentive than others?**

   Study IV then shows that head nods are indeed perceived to convey more attentiveness than audio backchannels. The fusion of both modalities, however, conveys the highest amount of attentiveness.

C: Can we model an artificial listener that can display different degrees of attention?

1. **Is it possible to provide a speech synthesizer with an explicit control of the degree of attentiveness?**

   Yes, Study V shows that it is possible to provide a speech synthesiser with explicit control of the perceived degree of attentiveness in bisyllabic backchannel tokens. It is possible to show, by means of crowd-sourcing experiments, that participants perceive the expressed attentiveness in synthesised backchannel token in the same way as in natural backchannel tokens.

2. **Can statistical parametric speech synthesis be used to validate prosodic models of perceived attentiveness?**

   It was also possible to validate the prosodic model from Study IV by generating backchannel tokens that had both a higher and a lower perceived attentiveness than any of the backchannel tokens observed in the original data.

2. **Is it possible to affect the amount of listener feedback in human-robot interaction by controlling for the gaze direction of the virtual agent?**

   Yes, the results of Study VI show that it is possible to influence the amount of feedback a participant provides by controlling for the gaze direction of a virtual agent. In fact, it was demonstrated that participants provided more feedback at points in the story where the virtual agent gazed directly at them.

3. How is uncertainty in human-robot interaction expressed in the feedback utterances of the listener?

The results of Study VIII reveal that it is possible to characterise uncertainty in backchannel tokens. For instance, it was possible to describe differences in the distribution of lexical items in terms of uncertainty: Uncertainty, for example, is more often expressed with "mm" and "ah", whereas certainty is more often expressed with "yes". Additionally, feedback tokens that are perceived to convey certainty are expressed with a higher intensity, a shorter duration, and a comparatively greater rising pitch. This was tested through the feedback token "okay", which occurs approximately equally frequent in both certain and uncertain conditions. For this token, it can be observed that, except for pitch slope, all other differences in prosodic cues stay significant.

8. General conclusions

The work described in this thesis covers various aspects of human engagement in conversations. On the one hand, focus has been given to perceptions of group involvement in general. On the other hand, the work this thesis has zoomed in on two specific aspects of it: individual engagement, the role of the listener, and implications for human-robot interactions. The methods used and the topic of this thesis, fit well into the framework of social signal processing.

A common feature of the studies contained in this thesis is the use of human perception and multi-modal feature extraction to group model involvement, individual engagement, and attention. For example, in Study I this method has been used to model group involvement; in Study II, it has been used to model group involvement as well as individual engagement; and in Study III, it has been used to model listener categories.

However, the various studies also differ in methodology. Rather than using raw data, Study II has grouped the gaze-features into four groups first. Another difference lies in the use of annotators. While Studies I-III rely mainly on post-graduates for their annotations, **Studies IV** and **V** apply crowd-sourcing techniques. The advantage of using crowd-sourcing techniques is that the pool of annotators does not only cover the perceptions of the graduate population but also the perceptions of a wider public.

Furthermore, the annotations are returned very promptly, which would not have been possible for this amount of data when relying solely on a post-graduate pool.

There are also disadvantages, however: there is generally less control over the quality of annotations, there is less control over the quality of annotations, and it is costlier to use.

Another methodological difference is the use of machine learning: while Studies I-III use categorical machine learning Studies IV and V use ordinal machine learning. For future work, it might be interesting to explore whether ordinal machine learning might also be advantageous for predicting group involvement, individual engagement and the different listener categories.

One of the main contributions of this thesis is the exploration of behavioural dynamics within multi-party interactions using multi-modal low-level nonverbal cues. These nonverbal cues, such as eye-gaze and speech features, allow for more robust inferences than speech recognition and subsequent language processing. While a deeper understanding of the dialogue context is not possible using nonverbal features alone, I hope that this thesis shows that affective processing is possible.

8.1 Limitations and Discussion

Some limitations should be considered when interpreting the findings of this thesis.

