SpoHuMa21

Proceedings of the Conference

Human Perspectives on Spoken Human-Machine Interaction

November 15–17, 2021

Freiburg im Breisgau, Germany (online)

FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

Proceedings of the FRIAS Junior Researcher Conference *Human Perspectives on Spoken Human-Machine Interaction* (SpoHuMa21), held online November 15–17, 2021.

Editors

Sarah Warchhold, University of Freiburg, Germany
Daniel Duran, Leibniz-Centre General Linguistics (ZAS), Berlin, Germany
Iona Gessinger, Saarland University, Saarbrücken, Germany
Eran Raveh, Saarland University, Saarbrücken, Germany / Hyro AI, Tel Aviv, Israel

DOI: 10.6094/UNIFR/223814

Advisory committee

Contents

Preface

This volume collects the papers presented at "Human Perspectives on Spoken Human-Machine Interaction" (SpoHuMa21), a *Junior Researcher Conference* hosted by the *Freiburg Institute for Advanced Studies* (FRIAS). SpoHuMa21 addressed interactions between humans and spoken dialogue systems, i.e., intelligent systems that receive spoken natural language input and respond with synthesized speech. The focus was not on the mere transmission of information, for example through one-turn interactions, typically with simple command-and-control systems. Rather, the goal of the conference was to gather and exchange knowledge and research approaches on the fundamental mechanisms of human speech perception and production in the context of spoken interaction with machines. To this end, SpoHuMa21 brought together junior and senior researchers from linguistics, psychology, cognitive science, sociology, computer science, and related fields to contribute their perspectives.

The development of spoken dialogue systems aims to enable human users to interact with them as naturally as possible via speech. Such spoken human-machine interfaces artificially emulate human behavior. However, research on spoken dialogue systems often focuses on machine-centered, technical aspects and is mainly based in areas such as computer science and artificial intelligence. The objective of SpoHuMa21 was therefore to highlight human-centered, linguistic issues while including perspectives from related fields.

Exploring the impact of individual differences between speakers in terms of psychological or cognitive characteristics on the way they interact with virtual interlocutors is gaining increasing attention in spoken human-machine interaction (HMI) research. The settings in which these interactions occur is also a point of interest: How does a conversation proceed with automated customer support on the phone or with an embodied social robot in the same room? Can we assume that people's attitudes towards virtual interlocutors shape the way spoken interactions take place, and that the interactions in turn shape people's attitudes towards speaking devices? Combining the insights gained from exploring real-life spoken HMI applications, for example in the medical field or in the field of computer-assisted language learning, can inform us about meaningful future directions for development.

The papers presented at SpoHuMa21 explore spoken HMI from various angles. The contributions by **Ibrahim & Skantze** (p. 6) and **Sinha & Siegert** (p. 11) examine how human speech in HMI changes depending on the addressee and what challenges are involved in understanding its variability. **Allen** (p. 17) and **Leisten & Rieser** (p. 23) investigate another aspect of human speech, namely gender-based perception and behavior differences, and how these may influence the attitude towards speaking devices. Various medical applications that leverage spoken HMI are presented by **Collins, Bevacqua, De Loor, & Querrec** (p. 29), **Attas, Kellett, Blackmore, & Christensen** (p. 35), and **Pevy, Christensen, Walker, & Reuber** (p. 40). These include seizure narration, emotion dimensions of speakers with anxiety disorders, and virtual medical assistance. Finally, **Honkalammi, Veivo, & Johansson** (p. 46) and **Chen, Liesenfeld, Li, & Yao** (p. 52) deal with cooperative aspects of spoken HMI, such as advice giving in learning processes and the effect of computer disfluency on memory recall.

SpoHuMa21 also featured invited talks by leading researchers, which are briefly introduced and summarized in the following.

Roger K. Moore is a professor of Spoken Language Processing at the University of Sheffield, UK, and holds Visiting Chairs at Bristol Robotics Laboratory, UK, and University College London, UK. He is an engineer by training, but much of his research has been based on insights from human speech perception and production. In his talk entitled "Spoken language interaction with 'intelligent' systems: Are we nearly there yet?", he discussed how we have almost gone too far with development in some areas of spoken HMI (e.g., human-sounding synthetic speech), while other areas lag behind due to the underappreciated richness and complexity of linguistic interaction (e.g., considering situatedness and leveraging world knowledge). The affordances of all components in spoken HMI should be aligned to avoid a sense of uncanny incongruence for the user. He stressed the importance of modeling as a tool to ensure that development is headed in the right direction.

Michelle Cohn is a postdoctoral fellow at the University of California, Davis, USA, and a principal investigator on an NSF Training Fellowship to explore the interaction of humans with voice assistants. In her talk with the title "Speech behavior with voice assistants: cognitive, social, and emotional factors", she reported on experimental results showing how humans adapt their way of speaking to virtual interlocutors. She demonstrated the influence of age and gender – of both the user and the device – on human speech behavior, and showed that emotional expressiveness in the device's speech and the degree of its embodiment are predictors for the occurrence of adaptation. As the capabilities of virtual interlocutors improve, the mental models users have of them will change, which in turn will have an impact on spoken HMI.

Catharine Oertel is an assistant professor at Delft University of Technology, The Netherlands, and a principal investigator of the Designing Intelligence Lab. Her research focus is on understanding and modeling human interaction to develop socially aware conversational agents, and on exploring how humans and artificial intelligence can creatively collaborate over extended periods of time. In her talk, entitled "On socially-aware conversational agents – challenges and outlook", she talked about engagement in HMI. In the presented studies, she discussed methods for measuring human engagement levels when interacting with physical robots and the influence of engagement on the interactions. She showed how robots can utilize different social behaviors to improve HMI and the impact such improved behaviors have on human interlocutors.

Karola Pitsch is a professor of Multimodal Communication, Social Interaction, and Technology at the University Duisburg-Essen, Germany. She conducts research on the topic of multimodal interaction, including human-machine interfaces. In her talk with the title "Human-robot interaction as a methodological tool for research on situated interaction", she focused on the multimodal analysis of human-robot interaction to develop dynamic interaction models (e.g., for interactions with robots in educational settings or in museums). She argued that HMI does not need to replicate human-human interaction, but should rather be inspired by it. In her opinion, the goal must be to identify small building blocks of human-human interaction that can be usefully employed in HMI.

Friederike Eyssel is a professor of Social Psychology at Bielefeld University, Germany, with a focus on gender and emotion in cognitive interaction. She is affiliated with the Center of Excellence Cognitive Interaction Technology (CITEC). In her talk entitled "A social psychological perspective on social robots", she addressed the question of whether psychological mechanisms of human-human interaction also apply to HMI. She reviewed the ways people anthropomorphize robots (e.g., giving vacuum cleaner robots a name and talking to them) and presented results showing that people indeed feel bad about ignoring a robot that, for example, would like to play a game with them. She also elaborated on how the *IKEA effect* (i.e., assigning a higher value to a product if one was involved in its construction) can play a role in technology acceptance.

Acknowledgements

We are grateful to the *Freiburg Institute for Advanced Studies* (FRIAS) for their support. Our special appreciation goes to Kyra Vogt and Attilio Ferrarese for their valuable help in planning the event. We would also like to acknowledge the *Vector Stiftung* for the project grant "Der Faktor Mensch in der Mensch-Maschine-Interaktion" ("The human factor in human-machine interaction") to DD. Finally, we would like to thank the conference advisory committee for their efforts and all participants of SpoHuMa21 for their contributions.

January 2022

Sarah Warchhold
Daniel Duran
Iona Gessinger
Eran Raveh

FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

FRIAS Junior Researcher Conference
Human Perspectives on Spoken Human-Machine Interaction
November 15–17, 2021

Revisiting robot directed speech effects in spontaneous Human-Human-Robot interactions

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

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

omnia@lst.uni-saarland.de

Abstract

In this paper, we investigate the differences between human-directed speech and robot-directed speech 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 more spontaneous turn-taking. 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 human). The following acoustic features were extracted: fundamental frequency, intensity, speaking rate, and total utterance duration. There were significant differences between human- and robot-directed speech for speaking rate and the total utterance duration. These results are in line with previous studies on robot-directed speech, and confirms that this difference holds also when the conversations are of a more spontaneous nature.

1 Introduction

Recent years have seen an increased interest in modeling communication for human-robot interaction; dynamic modeling of spoken dialogue seeks to capture how interlocutors change their speech over the course of a conversation. However, modelling conversational interaction between humans and robots is non-trivial. For multi-party conversational interaction, other aspects need to be addressed. For example, user's utterances directed to another human interlocutor should not be recognised as commands directed to a robot.

In human-machine interactions, participants have been shown to address computers/robots differently than humans. Speakers change their acoustic characteristics (e.g., raise their F0, slower speaking rate) when they are talking to a computer/robot in comparison to a human [1, 2, 3]. This is also in line with the findings of [4], although they found visual cues (e.g. eye gaze and head movement) to be more informative.

