Paper abstracts of Human Perspectives on Spoken Human-Machine Interaction (SpoHuMa) 2021

Freiburg, Germany

November 15–17, 2021

Revisiting robot directed speech effects in spontaneous H-H-R interactions

Omnia Ibrahim[1], *Gabriel Skantze*[2]

[1]Language Science and Technology dept., Saarland University, Germany
[2]Speech Music and Hearing dept., KTH Royal Institute of Technology, Sweden

In this paper, we investigate the differences between human-directed speech (HDS) and robot-directed speech (RDS) during spontaneous human-human-robot interactions. The interactions under study are different from previous studies, in the sense that the robot has a more similar role as the human interlocutors, which leads to a more spontaneous alternation of addressee (between human and robot). 20 conversations were extracted from a multi-party human-robot discussion corpus, where two humans are playing a collaborative card game with a social robot. Each utterance in the conversations was manually labeled according to addressee (robot or to human).The following acoustic features were extracted: fundamental frequency, intensity, speaking rate, and total utterances duration. There were significant differences between human- and robot-directed speech for speaking rate and the total utterances duration. Those results demonstrate that robot-directed speech effect is still robust when speakers spontaneously switch turns between human and robot.

Improving the Accuracy for Voice-Assistant conversations in German by combining different online ASR-API outputs

Yamini Sinha, Ingo Siegert

Mobile Dialog Systems, Institute for Information Technology and Communications,
Otto von Guericke University Magdeburg, Germany

The central issue for the wider use of speech-based technical systems is the proper recognition of speech. But as spontaneous human speech has a lot of disfluencies and variations, even state of the art ASR engines are posed with difficulties. One possibility to overcome this issue is the combination of different ASR outputs. The current paper analyses ROVER a popular ASR output combination method for spontaneous German device directed utterances distinguishing high quality clean and noisy spontaneous speech samples. Using ROVER, in this paper, a relative error reduction of about 10% could be achieved. For noisy speech data an absolute higher error rate, due to high variance in error rate among individual transcriptions, is observed.

An Assessment of Apple Siri's and Amazon Alexa's Perceived Vocal Attractiveness Based on Average F0

Alyssa Allen

Eastern Michigan University, United States

Machine learning technology has enabled voice-driven virtual assistants such as Amazons Alexa and Apples Siri to become commonplace in society. Popular voice assistants commonly default to female-sounding voices. As these computer-generated voices evolve to sound more human, the potential for female subservience stereotypes to be perpetuated increases. This study analyzed a set of responses from Siri and Alexa to see if the average fundamental frequency (F0) of each response matches the average F0 of a perceived attractive adult female voice. F0 fluctuation of responses with varying sentiment levels was also compared. Findings showed that on average, Siri and Alex- as average F0 ranges correlated with the ideal average F0 range of a perceived attractive adult female voice across fact-based, opinion-based, and emotion-based responses. This paper discusses how a correlation between virtual assistant average F0 ranges and the ideal average F0 of a perceived attractive adult female indicates vocal traits of a disembodied voice have the potential to perpetuate common gender stereotypes.

“I Like You, as a Friend”: Voice Assistants Response Strategies to Sexual Harassment and Their Relation to Gender

Luca M. Leisten[1], *Verena Rieser*[2]

[1]Radboud University, Nijmegen, The Netherlands
[2]Heriot-Watt University, Edinburgh, United Kingdom

Sexual harassment towards voice assistants continues to be prevalent with up to 10% of interactions being abusive often with sexual overtones. Voice assistants are predominantly modeled as female and often respond positively to sexual harassment, which according to UNESCO has the potential to reinforce negative gender biases and stereotypes. In the following, we evaluate alternative responses to sexual harassment and their relation to the assistants gender. In an online study, 77 participants rated the appropriateness of the assistants responses to sexual harassment while the gender of the artificial voice was manipulated and compared the ratings to appropriateness scores collected with no voice-based gender information present, i.e. text-only. Results showed an interaction between gender and the appropriateness level of a response. In particular, we were only able to confirm the previous results for male voices. We found that the perceived appropriateness changed as expected when spoken by a male voice, in accordance to previous no-voice ratings. However, we observed no clear difference in appropriateness levels when spoken by a female voice. We assume that this relationship is due to conflicting expectations regarding females responses to sexual harassment. Possibly, unrealistic stereotypical expectations towards females resulted in neither response being considered appropriate.

Designing Speech with Computational Linguistics for a Virtual Medical Assistant that uses Situational Leadership

Aryana Collins Jackson, Elisabetta Bevacqua, Pierre De Loor, Ronan Querrec

ENIB, Lab-STICC UMR 6285 CNRS, 29200, Brest, France

In emergency medical procedures, positive and trusting interaction between followers and leaders are imperative. That interaction is even more important when a virtual agent assumes the leader role and a human assumes the follower role. In order to manage the human-computer interaction, situational leadership is employed to match the human follower to an appropriate leadership style embodied by the agent. This paper explores how different leadership styles can be conveyed by a virtual agent through an analysis of utterances made by doctors and coordinators during emergency simulations. We create a corpus which comprises utterances from simulation videos of medical emergencies. Each utterance is annotated with a leadership style. After analysis involving k-means clustering, we compile easily reproducible rules that dictate how speech should appear in each leadership style for use in a virtual agent system.