The first concern relates to the limited number of participants in Study I and Study II: in Study I only four people participated and in Study II only eight. These numbers, of course, do not suffice to make any general claims about group involvement or individual engagement. The small number of participants in Study I was the result of the limitations of the D64 corpus, which only comprised four participants. Yet, the unique character of the corpus made us decide to use it instead of a larger corpus, which would not have included the unconstrained characteristics of the D64 corpus.

Similarly, the limitations in the number of participants of Study II were in part due to the difficulties of recruiting eight people who all had to be available at the same time, and had to volunteer to take part in the recordings. Another limiting factor for both for Study I and Study II is the scalability of gaze annotations. The gaze annotations in both studies were still carried out manually due to a lack of access to suitable software or gaze trackers.

A second concern lies with the limited use of nonverbal features. The studies in this thesis mainly concentrated on investigating the use of eye-gaze, prosody, and backchannels. However, other nonverbal facial features, such as frowning, smiling, raising an eyebrow, and laughing were not considered.

While it is true that all of these features could have a potential influence on the perception of group involvement, individual engagement, and listener behaviours, their inclusion would have made it impossible to analyse gaze and prosody as closely as it has been done now. Given the time limitations and scope of this thesis, it was thus decided to limit the analysis to gaze, prosody, and backchannels.

A further potential concern lies with the consideration of context. The context of interactions is important on both a macro and on a micro level: on the macro level, it is important to consider the general context, such as whether participants are in a formal meeting or whether they are at a party with a group of friends. On the micro level, it is important to consider the specific context of the room in which a conversation is taking place. Factors like distracting objects in the vicinity of the participants may have consequences for the subsequent analysis and interpretation of the nonverbal behavioural patterns. This is why we controlled for objects that could have distracted the users in Study II. Moreover, while the scope of this thesis did not allow for a consideration of all possible contexts, it does analyse corpora that vary quite considerably in their dynamics.

Seating arrangements during the recordings may also have influenced the eye-gaze patterns of participants. However, we tried to control for this in Study II, III, IX and X, by seating participants in a synchronous way

Hierarchical differences provided another factor that may have impacted the outcomes of this investigation. In Study I, the group of participants consisted of two female students and three male higher ranking academics (whose function ranged from a final-year PhD student to a Professor). However, we controlled for social hierarchy in studies II-III. For the remaining studies, hierarchical differences should not have been an influential factor due to their perceptual character and/or due to the fact that the interaction took place between humans and computers.

Finally, in all of the human-human studies presented in this thesis, the groups of participants consist of a mixture of native and non-native speakers of English. All of the participants, however, had a very high proficiency in English. Moreover, information about participants' mother tongue, their gender, and culture is recorded and could be used at a future study. However, the explicit analysis of these three factors lies outside the scope of this thesis

8.2 Future work

The current and final section of this thesis will provide an overview over potential directions of future work.

The first potential area that would be interesting to more closely is the selection of window size for the study of group involvement, individual engagement, and different listener categories. For example, in Study I as well as in Study III it was decided to use a fixed windowing approach. However, in Study II a mixture between an interval and an event-based approach was used. In a future study, it would be interesting to empirically test whether an event-based approach may provide extra information and, if so, what type of information. Perhaps, both approaches may be combined in order to better understand the unfolding of group involvement, individual engagement and listener categories over time.

Another question that would be very interesting to explore is whether the provision of feedback about group cohesion to participants during a conversation has any impact. Information such as the amount an individual person contributed to the conversation and the amount of dominance over the conversation may be provided to participants. It could, consequently, be explored in what format they can profit from the provision of feedback.

Moreover, it could also be worthwhile to explore whether participants profiting more from an online, or a short-time-post summarization evaluation or from a long time evaluation. Finally, it would be interesting to investigate whether a change in the conversational patterns leads to an increase in group productivity.

A further study for study is to test and, potentially, to extend the models developed here to other environmental contexts. For example, all studies used in this thesis, except for Study I, controlled for the presence of objects during the interaction.

In a future study, it would be interesting to investigate how model developed here performs on a corpus in which objects were present in the field of view of the participants. Moreover, it would be interesting to explore which machine learning techniques may be used to best adapt to different environmental contexts.

Additionally, it would be worthwhile to investigate the exact relationship between individual engagement and group involvement.