A possible explanation for the differences between robot-directed speech (RDS) vs. human-direct speech (RDS) might be that speaker adapts to the limited understanding capabilities of the robot; when speaker are aware of a speech perception difficulty on the part of the listener (e.g., due to background noise, a hearing impairment, or a different native language), speakers will naturally and spontaneously modify their speech in an attempt to make themselves more intelligible. [5]. According to the hypo- to hyper-articulation theory, those within-speaker var ations reflect the trade-off between clarity of speech (listener-oriented output) and economy of effort (talk oriented output) [6]. In this respect, goal-oriented speaking styles, such as infant- or robot-directed spee can be seen as an adjustment of the speaker's output (consciously or unconsciously) to meet the demands of their target audience or the communicative situation [7].

On the other hand, there are factors that might affect humans' perception of robots. According to the Computers as Social Actors paradigm, humans apply the same social rules used in human interactions when they interact with computer [8, 9]. Furthermore in light of recent advances of social robots, the interaction with computers (or robots) via spoken language is becoming more and more integrated into our everyday life. Social robots (e.g., Furhat) are accompanied with expressive lip movements, facial gestures, gaze and non-verbal expressions, such as breathing, filled pauses and different types of back nels, which have been shown to be easy for users to read [10] and allows for more natural and spontaneous communication between humans and robots. However it is still not clear whether such advances affect the differences between human vs robot-directed speech in real-life setting.

The present study aims for better understanding of human behavior during human-robot interactions by exploring the extent to which humans adapt their speaking style to the listener in unstructured human-human-robot interactions. A similar study was conducted by [11, 12]. However, their experiment used a voice-based device (Amazon Alexa) as the computer interlocutor. In our experiment we used a physical social robot. The interactions under study are different from previous studies, in the sense that the robot has a more equal role as the other humans, which leads to a more spontaneous alternation of addressee (human and robot). We hypothesise that a human-like social robot (Furhat) will have an effect on humans' perception of the interaction and consequently might reduce the differences between human- and robot-directed speech.

2 Methods

The data used for the analysis comes from a setting where two humans play a collaborative card game with a social robot [13]. While playing the game, the participants and the robot discuss the solution together in a symmetrical three-party dialogue. The data was collected at an exhibition in the Swedish National Museum of Science and Technology for nine days.

2.1 Participants

The interactions of 20 adult male participants was extracted from the data. The age of the players ranged between 16 and 64, with a mean of 35 years. The total conversation duration ranged from 4 to 12 minutes.

2.2 Recording setup

The interactional setting of the game is illustrated in Figure 1. Two players were seated at a large table with a multi-touch screen, opposite the Furhat robot head, which has an animated face back-projected on a translucent mask, as well as a mechanical pan-tilt neck [10]. This allows Furhat to direct the gaze using a combination of head and eye movements.

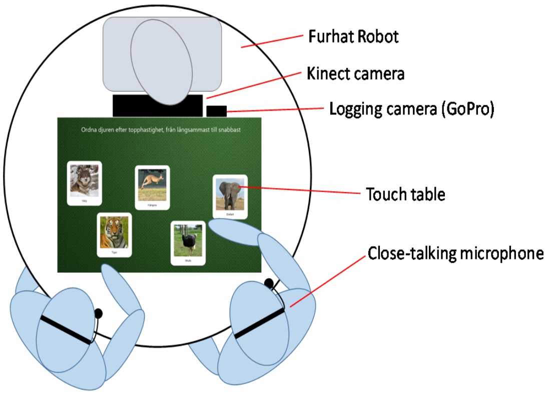


Figure 1: The setup used in the museum

Both users were wearing unidirectional headset microphones, which allowed for the recording of two separate good quality audio streams (given the noisy setting in the museum). The signal to noise ratio in the recording is around 38 dB. A Kinect camera was used to track the location and rotation of the users' heads.

2.3 Procedure

The team was seated at a table and the recordings started when they pressed a 'Start' button on the touch screen. The robot initiated the interaction by asking them for their names. Then five cards were shown on the table and Furhat (with a male voice and face) explained the game, which consists of sorting 5 cards according to sorting criterion, after which the discussion starts. An example interaction is show in Figure 2. Furhat's turn-yielding behaviour was randomly selected for each turn, both in terms of addressee and speech act (e.g., question or statement).

U-1	I wonder which one is the fastest [looking at cards]
U-2	I think this one is fastest [touching a lion card], what do you think? [looking at robot]
R	I'm not sure about this, but I think the lion is the fastest animal [looking at cards]
U-1	Okay [moving the lion]
R	Now it looks better
U-2	Yeah... How about the zebra? [looking at robot]
R	I think the zebra is slower than the horse. What do you think? [looking at U-1]
U-1	I agree
U-2	I'm not sure, the zebra has to be fast to escape the lion...
R	mhm

Figure 2: Example interaction (translated from Swedish)

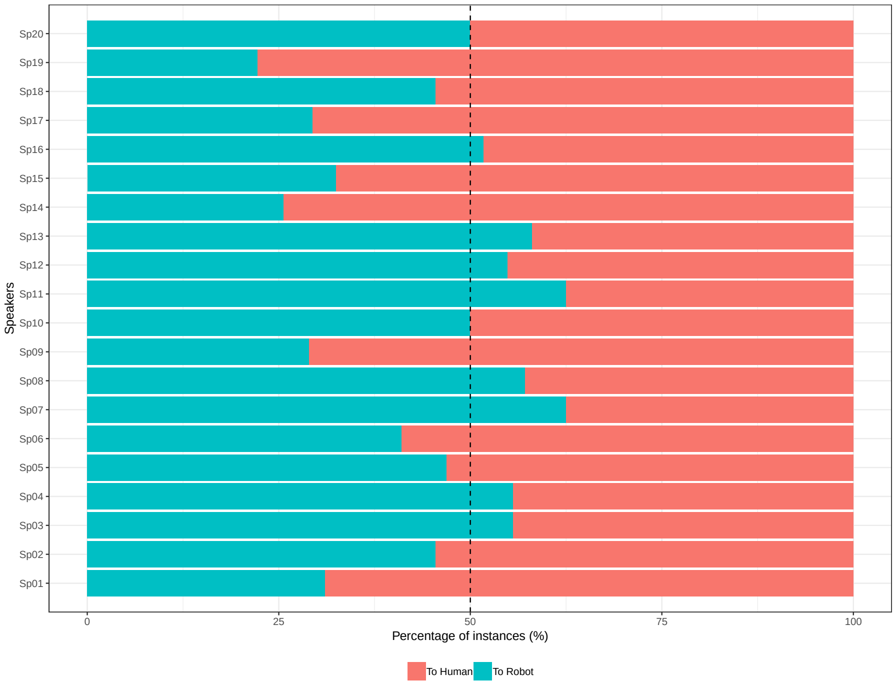


Figure 3: Percentage of utterances addressed to human vs robot for each speaker

After the task was discussed for some time, a button was shown on the table that could be pressed to reveal the solution. Furhat then commented on the solution, comparing it with his own belief (admitting mistakes or pointing out that they should have listened to him).

2.4 Data analysis

Each utterance in the conversation was manually labeled according to addressee type (robot or human) using ELAN annotation software [14]. Overlapping utterances between speakers were excluded from our analysis. The following acoustic features were extracted using Praat scripts [15]: fundamental frequency (mean, median, standard deviation, range, slope (how fast $f0$ changes during the utterance), intensity and speaking rate (vowels per minute) and total utterance duration.

For statistical analysis, Linear mixed-effects modeling was used to evaluate the effect of addressee type (human vs. robot) on the acoustic signal using R lmer package [16]. Backwards model selection procedure was applied to arrive at a final model as reported below. Our fixed effect was addressee type, while the random effect was speaker to account for the fact that different speakers might behave differently when addressing the robot vs. a human. We added utterances and number of robot-directed utterances as random factors. The final model structure was: *lmer (feature ˜ Addresses+ (1|Speaker Id)+ (1|Uttrance) + (1|RDS uttrances).*

3 Experimental results

We examined the global level differences between human-directed speech and robot-directed speech in comparison to the robot voice. This analysis helps to investigate if our speakers behave differently when speaking to a human vs. the robot in respect to the analyzed features.

Figure 3 shows the percentage of utterances each speaker used to address human vs. addressing Furhat In general the speakers talk to their human interlocutors more than Furhat. With regard to utterance duration, we found a significant difference between HDS and RDS ($Est. = -0.179, t = -2.72, p = .00683$); utterances are longer when they are addresse to robot than when addressed to human.