Automatic Time-Continuous Prediction of Emotional Dimensions During Guided Self Help for Anxiety Disorders

Dalia Attas[1], *Stephen Kellett*[1,2], *Chris Blackmore*[3], *Heidi Christensen*[1,4]

[1]Department of Computer Science
[2]Sheffield Health and Social Care NHS Foundation Trust
[3]School of Health and Related Research (ScHARR)
[4]Centre for Assistive Technology and Connected Healthcare (CATCH)
University of Sheffield, Sheffield, UK

Low-intensity psychological interventions, such as Cognitive Analytic Guided Self-help or Cognitive Behavioural Guided Self-help, depend on the patient engaging with a manualised approach to the treatment of their mental health problem. Throughout the process, the Psychological Wellbeing Practitioner (PWP) will be observing the patients engagement and communicative behaviours especially concerning their progress. These behaviours are closely related to the patients emotional state, and more competent PWPs use their attentive listening skills to be alert and responsive to the signs of emotions, and to pick up on any treatment and implementation issues live during the session. However, this can be challenging to do and speech-based automatic analysis could be a way to aid the PWP by providing conversation-based higher-level, complex analysis. This study is a step towards such automatic session analysis and explores the automatic prediction of the PWPs and patients emotions using real self-help session audio recordings. A system for continuously predicting emotions using a dimensional approach was explored along with different classifiers and acoustic feature extraction approaches. Qualitative analysis of the emotion dimensional value tracks throughout sessions revealed different patterns depending on PWP competency and session timing (early or late in the treatment process).

A comparison of speech during a seizure narration in human-human or human-computer interactions

Nathan Pevy[1], *Heidi Christensen*[2], *Traci Walker*[3], *Markus Reuber*[4]

[1]Department of Neuroscience, The University of Sheffield, United Kingdom

[2]Department of Computer Science, The University of Sheffield, United Kingdom

[3]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, United Kingdom

Researchers exploring the predictive performance of different properties of speech in medical interactions often use data from different contexts due to the scarcity of available speech data, for example telephone, in-person, or human-computer interactions. However, people may speak differently across these different contexts. The objective of this research project was to explore whether there are differences in spoken narratives depending on whether they occur in human-human or human-computer interactions. We compared differences in speech rate, word count, pause frequency, and total pause time for patient narratives about what happened during a seizure for human-human and human-computer interactions. We found that participants in human-human interactions spoke significantly faster and said significantly more than participants in human-computer interactions, but that there were no differences in the frequency of pauses or the total time spent pausing. These findings suggest that there are differences in how people speak with a human compared to a computer and that users of medical speech technology should consider these differences when changing methods of data collection.

Advice-Giving between Young Learners in Robot-Assisted Language Learning

Hilla-Marja Honkalammi, Outi Veivo, and Marjut Johansson

Department of French Studies, University of Turku, Finland

Social robots bring new possibilities to education. This article presents an analysis of young learners interactions in robot-assisted language learning (RALL) and seeks to describe how they give each other advice in speaking situations. Especially when some kind of problem arises, the learners engage in problem-solving negotiations. The data consist of eight video-recorded learning situations using English as a foreign language (EFL), where eight pairs of children between 10 and 12 years old interact with a robot. This article presents microanalyses on advice-giving situations where the learners help each other succeed in their common task of answering the robots questions correctly. These microanalyses show that the learners give each other normative and epistemic advice. It seems that the robot creates a learning situation where the learners encounter interaction problems that may encourage problem-solving interactions between them.

Effects of disfluent machine speech on memory recall in human-machine interaction

Xinyi Chen[1], *Andreas Maria Liesenfeld*[1,2], *Shiyue Li*[1], *Yao Yao*[1]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]Radboud University, Netherlands

In recent years, voice-AI systems have seen significant improvement in intelligibility and naturalness, but the human experience when talking to a machine is still significantly different from the experience of talking to a fellow human. In this paper, we explore one dimension of such differences, i.e., the occurrence of disfluency in machine speech and how it may impact human listeners processing and memory of linguistic information. We conducted a human-machine conversation task in Mandarin Chinese using a humanoid social robot (Furhat), with different types of machine speech (pre-recorded natural speech vs. synthesized speech, fluent vs. disfluent). During the task, the human interlocutor was tested in terms of how well they remembered the information presented by the robot. Preliminary results suggest that the participants remembered the information presented in disfluent speech (surrounded by um/uh) better than the information presented in fluent speech, which is consistent with previous findings from human-human interaction. We discuss the implications of current findings and possible directions of future work.