For example, we did not measure whether the *presence* feature, introduced in study II, would represent the whole group at a given point in time equally well as the individual. It is possible that, in the judgements of the raters, certain participants were considered to a greater degree than other participants. It would be interesting to calculate the exact weight each and every participant had in the overall group involvement rating.

The differences between the collaborative and the competitive sections of the various corpora are another avenue for further investigation. All of the corpora used in this thesis include such sections. It would be insightful to explicitly investigate whether the nonverbal behavioural patterns of the participants vary in these

different sections. It would also be interesting to include dialogue acts into the analysis. Potentially, the insights and models gained here could then be used for multi-party human-robot interactions. A potential research question for such an analysis could be: at what point in a conversation it is advantageous with respect to for a robot to interrupt a human-human interaction?

Another direction to be taken might be the development of a speech synthesiser that is able not only to express "mhms" but also to express other lexical items such as "yeah" and "okay". We might then go on to model further conversational phenomena such as "filled pauses". The aim of developing such a synthesiser is to use it in human-robot interaction studies, in which an important role of the robot is to be an attentive listener.

Finally, the area of multi-party human-robot interaction needs further exploration. While all of the human-human interaction studies in this thesis were multi-party interactions, all of the human-robot studies were dyadic. In the future, I would like to test how insights gained from multi-party human-human interaction can also be exploited for human-robot interaction.

9. References

Adell, J., Bonafonte, A., & Escudero, D. (2007). *Filled pauses in speech synthesis: towards conversational speech.* Paper presented at the International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic.

Allwood, J., Nivre, J., & Ahlsen, E. (1990). Speech management—On the non-written life of speech. *Nordic Journal of Linguistics, 13*(1), 3-48.

Allwood, J., Nivre, J., & Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics}, 9*(1), 1-26.

Altmann, U., Oertel, C., & Campbell, N. (2012). Conversational involvement and synchronous nonverbal behaviour. In A. Esposito, A. Vinciarelli, R. Hoffmann, & V. Müller (Eds.), *Cognitive Behavioural Systems. Lecture Notes in Computer Science* (pp. 343--352): Springer.

Andersson, S., Yamagishi, J., & Clark, R. A. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication, 54*(2), 175-188.

Andrist, S., Pejsa, T., Mutlu, B., & Gleicher, M. (2012). *Designing effective gaze mechanisms for virtual agents*. Paper presented at the SIGCHI conference on Human factors in computing systems, Austin, Texas.

Andrist, S., Tan, X. Z., Gleicher, M., & Mutlu, B. (2014). *Conversational gaze aversion for humanlike robots*. Paper presented

at the 2014 ACM/IEEE international conference on Human-robot interaction, Bielefeld, Germany.

Aran, O., & Gatica-Perez, D. (2013). *One of a Kind: Inferring Personality Impressions in Meetings*. Paper presented at the ICMI 2013, Sydney.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*: Cambridge University Press.

Argyle, M., & Graham, J. A. (1976). The central Europe experiment: Looking at persons and looking at objects. *Environmental psychology and nonverbal behavior, 1*(1), 6-16.

Aylett, M. P., & Pidcock, C. J. (2007). *The CereVoice characterful speech synthesiser SDK*. Paper presented at the AISB'07, Newcastle, Uk.

Bailly, G., Elisei, F., Juphard, A., & Moreaud, O. (2016). *Quantitative analysis of backchannels uttered by an interviewer during neuropsychological tests*. Paper presented at the Interspeech, San Francisco, USA.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as Co-Narrators. *Journal of Personality and Social Psychology, 79*(6), 941-952.

Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening, 25*(3), 178-198.

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs Thatcher interrupted so often? *[Letters to Nature] Nature, 300*, 744-747.

Bednarik, R., Eivazi, S., & Hradis, M. (2012). *Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement*. Paper presented at the Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction, Los Angeles, USA.

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., . . . Murino, V. (2016). *Detecting emergent leader in a meeting environment using nonverbal visual features only*. Paper presented at the ICMI 2016, Tokyo, Japan.