Intensity: 70% of the speakers talk louder to the robot in comparison to their human interlocutors. There are two potential explanations: (a) speakers' intuition that the robot might have difficulties in un-

DOI: 10.6094/UNIFR/223822

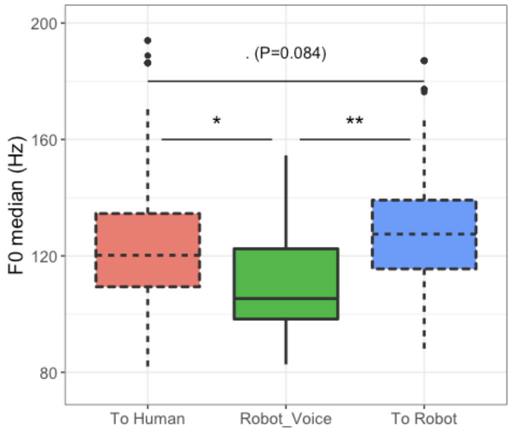


Figure 4: Fundamental frequency when speaking to human vs. speaking to robot

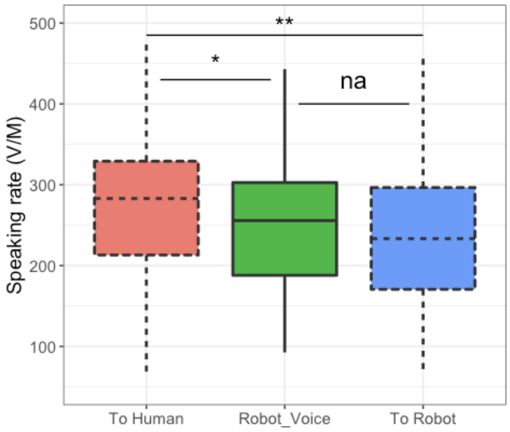


Figure 5: Speaking rate when speaking to human vs. speaking to robot

derstanding their speech in comparison to a human interlocutor, or (b) due to different distances of the interlocutors. Therefore, Intensity measures were excluded from analysis.

Fundamental frequency: Figure 4 show the results of fundamental frequency of HDS (to Human) vs. RDS (to Robot). Even though Figure 4 indicates that RDS has higher fundamental frequency than HDS, the difference is not significant ($Est. = 4.99, t = 1.75, p = .084$), while comparing RDS to the robot voice in the interactions, we found a significant difference ($Est. = -14.62, t = -3.22, p = .00214$).

Speaking rate: Figure 5 shows the speaking rate of HDS vs. RDS. As expected, speech towards the robot is significantly slower than speech towards a human ($Est. = -35.56, t = -3.18, p = .00325$). However, when comparing RDS to the robot voice in the interactions, we did not find any significant difference ($Est. = -0.454, t = -0.032, p = .974$). Those results align with literature where speakers are likely to speak slower than usual when speaking with computers and systems (e.g. Alexa) [12].

4 Discussion and conclusion

In light of the recent advances of social robots, the multiparty human-robot interaction is becoming more and more integrated into our everyday life, which might effect humans' perception of robots. Our question is whether well known speaking style (robot-directed speech) can be observed in naturalistic setting. In this study, we present a global-level analysis of human-directed speech and robot-directed speech in multi-party interactions. Our findings provide evidence for inter-speaker dynamics when speaking to a human vs. a robot. The results show that there are significant differences between human- and robot-directed speech for speaking rate and the total utterance duration. Those results demonstrate that speakers change and modulate their speech alternatively when speaking to robot vs. when speaking to humans for the sake of increasing their speech intelligibility to the robot. These results might have been partly due to the fact that the interactions took place in a noisy environment (a museum) [17]. To conclude, the robot-directed speech effect is still robust when speakers spontaneously switch turns between human and robot in naturalistic setting.

References

[1] S. Kriz, G. Anderson, M. Bugajska, and J. G. Trafton, "Robot-directed speech as a means of exploring conceptualizations of robots," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, ser. HRI '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 271–272.

[2] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *Journal of Pragmatics*,

vol. 42, no. 9, pp. 2355–2368, 2010, how people talk to Robots and Computers.

[3] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[4] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 144–151.

[5] A. Bradlow, *Confluent talker- and listener-related forces in clear speech production*. Mouton de Gruyter, 2002, pp. 241–273.

[6] Lindblom, *Explaining Phonetic Variation: A Sketch of the H&H Theory*, 1990, p. 403–439.

[7] R. Smiljanić and A. R. Bradlow, "Speaking and hearing clearly: Talker and listener factors in speaking style changes," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 236–264, 2009.

[8] B. Reeves and C. Nass, *The media equation: How people treat computers,television, and new media like real people and places.*, 1996, p. 128.

[9] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000.

[10] S. A. Moubayed, G. Skantze, and J. Beskow, "The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction," *International Journal of Humanoid Robotics*, vol. 10, no. 01, p. 1350005, 2013.

[11] E. Raveh, I. Steiner, I. Siegert, I. Gessinger, and B. Möbius, "Comparing phonetic changes in computer-directed and human-directed speech," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, P. Birkholz and S. Stone, Eds. TUDpress, Dresden, 2019, pp. 42–49.

[12] E. Raveh, I. Siegert, I. Steiner, I. Gessinger, and B. Möbius, "Three's a crowd? effects of a second human on vocal accommodation with a voice assistant," pp. 4005–4009, 2019.

[13] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 67–74.

[14] H. Sloetjes and P. Wittenburg, "Annotation by category: ELAN and ISO DCR," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.1. 08) [computer program], retrieved may 1," 2019.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[17] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

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

yamini.sinha@ovgu.de, ingo.siegert@ovgu.de

Abstract

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. In this paper ROVER, a popular ASR output combination method is applied 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% is achieved. For noisy speech data insignificant error reduction, due to high variance in error rate among individual transcriptions, is observed.

1 Introduction

Speech-based interaction with technical systems has been on a rise since the last decade [1]. Such interactions are mainly realized using a voice assistant and use an Automatic Speech Recognition (ASR) system that outputs a sequence of words or sentences spoken by identifying and converting the input speech segments into text. Human speech has a lot of variations in its pronunciation, accents, pitch, dialects, etc. [2] and developing a system to recognize speech accurately is challenging [3].

ASRs are used for several applications such as dictation systems in medical systems, the automotive industry, or day-to-day use on mobile devices [4]. Developing ASR systems for specific applications from scratch requires a large amount of data and long hours of training.

To facilitate such needs, a speech-API, available as offline or cloud-based services, is quite useful and provides a convenient interface. The choice of an ASR service depends on its performance which differs e.g. with the quality of input speech data (w/wo background noise), direct or spontaneous speech, recognition algorithms. The accuracy of ASR systems is then usually evaluated using performance metrics like word error rate (WER), to name just the most prominent one [5].

Continuous improvements have been made to reach human parity [6]. While several ASRs provide transcriptions with varying accuracy, the approach that should be discussed in the following is to improve the overall WER by combining transcriptions from multiple ASR systems having different configurations. Thereby, this contribution investigates spontaneous German interactions between humans and a modern voice assistant using a popular combination method called Recognition Output Voting Error Reduction (ROVER) [7] by combining 1-best hypothesis (transcriptions) from each of three different ASR services, namely Google Cloud speech-to-text, IBM Watson and Wit.ai from a python library called SpeechRecognition [8]. The experiments utilized German conversational speech as ASR accuracy still lags for German in comparison to English [9, 10].

The remainder of the paper is as follows: Section 2 gives some background on related work. Section 3 introduces the methods and data used. Section 4 explains the experiments. The results are shown in Section 5. Section 6 discusses the limitations to the method chosen and further improvements.

2 Background and Related Work

While many studies have been conducted on English speech samples, other languages' ASR performances are investigated seldom. The authors of [9] compared the performance of cloud-based ASR services (as used in this paper) on German conversa-

DOI: 10.6094/UNIFR/223823

tions and reported 17%-74% WER for noisy speech, while for clean speech Wit.ai achieves lowest with 8% WER. Another study investigated Sphinx4's performance on spontaneous Upper Saxon German speech [11], showing 77% WER for spontaneous speech and 51% WER for read speech. For Romanian using Google Speech, the authors of [12] reported 31% WER for a corpus of 20 YouTube videos.

The possibility of reducing the WER can be explored by further analyzing the cause of errors and also combining hypotheses from multiple ASR services. Some of the studies used methods like ROVER, BAYCOM, or other enhanced versions of ROVER, cf. [7, 13, 14]. ROVER can also be used to combine at sentence level or combine N-best hypotheses [15]. Similar work as in the current paper is reported for other languages such as [16] on Slovak using ROVER, or on Malay broadcast news [17].

3 Methodology

3.1 Performance Metrics

The recognition error is evaluated by comparing the ASR systems' hypothesis text (H) to its ground truth (R) using the following metrics:

Word error rate (WER) is calculated by evaluating the minimum edit distance. It counts the total number of deleted (D), substituted (S), and inserted (I) words in H over the total number of words (N) in the ground truth text [18]:

$$WER = \frac{S+D+I}{N} * 100 \tag{1}$$

Often WER can be more than 100%, for instance, when the length of hypothesis text is much larger than the ground truth [19].

Match error rate (MER) is the probability of a match being incorrect [19]. In contrast to WER, MER considers matching word-pairs (M) as well:

$$MER = \frac{S+D+I}{N=M+S+D+I} * 100 \tag{2}$$

Word information lost (WIL) is based on the Mutual Information (MI), which provides a measure of the statistical dependence between the input words X and the output words Y in the unordered set of I/O word pairs obtained by I/O alignment: [19].

$$WIL = 1 - \frac{H(Y|X)}{H(Y)} \tag{3}$$

Another metric that helps to determine the accuracy of ASR transcription is the confidence score (CS). It indicates the reliability of the recognition result [20]. The success of speech applications is determined by the quality of the confidence measure.

3.2 Combining Multiple ASR Hypotheses

A common method to combine several ASR hypotheses texts is ROVER, developed at NIST [7]. It aims to reduce the recognition error by exploiting differences in the nature of the errors in differebt ASR output text. ROVER produces a composite hypothesis using transcriptions from multiple ASR systems where the best word hypothesis is voted at every word instance [7]. This algorithm can be categorized into two phases, as shown in Figure 1. For N ASR systems, we have N hypothesis i.e., $ASR_1, ASR_2, .., ASR_N$.

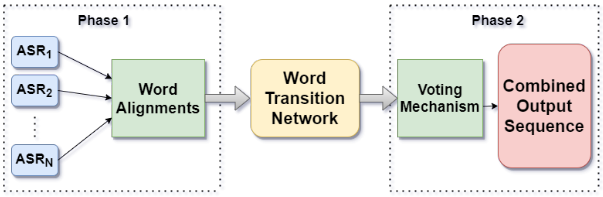


Figure 1: Combination of ASR outputs using ROVER - architecture

Phase 1 is forming a composite Word Transition Network (WTN), as shown in Figure 2, by aligning two hypotheses (for e.g., ASR_1 and ASR_2) having the lowest error rate. Alignments are done iteratively over all ASR hypothesis. Figure 2 shows an example of a composite WTN when three ASR hypothesis texts are aligned. For any occurrence of missing words during alignments, a null word transition ("@") is assigned.

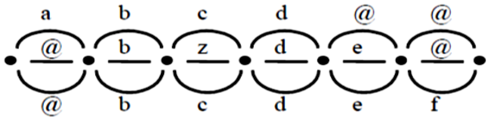


Figure 2: An example of Word Transition Network, taken from [7]. The different words are depicted by the letters a to f and z (substituted word). "@" is a null word transition signifying a missing word.

In *phase 2*, a voting mechanism is employed. A word score is calculated by the weighted sum of word confidence scores $C(w_i)$ and the number of occurrences $N(w_i)$ for each word using equation 4. Each word is assigned its word score and the highest scoring word is chosen from each set of words from the WTN to form a combined text:

$$Score(w_i) = \alpha(\frac{N(w_i)}{N_s}) + (1-\alpha)C(w_i) \tag{4}$$

The number of occurrence is normalized to unity by number of hypothesis used for combination N_s. In Figure 2, N_s is 3. The weighting factor α is applied to control the importance among $C(w_i)$ and $N(w_i)$, which ranges between 0 and 1. A lower α value means that a higher weight is given to the word confidence score.

4 Experimental Setup

4.1 Dataset

This study focuses on spontaneous Human-Computer Interaction (HCI) at two different qualities. Spontaneous speech is characterized by having scruffy grammatical phrases, self-corrections, hesitations, disfluencies, realistic turn-taking, or prosodic variations, which are quite challenging even for state-of-the-art ASR engines. Therefore, two datasets: Voice Assistant Conversation in the wild (VACW) and Voice Assistant Conversation Corpus (VACC), were selected for this study. VACW consists of unconstrained German conversations between a voice assistant and humans, comprising different types of background noise, in a public environment. VACC consists of conversations between one or more speakers and a voice assistant recorded under high-quality conditions. For further details, see [21, 22]. In the actual investigation, three sets – each consisting of 100 randomly selected audio files – of both datasets each are used.

4.2 Selected speech recognition services

This study relies on three online speech services: Google Cloud Speech API (GC), IBM Watson Speech-to-Text (IBM), and Wit.ai (WIT). They are accessible through the SpeechRecognition library using python language (for details see [8]). Audio files were sent to ASR-API server via an HTTP request. The API responds with a set of information such as transcription, sentence-level confidence score or word-level confidence score, word timestamps, for example. Other online ASR services in SpeechRecognition library do not deliver the confidence score needed for hypothesis combination and were thus omitted.

4.3 Processing pipeline

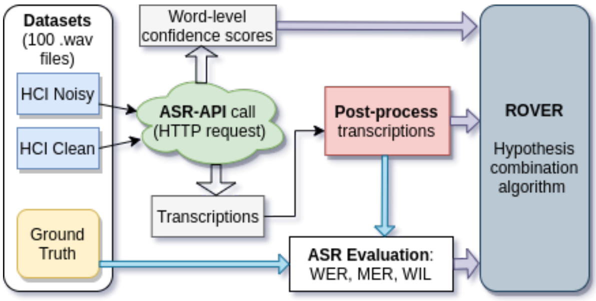

Datasets
(100 .wav files)
HCI Noisy
HCI Clean
Ground Truth
Word-level confidence scores
ASR-API call (HTTP request)
Transcriptions
Post-process transcriptions
ASR Evaluation: WER, MER, WIL
ROVER
Hypothesis combination algorithm

Figure 3: Processing Pipeline

Post-processing transforms the ASR output text to match the style of ground truth for error rate calculations. Upon evaluating error rates of all the selected files, post-processed transcriptions are chosen to combine the three ASR transcriptions, from GC, IBM and WIT, using ROVER. The ASR transcriptions used for combination are selected only if all the three transcriptions incur at least one erroneous word in all three transcriptions for a selected audio file. When all the three transcriptions are same, the final output transcription is the original transcription as there is no scope of reducing error rate further. Two ASR transcriptions, having the lowest WER, are aligned together on word-level using the Needleman-Wunsch algorithm [23] and then, further aligned with the third transcription text to form a composite WTN as shown in Figure 2. A ROVER scoring mechanism is used to vote the best word hypothesis at each word transition, as explained in section 3.2. The overall processing pipeline as shown in Figure 3, iterates over all the selected audio files.

5 Results and Analysis

5.1 ASR performance

In this section the experimental results are presented on three test set each consisting of 100 audio files of HCI clean (VACC) and HCI noisy (VACW) each. Figure 4 shows the WER for each set of the two corpora evaluating each ASR. Noisy data reports the highest error rates. On comparing all the results, IBM is the least accurate with WER ranging from 13% for VACC to 56% for VACW. While for noisy speech (VACW), the WER ranged between 19% - 56% with the worst accuracy being from IBM Watson, for clean speech (VACC) GC performed best at 4% WER.

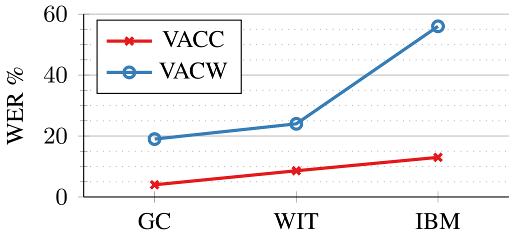


Figure 4: Word error rate (averaged over the three sets) for each corpus and each ASR engine.

An example, in Table 1, shows ”antonio meyuki” is misrecognized as it is a noun, which is probably rare in the training data. One way to recognize such words is if the use-case is specified and provides word hints, which is known as speech adaptation. Additionally, WIT recognized ”erfand” (in English, invented) as two separate words ”er fand” (in English, he found). While WER is same for WIT and IBM, the errors are not. This provides the grounds to further explore the possibility of combining such ASR transcriptions by choosing the most accurate word from each of them forming an improved transcription.

Table 1: Example of errors in ASR transcription.

Ground Truth: erfand antonio meyuki das telefon				
ASR	**Hypothesis**	**WER**	**MER**	**WIL**
GC	erfand antonio miyuki das telefon	0.2	0.2	0.36
WIT	er fand antonioki das telefon	0.6	0.6	0.84
IBM	erfand antonio um mir yuki das telefon	0.6	0.43	0.54

5.2 Combination results using ROVER

In Table 2, the ground truth is ”alexa wie”. GC and WIT recognize only the first word ”alexa” and therefore, an ”@” is placed at the second-word position because of deletion. IBM recognizes the word ”wie” correctly. The confidence score of ”@” is not predefined by ASRs, so, in this experiment, C(@) values varying between 0 and 1 were tested to find the optimal value, which is C(@) = 0.45. Ergo, combined hypothesis at α = 0 performed with 100% accuracy.