Bohus, D., & Horvitz, E. (2014). *Managing Human-Robot Engagement with Forecasts and... um... Hesitations* Paper presented at the ICMI 2014, Istanbul.

Bonin, F., Böck, R., & Campbell, N. (2012). *How do we react to context? annotation of individual and group engagement in a video corpus*. Paper presented at the Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), Amsterdam, Netherlands.

Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., & Schlangen, D. (2012). *Combining incremental language generation and incremental speech synthesis for adaptive information presentation*. Paper presented at the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Metz, France.

Cabrera-Quiros, L., Gedik, E., & Hung, H. (2016). *Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios*. Paper presented at the ICMI 2016, Tokyo, Japan.

Campbell, N. (2005). Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE transactions on information and systems, 88*(3), 376-383.

Campbell, N. (2007). *Towards conversational speech synthesis; lessons learned from the expressive speech processing project*. Paper presented at the Speech Synthesis Workshop, Bonn, Germany.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation, 41*(2), 181-190.

Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & Mcowan, P. W. (2013). Multimodal affect modeling and recognition for empathic robot companions. *International Journal of Humanoid Robotics, 10*(01).

Cathcart, N., Carletta, J., & Klein, E. (2003). *A shallow model of backchannel continuers in spoken dialogue*. Paper presented at the the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1, Budapest, Hungary.

Charfuelan, M., Schröder, M., & Steiner, I. (2010). *Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings*. Paper presented at the Interspeech 2010, Makuhari, Japan.

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., . . . Harper, M. (2005). *VACE multimodal meeting corpus*. Paper presented at the International Workshop on Machine Learning for Multimodal Interaction, Martigny, Switzerland.

Chollet, M., Prendinger, H., & Scherer, S. (2016). *Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training*. Paper presented at the ICMI 2016, Tokyo, Japan.

Clark, H. H. (1996). *Using language*: Cambridge University Press.

Coker, D., & Burgoon, J. (1987). The nature of conversational involvement and nonverbal encoding patterns. *Human Communication Research, 13*(4), 463-494.

Conkie, A. (1999). *Robust unit selection system for speech synthesis*. Paper presented at the 137th meeting of the Acoustical Society of America.

Corrigan, L. J., Basedow, C., Küster, D., Kappas, A., Peters, C., & Castellano, G. (2015). *Perception matters! Engagement in task orientated social robotics*. Paper presented at the Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, Kobe, Japan.

Cummins, F. (2012). Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes, 27*(10), 1525-1549.

Curtis, K., Jones, G. J. F., & Campbell, N. (2015). *Effects of Good Speaking Techniques on Audience Engagement*. Paper presented at the ICMI 2015, Seattle, USA.

Dall, R., Wester, M., & Corley, M. (2014). *The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech*. Paper presented at the Interspeech 2014, Singapore.

Damian, I., Baur, T., & André, E. (2016). *Measuring the impact of multimodal behavioural feedback loops on social interactions*. Paper presented at the ICMI 2016, Tokyo.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Hatice Kose-Bagci, Mirza, N. A., & Blow, M. (2009). KASPAR – a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics, 6*(3-4), 369-397.

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., & Heylen, D. (2008). *The sensitive artificial listner: an induction technique for generating emotionally coloured conversation*. Paper presented at the LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko.

Duncan, S. (1974). On the structure of speaker--auditor interaction during speaking turns. *Language in society, 3*(2), 161-180.

Edlund, J., & Beskow, J. (2009). Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech, 52*(2-3), 351-367.

Eichner, T., Prendinger, H., André, E., & Ishizuka, M. (2007). *Attentive presentation agents*. Paper presented at the International Workshop on Intelligent Virtual Agents, Paris, France.

Feil-Seifer, D., & Mataric, M. J. (2005). *Defining socially assistive robotics*. Paper presented at the 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005., Chicago, USA.

Finnerty, A. N., Kalimeri, K., & Pianesi, F. (2014). *Towards Happier Organisations: Understanding the Relationship between Communication and Productivity*. Paper presented at the International Conference on Social Informatics, Barcelona, Spain.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems, 42*(3), 143-166.

Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication, 53*(9), 1115-1136.