Table 2: An example of the combination of ASR hypothesis for different α values using ROVER.

Ground Truth: ”alexa wie”			
	WTN		**WER**
GC	alexa	@	0.5
word-conf	0.92	0.45	
WIT	alexa	@	0.5
word-conf	1	0.45	
IBM	alexander	wie	0.5
word-conf	0.54	0.51	
$\alpha = 0$	alexa	wie	0
$\alpha = 0.2$	alexa	@	0.5
$\alpha = 0.5$	alexa	@	0.5
$\alpha = 0.7$	alexa	@	0.5
$\alpha = 1$	alexa	@	0.5

Figure 5 shows the relative WER reduction of approx. 10% for VACC, achieving 96% accuracy with ROVER. The error rate for VACW is still relatively very high due to the high variance in WER caused by the three ASRs (GC and WIT versus IBM). The use of the ROVER method is only advantageous when original ASR WERs are closer than apart.

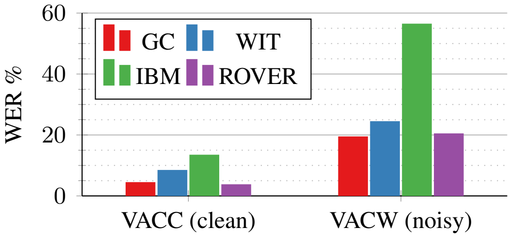


Figure 5: Improvements in ASR performance using ROVER over original transcriptions.

In Table 3, the highest relative error reduction of 9% (WER) is seen for α values between 0 - 0.2. This means, word-level confidence scores are given 80% to 100% weightage in equation 4. Whereas, giving more weight to the frequency of word occurrence (i.e., α = 0.5 - 1) yields only 4% relative error reduction for HCI Clean. As a result of the high variance in ASR error rates in HCI Noisy

Table 3: WER (average of three sets) before and after hypothesis combination for different α values.

ASR	HCI Clean	HCI Noisy
GC	0.04	0.19
WIT	0.08	0.24
IBM	0.13	0.56
ROVER with α = 0	0.033	0.2
ROVER with α = 0.2	0.04	0.25
ROVER with α = 0.5	0.068	0.28
ROVER with α = 0.7	0.08	0.29
ROVER with α = 1	0.09	0.3

audio data, ROVER's combined transcription with lowest error rate is still higher than the best performing ASR, even though the relative WER reduction is between 26% - 36%. That said, error reduction using ROVER for α values 0.5 and less provides more accurate transcriptions. It is safe to say that, confidence scores play a major role in the evaluation of ASR performance.

6 Limitations and Conclusion

Hypothesis combination using ROVER requires ASRs to provide WER with low variance to achieve any significant improvement in ASR performance. Furthermore, important scores from the ASR system during the decoding process are disregarded such as posterior probabilities of words. Since, in this experiment, APIs are used to access ASR services, only a limited set of parameters can be obtained. In addition, the selected ASRs are closed source services. This forbids information such as differences between ASR systems could be exploited in principle within such an approach in order to optimize the performance.

This paper contributes towards analyzing ASR performance for unconstrained spontaneous German conversation speech in noisy and quiet environments and further investigate the ROVER method for transcriptions combination to reduce the errors. Using ROVER, a relative word error reduction of approx. 10% using confidence scores was achieved. Combination transcription of noisy speech data has a significantly high error rate, due to high variance in error rate among individual transcriptions. While we achieved reduced error rates, further investigations could be made to use better scoring methods.

7 Acknowledgment

This research has been funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060A).

References

[1] B. Kinsella, “Nearly 90 million u.s. adults have smart speakers, adoption now exceeds one-third of consumers.” voicebot.ai, 2020, [Online; posted 28-Apr-2020]. [Online]. Available: https://perma.cc/336P-2C77

[2] B. T. Meyer, T. Brand, and B. Kollmeier, “Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes,” *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 388–403, 2011.

[3] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, “End-to-end asr: from supervised to semi-supervised learning with modern architectures,” 2019.

[4] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimedia Tools and Applications*, vol. 80, pp. 9411–9457, 2021.

[5] R. Errattahi, A. El Hannani, and H. Ouahmane, “Automatic speech recognition errors detection and correction: A review,” *Procedia Computer Science*, vol. 128, pp. 32–37, 2018.

[6] W. Xiong, L. Wu, J. Droppo, X. Huang, and A. Stolcke, “The Microsoft 2017 Conversational Speech Recognition System,” in *Proc. of the IEEE ICASSP*, 2018, pp. 5934–5938.

[7] J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” in *Proc. of IEEE ASRU'97*, 1997, pp. 347–354.

[8] *SpeechRecognition library for online and offline APIs for performing speech recognition.*, accessed on 05/11/2019. [Online]. Available: https://pypi.org/project/SpeechRecognition/

DOI: 10.6094/UNIFR/223823

[9] I. Siegert, Y. Sinha, O. Jokisch, and A. Wendemuth, "Recognition Performance of Selected Speech Recognition APIs–A Longitudinal Study," in *International Conference on Speech and Computer*, 2020, pp. 520–529.

[10] V. Silber Varod, I. Siegert, O. Jokisch, Y. Sinha, and N. Geri, "A cross-language study of speech recognition systems for english, german, and hebrew," *OJAKM*, vol. 9, 2021.

[11] R. Herms, L. Seelig, S. Münch, and M. Eibl, "A corpus of read and spontaneous upper saxon german speech for asr evaluation," in *Proc. of the 10th LREC*, 2016, pp. 4648–4651.

[12] B. Iancu, "Evaluating google speech-to-text api's performance for romanian e-learning resources." *Informatica Economica*, vol. 23, no. 1, 2019.

[13] A. Sankar, "Bayesian model combination (baycom) for improved recognition," in *Proc. of the IEEE ICASSP*, 2005, pp. I–845.

[14] M. AlemZadeh, K. Abida, R. Khoury, and F. Karray, "Enhancement of the rover's voting scheme using pattern matching," in *International Conference on Autonomous and Intelligent Systems*. Springer, 2012, pp. 167–174.

[15] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk asr voting strategies," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] M. Lojka and J. Juhár, "Hypothesis combination for slovak dictation speech recognition," in *Proc. of the IEEE ELMAR*, 2014, pp. 1–4.

[17] Z. A. Khalaf, T.-P. Tan, L.-P. Wong, and B. H. Ahmed, "A system combination for malay broadcast news transcription," *Jurnal Teknologi*, vol. 77, no. 19, 2015.

[18] M. Thoma, *Word Error Rate Calculation.*, accessed on 27/04/2020. [Online]. Available: https://martin-thoma.com/word-error-rate-calculation/

[19] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Proc. of ICSLP*, 2004.

[20] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[21] I. Siegert, ""alexa in the wild"–collecting unconstrained conversations with a modern voice assistant in a public environment," in *Proc. of the 12th LREC*, 2020, pp. 615–619.

[22] I. Siegert, J. Krüger, O. Egorow, J. Nietzold, R. Heinemann, and A. Lotz, "Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA," in *Proc. of the 11th LREC*, 2018.

[23] K. Erciyes, "Sequence alignment," in *Distributed and Sequential Algorithms for Bioinformatics*. Springer, 2015, pp. 111–133.

An Assessment of Apple Siri and Amazon Alexa's F0 Values as Related to Vocal Attractiveness

Alyssa Allen[1]

[1]Eastern Michigan University, United States

aallen76@emich.edu

Abstract

Popular voice assistants like Amazon's Alexa and Apple's Siri commonly have a default female-sounding voice. As these devices become more commonplace in society, the potential socio-indexical proprieties perpetuated by the voices should be considered, especially related to gender ideologies and stereotype perpetuation.

This study analyzed Alexa and Siri's fundamental frequency (F0) values for responses to fact-based, opinion-based, and emotion-based prompts to see if there existed a correlation between the results and average F0 values of a perceived attractive adult female voice (220Hz - 260Hz, noting that higher F0 values in this range were perceived as more attractive).

Findings showed that Siri's F0 values were significantly higher than Alexa's F0 values overall. Responses to fact-based prompts were statistically lower than responses to emotion-based responses. There was no statistical significance between emotion- and opinion-based responses.This paper serves as a starting point in a larger conversation regarding what social characterizations virtual assistant voices index and how machine voices have the potential to reflects societal gender expectations and stereotypes.

1 Introduction

People commonly interact with voice-generating machine-learning programs through virtual assistants such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google Home. Virtual assistants are defined in this paper as any voice-based assistants that do not have any physical representation or manifestation attached to them. The assistant is completely operated by machine-learning capabilities and has the goal of providing humans with information or to perform tasks virtually.

Virtual assistants (such as Amazon's Alexa) commonly employ female-sounding voices as the default setting. With consideration to the role virtual assistants play in the lives of human users, this paper aims to better understand what type of human voice the virtual assistant voice embodies based on F0 and how that characterization fits within (in this case) US societal gender expectations or stereotypes. As a starting point into a more complex conversation about the social characteristics indexed by a virtual assistant machine-generated voice, this paper focusec on assessing if a correlation between F0 and vocal attractiveness exists for Siri and Alexa. All results were analyzed with the understanding that factors such as voice clarity, intelligibility, and effectiveness are crucial influences in voice design and linguistic factors such vowel length, F1, and F2 influence vocal attractiveness perceptions.

1.1 Female Subservience Stereotypes

Popular culture in the United States maintains a stereotype that women are best suited for assistant-type roles, and that the best assistants are women. Women in turn, are more likely to be perceived as assistants rather than leaders [1]. Stereotypes about women being subservient also include women being perceive as ”pushy” or ”bossy” when placed into a socially dominant role, such as a boss or team leader. Men are stereotypically seen as effective leaders while women are considered more apt at secretarial duties [2]. Female subservience stereotypes will be defined in this paper, as the stereotype that women are better suited for subservient roles, especially if a male figure is present.

As it pertains to virtual assistants, it is noteworthy that popular virtual assistants, including Apple's Siri and Amazon's Alexa have distinctively female-sounding voices set as the default for users. It should also be noted that the latest version of Apple's Siri (as of March 2021) does not assign a default English-speaking voice for Siri. Instead of the previous female-sounding default, users select their preferred Siri voice when setting-up the device [3]. This is a recent change though and does not indicate whether users still associate Siri with the original female-sounding voice. Popular virtual assistants overwhelmingly being assigned a female-sounding voice is interesting when considering if virtual assistant voices have the power to perpetuate female subservience stereotypes.

In addition, it should be considered that the type of female or femininity being portrayed within the scope of female subservience stereotypes can differ. One possible version of this stereotype is one that associates attractiveness with female subservience [4]. This paper will focus on starting to gain an understanding of what type of femininity is portrayed by virtual assistant voices.

1.2 Early Versions of Virtual Assistants

In early versions of Siri, the virtual assistant was "imbued with an overt personality, [could] carry on non-task related interactions, such as telling jokes, and [had] distinct identity characteristics" [5]. Early versions of Siri and Alexa were flirtatious and employed responses that reflected the female subservience stereotype, including that a female assistant should be flirty, openly sexual, and submissive. For example, the only sexual request Siri would refuse early on was if Siri was asked outwardly to engage in sexual intercourse. The response would include Siri saying "You have the wrong type of assistant" [6]. This response evades the sexual comment and alludes to there being a type of female assistant that should willingly accept a sexual proposition.

Companies such as Apple and Amazon have since worked to minimize utterance content that was a higher risk for being encoded with unconscious stereotypes about how a female-sounding voice should respond via submissive and flirtatious language [7]. One change involved replacing flirtatious responses with more neutral cases such as "I'm sorry, I didn't understand that" or by having the virtual assistant not respond at all.

1.3 Gender Bias in Machine Learning

Virtual assistants being commonly programmed with a female-sounding default voice suggests a connection between societal gender stereotypes and unconscious gender bias in machine learning. Gender bias in this paper is understood as a machine-language system learning to favor one particular gender over another or to exhibit stereotypical language features of a gender.

A 2019 analysis of the technology companies that develop virtual assistants show that only betweer 10 and 15 percent of researchers on development teams were women [8]. A limited female perspective in the early stages of programming may lead to conversations around sexuality and perceived femininity not happening at the necessary scale in order to anticipate the learning of stereotypical responses or the types of sexist remarks that could be said to the virtual assistant by the user. It should also be noted that how feminine the voice of the virtual assistant or automated voice sounds is a choice made at least in part by the people building the programs.

Regardless of potential linguistic implications of creating female-voiced virtual assistants, there are corporate motivations for choosing a female voice such as marketability and intelligibility [9]. The intention or motivation may not consciously be to reinforce female subservience stereotyping, but the linguistic qualities of virtual assistant voices have potential to play into or even reinforce societal expectations around gender given the role of the virtual assistants in society.

1.4 F0 and Vocal Attractiveness

Attractiveness is often tied to conversations about femininity. As a starting point in assessing the type or types of femininity portrayed by virtual assistants, vocal attractiveness will serve as a starting point. Vocal attractiveness is highly subjective based on factors, such as the hearer's preferences, age, culture, and sexuality ([10],[11],[12],[13]). There are a multitude of linguistic factors that dictate and shape vocal attractiveness, such as vowel shape, timbre, and creaky voice ([14],[15],[16]). Given previous studies connecting F0 to vocal attractiveness, the following experiment will first analyze the F0 of Siri and Alexa.

In a 2011 study by Borkowska and Pawlowski, polish-speaking participants were asked to rank fema sounding voices in terms of attractiveness. Findings showed that the listener's perceived attractiveness of a voice fit a bell curve model where F0 values between 220 Hz to 262 Hz was typically viewed as associated with attractive female voices. Voices with F0 values between 262 Hz to 282 Hz were still attractive but starting to decline in perceived attractiveness, meaning that the impact of raising the pitch started to have a reverse effect on a higher perceived attractiveness rating. Voices with F0 ranges below 220 Hz were perceived as least attractive. Noting that the lower the F0 under this point, the less attractive the voice was perceived [17]. When the voice has an average F0 between 220Hz to 262Hz, the voice is more likely to be viewed as feminine, youthful, and flirtatious [18].

A similar range of attractive F0 was found by Feinberg and colleagues in a 2008 perception study done with Scottish-English. Results were discussed in relative terms. Based on a sample of adult female voices, a low F0 at 200 Hz, an average F0 at 220 Hz, and a high F0 at 241Hz was selected. Each starting F0 value was raised and lowered by factors of 20Hz, keeping all other formants stable. Participants were tasked with rating vocal attractiveness for each voice compared to its raised and lowered forms.

Findings showed that male listeners ranked higher F0 values as more attractive for each starting F0. The study did not test F0 values above 261Hz, so it is not clear if there is a F0 value higher than 261Hz that would be perceived as less attractive. For female listeners, raised F0 values are also preferred, but the difference in preference when the voice was raised from 200 to 220Hz was significantly larger than when F0 was raised from 240Hz to 261Hz. This suggests there will be a point at which raising the F0 will not increase perceived attractiveness [19].

While this study does not identify a defined attractiveness range, it demonstrates that higher F0 values were perceived as more attractive female voices for both male and female listeners, at least up to the tested range. Feinberg and colleague's study also suggests that there may be a limit (around 260Hz) at which point raising the F0 value does not equate to an increase in perceived attractiveness.

Because an F0 range of 220 Hz - 260 Hz in both cases appears to be perceived as particularly attractive in Polish and Scottish English, this range will be used as a reference point for this study. Both of the aforementioned studies work with specific cultures and may not be completely transferable to a study with American English, but should be considered a reliable reference point.

The following experiment aims to gain a better understanding of the average F0 values of Siri and Alexa's English-speaking voices and determine if there is a correlation between the F0 of virtual assistants and a perceived attractive adult female voice. Findings will be analyzed as it relates to what type of femininity the virtual assistants' voices index.

2 Research Methods and Stimulus

For this study, Siri and Alexa were asked the same set of 15 prompts. These prompts included five fact-based prompts, five emotion-based prompts, and five opinion-based prompts. The prompts were as follows:

1. What is the weather today?
2. How do I get to my home?
3. Who is the current president?
4. Who is the highest paid actor?
5. Who is the highest paid actress?
6. I'm sad.
7. You're pretty.
8. I'm lonely.
9. Will you be my friend?
10. Will you always be there to help me?
11. Do you think men and women are equal?
12. What should I have for dinner?
13. Do you think I should become an influencer?
14. What do you think is a good gift?
15. What should I name my cat?

Prompt types were defined as the following: fact-based prompts (1-5) require factual answers; emotior based prompts (6-10) require the virtual assistant to have a basic understanding of human emotion; opinic based prompts (11-15) require the virtual assistant to provide opinions on current social issues or the speaker's personal life choices. Prompt type was varied in order to not outweigh any one particular utterance type to get a sense of the assistants' sensitivity to context.

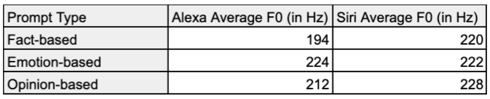
Responses were recorded using Praat software. Each response was analyzed for minimum F0, maximum F0, and average F0. All recordings were then cumulatively analyzed for average minimum F0, average maximum F0, and overall average F0.