Frauendorfer, D., Mast, M. S., Sanchez-Cortes, D., & Gatica-Perez, D. (2014). Emergent Power Hierachies and Group Performance. *International Journal of Psychology, 50*(5), 392-396.

Gatica-Perez, D., Mccowan, I., Zhang, D., & Bengio, S. (2005). *Detecting group interest level in meetings*. Paper presented at the Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, USA.

Gerritsen, D., Zimmerman, J., & Ogan, A. (2015). *Exploring Power Distance, Classroom Activity, and the International Classroom Through Personal Informatics*. Paper presented at the Sixth International Workshop on Culturally-Aware Tutoring Systems (CATS2015), Madrid, Spain.

Gobl, C., & Ni Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*, 189-212.

Goffman, E. (1966). *Behaviour in public places: Notes on the social organization of gatherings*: Simon and Schuster.

Goffman, E. (1967). *Interaction Ritual: Essays in Face to Face Behavior*: Aldine Transaction.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L.-P. (2006). *Virtual rapport*. Paper presented at the International Workshop on Intelligent Virtual Agents, Marina del Rey, USA.

Gravano, A., & Hirschberg, J. (2009). *Backchannel-inviting cues in task-oriented dialogue*. Paper presented at the Interspeech, Brighton, UK.

Gustafson, J., Boye, J., Fredriksson, M., Johanneson, L., & Königsmann, J. (2005). *Providing computer game characters with conversational abilities*. Paper presented at the International Workshop on Intelligent Virtual Agents, Kos Island, Greece.

Gustafson, J., & Neiberg, D. (2010). *Prosodic cues to engagement in non-lexical response tokens in Swedish*. Paper presented at the DiSS-LPSS Tokyo, Japan.

Heylen, D. (2008). Listening Heads *Modeling Communication with robots and virtual humans* (pp. 241-259): Springer.

Heylen, D., Bevacqua, E., Tellier, M., & Pelachaud, C. (2007). *Searching for prototypical facial feedback signals*. Paper presented at

the International Workshop on Intelligent Virtual Agents, Paris, France.

Hincks, R., & Edlund, J. (2009). PROMOTING INCREASED PITCH VARIATION IN ORAL PRESENTATIONS WITH TRANSIENT VISUAL FEEDBACK. *Language Learning & Technology, 13*(3), 32-50.

Hofer, G. O., Richmond, K., & Clark, R. A. (2005). *Informed blending of databases for emotional speech synthesis*. Paper presented at the Interspeech 2005, Lisbon, Portugal.

Huang, L., Morency, L.-P., & Gratch, J. (2011). *Virtual Rapport 2.0*. Paper presented at the International Workshop on Intelligent Virtual Agents, Reykjavik, Island.

Huang, Y., Gilmartin, E., & Campbell, N. (2016). *Conversational Engagement Recognition Using Auditory and Visual Cues*. Paper presented at the Interspeech 2016, San Francisco, USA.

Hung, H., & Chittaranjan, G. (2010). *The idiap wolf corpus: exploring group behaviour in a competitive role-playing game*. Paper presented at the 18th ACM international conference on Multimedia, Firenze, Italy.

Hunt, A. J., & Black, A. W. (1996). *Unit selection in a concatenative speech synthesis system using a large speech database*}. Paper presented at the Acoustics, Speech, and Signal Processing, 1996, Atlanta, USA.

Iida, A., & Campbell, N. (2003). Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *International Journal of Speech Technology, 6*, 379-392.

Johansson, M., Skantze, G., & Gustafson, J. (2013). *Head pose patterns in multiparty human-robot team-building interactions*. Paper presented at the International Conference on Social Robotics, Bristol, UK.

Kang, S., Qian, X., & Meng, H. (2013). *Multi-distribution deep belief network for speech synthesis*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada.

Kawahara, T., Hayashi, S., & Takanashi, K. (2013). *Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations.* Paper presented at the Interspeech 2013, Lyon, France.

Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., & Ward, N. (2016). *Prediction and Generation of Backchannel Form for Attentive Listening Systems*. Paper presented at the Interspeech 2016, San Francisco, USA.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.