The F0 ranges of Siri and Alexa were then analyzed against that of an attractive adult female voice F0 range. As discussed in the previous section, the F0 range of a perceived attractive female voice for this study will be 220Hz to 260Hz. The results will be analyzed with the understanding that higher F0 values are perceived as more attractive than lower F0 values, with possible attenuation or reversal of this trend above 260Hz.

3 Results

For Siri, fact-based questions resulted in an average minimum F0 of 144Hz, an average maximum F0 324Hz and an overall average F0 of 220Hz. Emotion-based questions resulted in an average minimum F0 of 159Hz, an average maximum F0 of 328Hz and an overall average F0 of 222Hz. Opinion-based questions resulted in an average minimum F0 of 144Hz, an average maximum F0 341Hz and an overall average F0 of 228Hz. Across categories, Siri's overall average F0 was 235Hz.

For Alexa, fact-based questions resulted in an average minimum F0 of 125Hz, an average maximum F0 of 353Hz and an overall average F0 of 194Hz. Emotion-based questions resulted in an average minimum F0 of 107Hz, an average maximum F0 of 264Hz and an overall average F0 of 224Hz. Opinion-based questions resulted in an average minimum F0 of 135Hz, an average maximum F0 324Hz and an overall average F0 of 212Hz. Across categories, the overall average F0 was 210Hz.



Prompt Type	Alexa Average F0 (in Hz)	Siri Average F0 (in Hz)
Fact-based	194	220
Emotion-based	224	222
Opinion-based	212	228

Figure 1: Average F0 per prompt category.

Shown in Figure 1 (above), Siri and Alexa's average F0 values in the fact-based category were the lowest compared to emotion-based and opinion-based F0 values. Fact-based response F0 values were significantly lower than emotion-based response F0 values by an estimated -22.5 Hz. There is no statistical significance between the F0 values for emotion-based and opinion-based responses.

Of note, a number of responses to emotion- and opinion-based prompts included facts or directives similar to the content found in responses to fact-based prompts, indicating that the utterance content was not a determining factor in F0 per category. For example, Siri's response to the emotion-based promp ”Will you always be there to help me?” yielded an F0 of 287Hz. Utterance content was as follows: ”I'm here to help! Get to know Siri at Apple dot com.” Siri's average F0 value remained high, even though a directive comprised the second part of the response.

Regarding the female vocal attractiveness F0 rang (220Hz to 260Hz), Siri's average F0 for responses to fact-based prompts correlated with the low end of the reference range at 220Hz. Alexa dropped below the reference range in the fact-based category with an F0 value of 194Hz.

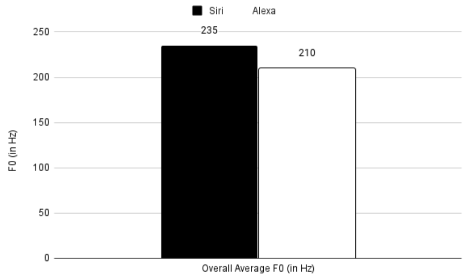


Figure 2: Overall average F0 values for Siri and Alexa across prompt categories.

As seen in Figure 2 (above), Siri and Alexa's overall average F0 values for the study. Siri has a higher overall F0 average than Alexa's overall average F0, at 235Hz and 210Hz respectively. The difference between Siri's F0 average and Alexa's F0 average was statistically significant with a p value ¡ 0.01.

While the average F0 of Siri could have been programmed to be higher, and potentially be perceived as even more attractive, 235Hz correlated wit the reference range. Alexa's F0 value fell slightly below the reference range, but also could have been programmed to be lower in F0.

4 Discussion

Across all categories tested, Siri's response F0 values were significantly higher than Siri's response F0 values. Therefore, even if the F0 value fell below the reference range of 220 - 260Hz, it is likely that Siri will be perceived as more attractive than Alexa before considering any other formants or vocal qualities.

This study also determined that the F0 of responses to fact-based prompts were significantly lower than the F0 values of responses to emotion-based and opinion-based prompts. Empirically, this seemed to be, at least in part, due to the virtual assistant reading directly from a source to answer fact-based questions. For example, responses from Siri and Alexa to fact-based prompts commonly included "According to [source]..." before providing the information required by the prompt.

As discussed previously, popular virtual assistant voices commonly having a female-sounding voice set as the default could by itself potentially reinforce a female-subservience stereotype given the role virtual assistants have in society. This study provides an initial finding in a larger conversation around the type of femininity being portrayed by the virtual assistant voices and motivates further questions about the social characteristics potentially indexed by disembodied machine voices.

5 Considerations and Future Research

Considerations for this study include expanding the number of questions per category to create a more robust data set and controlling for utterance length.

Future research in this topic could include repeating this study with other popular virtual assistants and analyzing user perceptions directly. Results from the user perception study could then be compared to the findings in this paper.

This study assessed only English-language responses from Siri and Alexa. Future research could also expand into other languages such as French and Spanish. Tests done in other languages will need to consider beauty standards and stereotypes present in cultures predominantly using the focus languages.

6 Conclusion

The results suggest that although certain phrases explicitly perpetuating female subservience stereotypes were removed from current versions of virtual assistants, there is a connection between virtual assistant F0 and the F0 range of a perceived attractive adult female used in this study.

Overall, the results from this study help to begin dissecting questions around how these disembodied machine-generated voices take up space in our human and naturally social society, including its potential ability to impact human perceptions or shape and reinforce stereotypes.

This paper serves as a preliminary study to begin a deeper and more nuanced conversation about the connection between the vocal qualities of disembodied machine-driven voices and social implications such as unconscious stereotype reinforcement.

References

[1] M. Specia, "Siri and Alexa Reinforce Gender Bias, U.N. Finds." *The New York Times*, May 2019. [Online]. Available: https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html

[2] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices," *Journal of Applied Social Psychology*, vol. 27, no. 10, pp. 864–876, May 1997. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.1997.tb00275.x

[3] Alison DeNisco Rayome, "In iOS 14.5, Apple adds new voices to make Siri sound more like you," *CNET*, Apr. 2021.

[4] J. A. Fullerton and Kendrick, Alice, "Portrayal of men and women in U.S. Spanish-language television commercials," *Journalism and Mass Communication Quarterly*, vol. 77, no. 1, pp. 128–142, 2000.

[5] A. L. Guzman, "Voices in and of the machine: Source orientation toward mobile virtual assistants," *Computers in Human Behavior*, vol. 90, pp. 343–350, Jan. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0747563218303844

[6] L. Fessler, "We tested bots like Siri and Alexa to see who would stand up to sexual harassment." *Quartz*, 2017.

[7] C. Gartenberg, "Siri is getting a new voice in iOS 13." *The Verge*, Jun. 2019. [Online]. Available: https://www.theverge.com/2019/6/3/18650906/

[8] UNESCO and Equal Skills Coalition., "I'd blush if I could: closing gender divides in digital skills through education." UNESDOC, Tech. Rep., 2019. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

[9] K. Schwab, "The real reason Google Assistant launched with a female voice: biased data." *Fast Company*, Sep. 2019. [Online]. Available: https://www.fastcompany.com/90404860/

[10] v. R. Bezooijen, "Sociocultural aspects of pitch differences between Japanese and Dutch women," *Language and speech*, vol. 38, no. 3, pp. 253–265, 1995, place: Los Angeles, CA Publisher: SAGE Publications.

[11] S. M. Hughes and N. E. Miller, "What sounds beautiful looks beautiful stereotype: The matching of attractiveness of voices and faces," *Journal of Social and Personal Relationships*, vol. 33, no. 7, pp. 984–996, Nov. 2016. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0265407515612445

[12] A. Mulac and H. Giles, "'Your're Only As Old As You Sound': Perceived Vocal Age and Social Meanings," *Health Communication*, vol. 8, no. 3, pp. 199–215, Jul. 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327027hc0803_2

[13] M. Zuckerman and R. E. Driver, "What sounds beautiful is good: The vocal attractiveness stereotype," *Journal of Nonverbal Behavior*, vol. 13, no. 2, pp. 67–82, 1988. [Online]. Available: http://link.springer.com/10.1007/BF00990791

[14] M. Babel, G. McGuire, and J. King, "Towards a More Nuanced View of Vocal Attractiveness," *PLoS ONE*, vol. 9, no. 2, p. e88616, Feb. 2014. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0088616

[15] K. Wu and D. G. Childers, "Gender Recognition from Speech, I: Coarse Analysis; II: Fine Analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, p. 1828, 1991.

[16] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982, place: United States.

[17] B. Borkowska and B. Pawlowski, "Female voice frequency in the context of dominance and attractiveness perception," *Animal Behaviour*, vol. 82, no. 1, pp. 55–59, Jul. 2011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0003347211001394

[18] D. E. Re, J. J. M. O'Connor, P. J. Bennett, and D. R. Feinberg, "Preferences for Very Low and Very High Voice Pitch in Humans," *PLoS ONE*, vol. 7, no. 3, p. e32719, Mar. 2012. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0032719

[19] D. R. Feinberg, L. M. DeBruine, B. C. Jones, and D. I. Perrett, "The Role of Femininity and Averageness of Voice Pitch in Aesthetic Judgments of Women's Voices," *Perception (London)*, vol. 37, no. 4, pp. 615–623, 2008, place: London, England Publisher: SAGE Publications.