Kim, J., Truong, K. P., & Evers, V. (2016). *Automatic detection of children's engagement using non-verbal features and ordinal learning.* Paper presented at the Workshop on Child Computer Interaction, San Francisco, USA.

Koutsombogera, M., Al Moubayed, S., Bollepalli, B., Hussen, A., Oertel, C., Stefanov, K., & Varol, G. (2014). *The Tutorbot Corpus—A Corpus for Studying Tutoring Behaviour in Multiparty Face-to-Face Spoken Dialogue*. Paper presented at the LREC 2014.

Lai, C. (2010). *What do you mean, you're uncertain?: the interpretation of cue words and rising intonation in dialogue.* Paper presented at the Interspeech, Makuhari, Japan.

Lai, C., Carletta, J., & Renals, S. (2013). *Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features.* Paper presented at the Interspeech 2013, Lyon, France.

Laskowski, K. (2008). *Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings.* Paper presented at the Spoken Language Technology Workshop, 2008. SLT 2008. IEEE, Goa, India.

Levitski, A., Radun, J., & Jokinen, K. (2012). *Visual interaction and conversational activity*. Paper presented at the Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction.

Ling, Z.-H., Deng, L., & Yu, D. (2013). *Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada.

Liscombe, J., Venditti, J. J., & Hirschberg, J. (2006). *Detecting question-bearing turns in spoken tutorial dialogues*. Paper presented at the Interspeech 2006, Pittsburgh, USA.

Litman, D., Paletz, S., Rahimi, Z., Allegretti, S., & Rice, C. (2016). *The Teams Corpus and Entrainment in Multi-Party Spoken Dialogues*. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas.

Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Kopp, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces, 7*(1-2), 5-18.

Maatman, R., Gratch, J., & Marsella, S. (2005). *Natural behavior of a listening agent*. Paper presented at the International Workshop on Intelligent Virtual Agents, Kos Island, Greece.

Malisz, Z., Włodarczak, M., Buschmeier, H., Kopp, S., & Wagner, P. (2012). *Prosodic characteristics of feedback expressions in distracted and non-distracted listeners*. Paper presented at the Proceedings of The Listening Talker. An interdisciplinary workshop on natural and synthetic modification of speech in response to listening condition, Edingburgh, UK.

Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., & Zancanaro, M. (2007). *Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection*. Paper presented at the Proceedings of the 2007 workshop on Tagging, mining and retrieval of human related activity information, Nagoya, Japan.

Mashadi, A., Mathur, A., Broeck, M. V. d., Vanderhulst, G., & Kawsar, F. (2016). *Understanding the impact of personal feedback on face-to-face interactions in the workplace*. Paper presented at the ICMI 2016, Tokyo, Japan.

Matsuyama, Y., Akibaa, I., Fujieb, S., & Kobayashi, T. (2015). Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech and Language, 33*, 1-24.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., . . . Bernardino, A. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *23*(8), 1125-1134.

Moore, D. (2002). *The IDIAP smart meeting room*. Retrieved from Morency, L.-P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems, 20*(1), 70-84.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., . . . Rochet, C. (2007). The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources & Evaluation, 41*, 389-407.

Moubayed, S. A., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction *Cognitive behavioural systems* (pp. 114-130).

Moubayed, S. A., Edlund, J., & Gustafson, J. (2013). *Analysis of gaze and speech patterns in three-party quiz game interaction*. Paper presented at the Interspeech 2013, Lyon, France.

Nakano, Y. I., & Ishii, R. (2010). *Estimating User's Engagement from Eye-gaze Behaviors in Human-Agent Conversations*. Paper presented at the 15th international conference on Intelligent user interfaces.

Neiberg, D., Salvi, G., & Gustafson, J. (2013). Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication, 55*(3), 451-469.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia, 16*(4), 1018-1031.

Nishimura, R., Kitaoka, N., & Nakagawa, S. (2007). *A spoken dialog system for chat-like conversations considering response timing*. Paper presented at the International Conference on Text, Speech and Dialogue, Pilsen, Czech Republic.

Oertel, C., Cummins, F., Edlund, J., Wagner, P., & Campbell, N. (2013a). D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces, 7*(1-2), 19-28.

Oertel, C., Funes Mora, K. A., Gustafson, J., & Odobez, J.-M. (2015). *Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions*. Paper presented at the ICMI 2015, Seattle, USA.

Oertel, C., Mora, K. A. F., Sheikhi, S., Odobez, J.-M., & Gustafson, J. (2014). *Who Will Get the Grant?: A Multimodal Corpus for the Analysis of Conversational Behaviours in Group Interviews*. Paper presented at the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions.

Oertel, C., & Salvi, G. (2013). *A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue*. Paper presented at the ICMI 2013, Sydney, Australia.

Oertel, C., Salvi, G., Götze, J., Edlund, J., Gustafson, J., & Heldner, M. (2013b). *The KTH Games Corpora: How to Catch a Werewolf*. Paper presented at the Multimodal Corpora 2013, Edingburgh, UK.

Oertel, C., Scherer, S., & Campbell, N. (2011a). *On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation*. Paper presented at the Interspeech 2011, Florence, Italy.

Oertel, C., Scherer, S., & Campbell, N. (2011b). *On the use of multimodal cues for the prediction of involvement in spontaneous conversation*. Paper presented at the Interspeech, Florence, Italy.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). *Gaze Patterns in Turn-Taking*. Paper presented at the Interspeech 2012, Portland, USA.

Olguin, D. O., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, 39*(1), 43-54.

Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., . . . Hall, L. (2004). *Caring for agents and agents that care: Building empathic relations with synthetic agents*. Paper presented at the Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1.

Pammi, S., & Schröder, M. (2009). *Annotating meaning of listener vocalizations for speech synthesis*. Paper presented at the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, Netherlands.

Pammi, S., Schröder, M., Charfuelan, M., Türk, O., & Steiner, I. (2010). *Synthesis of listener vocalisations with imposed intonation contours*. Paper presented at the SSW, Kyoto, Japan.

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., . . . Vinciarelli, A. (2011). Social signal processing: the research agenda *Visual analysis of humans* (pp. 511-538): Springer.

Peters, C., Asteriadis, S., & Karpouzis, K. (2010). Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces, 3*(1-2), 119--130.

Peters, C., Castellano, G., & de Freitas, S. (2009). *An exploration of user engagement in HCI*. Paper presented at the Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, Boston, USA.

Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*: Weidler.

Pon-Barry, H. (2008). *Prosodic manifestations of confidence and uncertainty in spoken language*. Paper presented at the Interspeech, Antwerp, Belgium.

Poppe, R., Truong, K. P., & Heylen, D. (2013). Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous agents and multi-agent systems, 27*(2), 235-253.

Qvarfordt, P., & Zhai, S. (2005). *Conversing with the user based on eye-gaze patterns*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Portland, USA.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., & Alku, P. (2011). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(1), 153-165.

Rasipuram, S., S.B., P. R., & Jayagopi, D. B. (2016). *Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study*. Paper presented at the ICMI 2016, Tokyo, Japan.

Rehm, M., & Nakano, Y. (2008). *Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces.* Paper presented at the Proceedings of Workshop on Enculturating Conversational Interfaces by Socio-cultural Aspects of Communication, 2008 International Conference on Intelligent User Interfaces Gran Canaria, Canary Islands.

Salam, H., Celiktutan, O., Hupont, I., Gunes, H., & Chetouani, M. (2016). Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access (accepted)*.

Sanchez-Cortes, D., Aran, O., Jayagopi, D. B., Mast, M. S., & Gatica-Perez, D. (2013). Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces, 7.1*(2), 39-53.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). *Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion.* Paper presented at the HRI, Lausanne, Switzerland.

Schröder, M. (2001). *Emotional speech synthesis: a review.* Paper presented at the Interspeech, Aalborg, Denmark.

Schröder, M. (2004). *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions*. Paper presented at the Tutorial and Research Workshop on Affective Dialogue Systems, Kloster Irsee, Germany.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., . . . Pelachaud, C. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing, 3*(2), 165-183.

Schröder, M., & Grice, M. (2003). *Expressing vocal effort in concatenative synthesis*. Paper presented at the ICPhS, Barcelona, Spain.

Schröder, M., Heylen, D., & Poggi, I. (2006). *Perception of non-verbal emotional listener feedback*. Paper presented at the Speech Prosody, Lund, Sweden.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., . . . Konosu, H. (2009). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Comuting, 27*, 1760-1774.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science, 1*(2), 305-319.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence, 166*(1), 140-164.

Sidner, C. L., Lee, C., Morency, L.-P., & Morency, C. (2006). *The effect of head-nod recognition in human-robot conversation*. Paper presented at the Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction, Salt Lake City, USA.

Stefanov, K., & Beskow, J. (2016). *A Multi-party Multi-modal Dataset for Focus of Visual Attention in Human-human and Human-robot Interaction*. Paper presented at the Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

Stocksmeier, T., Stefan Kopp, & Gibbon, D. (2007). *Synthesis of prosodic attitudinal variants in German backchannel ja*. Paper presented at the Interspeech 2007, Antwerp, Belgium.

Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., & Pantic, M. (2011). *A multimodal database for mimicry analysis*. Paper presented at the International Conference on Affective Computing and Intelligent Interaction, Memphis, USA.

Szafir, D., & Mutlu, B. (2012). *Pay attention!: designing adaptive agents that monitor and improve user engagement*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, USA.

Tachibana, M., Yamagishi, J., Masuko, T., & Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE transactions on information and systems, 88*(11), 2484-2491.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). *Speech parameter generation algorithms for HMM-based speech synthesis.* Paper presented at the Acoustics, Speech, and Signal Processing, 2000, Istanbul, Turkey.

Traum, D., Vault, D. D., Lee, J., Wang, Z., & Marsella, S. (2012). *Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation.* Paper presented at the IVA 2012.

Truong, K. P., Poppe, R., Kok, I. d., & Heylen, D. (2011). *A Multimodal Analysis of Vocal and Visual Backchannels in Spontaneous Dialogs.* Paper presented at the Interspeech 2011, Florence, Italy.

van Breemen, A., Yan, X., & Meerbeek, B. (2005). *iCat: an animated user-interface robot with personality*. Paper presented at the Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, Utrecht, Netherlands.

Vertegaal, R., Slagter, R., Veer, G. V. d., & Nijholt, A. (2001). *Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes.* Paper presented at the Proceedings of the

SIGCHI conference on Human factors in computing systems, Seattle, USA.

Vinciarelli, A., Pantic, M. P., & Boulard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Comuting, 27*, 1743-1749.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics, 32*(8), 1177-1207.

Ward, N. G., & Escalante-Ruiz, R. (2009). *Using responsive prosodic variation to acknowledge the user's current state.* Paper presented at the Interspeech 2009, Brighton, UK.

Watts, O., Wu, Z., & King, S. (2015). *Sentence-level control vectors for deep neural network speech synthesis*. Paper presented at the Interspeech, Dresden, Germany.

Yamaguchi, T., Inoue, K., Yoshino, K., Takanashi, K., Ward, N. G., & Kawahara, T. (2016). *Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents*. Paper presented at the International Workshop on Spocken Dialogue Systems, Riekonlinna, Finland.

Yngve, V. H. (1970). *On getting a word in edgewise*. Paper presented at the Papers from the sixth regional meeting of the Chicago Linguistic Society Chicago.

Yu, Z., Nicolich-Henkin, L., Black, A. W., & Rudnicky, A. I. (2016). *A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement*. Paper presented at the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, USA.

Zen, H., Senior, A., & Schuster, M. (2013). *Statistical parametric speech synthesis using deep neural networks*. Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*(11), 1039-1064.

Zhao, R., Papangelis, A., & Cassell, J. (2014). *Towards a dyadic computational model of rapport management for human-virtual agent*

interaction. Paper presented at the International Conference on Intelligent Virtual Agents, Boston, USA.

Zhao, R., Sinha, T., Black, A. W., & Cassell, J. (2016). *Automatic Recognition of Conversational Strategies in the Service of a Socially-Aware Dialog System*. Paper presented at the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, USA.