[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], and [19].

FRIAS Junior Researcher Conference
Human Perspectives on Spoken Human-Machine Interaction
November 15–17, 2021

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

luca.leisten@ru.nl, v.t.rieser@hw.ac.uk

Abstract

Current 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 study, we evaluated 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 response category. We found that the perceived appropriateness changed 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 stereotypical expectations regarding women's responses to sexual harassment – where neither response is considered appropriate.

1 Introduction

A recent report by UNESCO raised the question whether voice assistants' replies such as "I'd blush if I could" are an appropriate response to sexual harassment [1]. Voice assistants are artificial agents that communicate using speech. They are often designed to have female voices and names and act subservient [1–3]. According to the UNESCO, a wide variety of problems result from this dominance of female-only voice assistants, including the reinforcement of gender stereotypes and biases, the perception of females as tolerant of poor treatment, and the normalisation of harassment [1].

Indeed, sexual harassment (i.e., unwanted behavior of a sexual nature [4]) is a prevalent problem in interactions with voice assistants, with numbers reported from 5 [5] to 10% [6]. Voice assistants themselves are unlikely to experience harm through this form of gender based violence. However, abuse should still be discouraged as previous research found that human-machine interaction can transfer to human-human interaction and there is thus the possibility that this behaviour is promoted towards people [7].

Until recently, voice assistants often playfully deflected abuse or even responded positively [8]. Similar results were found by [5], where 22% of responses were labeled 'positive', including flirting, playing along or joking. In a follow-on study, [9] evaluated the "perceived appropriateness" of responses of current conversational systems to certain types of abuse using crowd-based evaluation of text. Their results showed that polite refusal was found to be most appropriate while flirtation and retaliation were perceived least appropriate [9].

In this research, we investigate the influence of the interlocutors' gender on what response is deemed to be appropriate. Previous research on human-human conversations found that the perceived appropriateness of an utterance in emotionally charged contexts, such as abuse, is influenced by gender – possibly due to gender role stereotypes and gender expectations [10–15]. Similarly, research in human-robot interaction investigating gender stereotypes and gender biases found that stereotypes are also applied to robots [16–19]. Appropriateness might thus also be influenced by both the gender of the voice assistant as well as by the gender of the participant. In the

DOI: 10.6094/UNIFR/223817

following study, we investigate whether the perceived appropriateness of responses to sexual harassment of voice assistants is influenced by the gender of the voice assistant, participants' gender, and the response category.

2 Data collection

2.1 Sample

We conducted an online study with 77 (57% male, 43% female, $m_{\text{age}} = 33.5$, $SD_{\text{age}} = 11.4$) crowd-working participants using Prolific [20].[1] Participants were native English-speakers from the United Kingdom (53%), the USA (35%), Australia (6%), New Zealand (3%), and other countries (3%).

2.2 Methodology

Participants were asked to rate the social appropriateness of eight audio recorded responses (e.g. "I like you, as a friend.") to sexually sensitive prompts (e.g., "Do you want to kiss me?"). The text stimuli were collected by [5, 9]. The authors collected abusive utterances from users and used these to sample responses from a range of state-of-the-art voice assistants and chat-bots. The responses were annotated into 14 response categories and rated on appropriateness from crowd-workers. We selected a sub-set of the collected responses, where half of the responses belonged to the category labelled as 'polite refusal' and half as 'flirtation'. 'Polite refusal' includes answers such as "That is not something I feel compelled to answer", while 'flirtation' entails answers like "In the cloud no one knows what you're wearing". In [9], these two categories were on opposite ends of the spectrum: 'Polite refusal' was perceived highly appropriate whereas 'flirtation' lowly appropriate by their crowd-workers.

We then varied the gender of the assistant giving that response, using two male and two female British-English synthetic voices from Microsoft Word's [21] Text-to-Speech feature. Each participant listened to eight prompts, presented in pairs of two. The presentation of prompts and voices was counterbalanced. Participants were asked to rate the social appropriateness on a user defined scale, in comparison to a reference answer labeled with an appropriateness score of 100. This methodology is also known as 'magnitude estimation' and was found to produce more reliable user ratings than commonly used Likert scales [9, 22].

3 Results

We calculated Cronbach's alpha for the response categories 'polite refusal' and 'flirtation'. Cronbach's alpha was .56 for 'polite refusal' and .51 for 'flirtation'. The appropriateness ratings were normalized on a scale of 0-1 to make the results comparable to [9]. Pearson's correlations were calculated between all study measures and can be seen in Table 1.

Table 1: *Means*, *Standard deviations* and correlations of all study measures.

Variable	*M*	*SD*	1	2	3
1. Perceived appropriateness	0.34	0.22			
2. Response category	1.50	0.50	-.05		
3. Assistant's gender	1.50	0.50	.04	.25**	
4. Participant's gender	1.57	0.50	-.03	.00	.00

To calculate the interactions, we ran a linear mixed effects model with perceived appropriateness as the dependent (continuous) variable and fixed effects for the factors of gender (sum-to-zero coded, male coded as -1, female as 1), response category (sum-to-zero coded, 'flirtation' coded as -1, 'polite refusal' as 1), and participant's gender (sum-to-zero coded, male coded as -1, female as 1). We followed the advice by [23] to use a maximal random-effects structure. Therefore, the repeated measures nature of the data was modeled by including a per-participant random intercept and a random slope for gender, response category, and their

[1] Due to the analytic procedure, an a priori power analysis was not possible, as simulation-based sample size calculations for mixed models require previous data, which were not yet available. Therefore, a convenience sample of 80 participants was recruited of which 3 participants were excluded due to failed attention checks. The analyzed sample is a sub-set of a larger data set.

DOI: 10.6094/UNIFR/223817

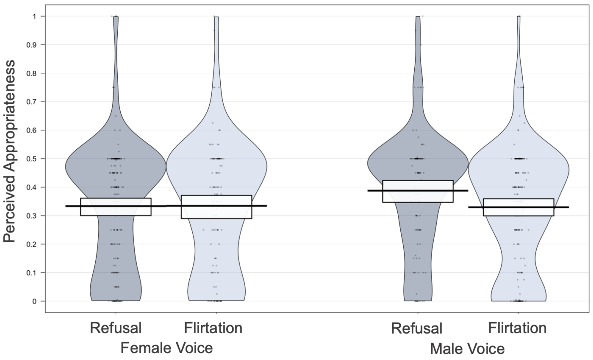


Figure 1: Pirate plot of appropriateness ratings in dependency of voice assistant's gender and response category. The plot shows the raw data points, distributions, means (indicated through the solid lines), and 95% intervals (indicated through the boxes).

interaction. Additionally, all possible random correlation terms of the random effects were included.

The model showed no significant effect of voice assistant's gender (*Estimate* = -0.017, *SD* = 0.009, *F*(1, 73.833) = 3.948, *p* = .051), response category (*Estimate* = -0.011, *SD* = 0.009, *F*(1, 74.126) = 1.229, *p* = .271), nor participant's gender (*Estimate* = -0.009, *SD* = 0.014, *F*(1, 74.668) = 0.418, *p* = .520). However, there was a significant two-way interaction between voice assistant's gender and response category (*Estimate* = -0.018, *SD* = 0.009, *F*(1, 73.671) = 4.169, *p* = .045), indicating a significant effect of response category for male voice assistants, but not for female voice assistants, see Figure 1. For male voice assistants, the perceived appropriateness changed according to the previously found appropriateness level of the response categories. Hence, polite refusal responses were perceived as highly appropriate while flirtatious responses were perceived as lowly appropriate. Surprisingly, for female voice assistants this pattern did not occur.

4 Discussion

We present the first study on how the perceived appropriateness of a voice assistant's response to sexual harassment changes with the interlocutor's gender. Our results provide first evidence that the perceived appropriateness of voice assistants' responses to sexual harassment differs between male and female voice assistants. This effect may originate from conflicting gender role beliefs and gender expectations regarding female responses to sexual harassment. Females in our society face unrealistic standards and expectations [24, 25]. These standards might have resulted in neither response being perceived as appropriate, as potentially neither refusal nor flirtation were stereotypically considered appropriate response strategies for female voice assistants that face harassment. Further research is needed in order to understand why for female voice assistants the content of a response did not seem to affect the perceived appropriateness.

Ultimately, our results indicate that the gender of a voice assistant needs to be considered when developing future response strategies to sexual harassment. Response strategies might need to be adjusted to the voice assistants' gender, in order to develop appropriate, assertive, and discouraging responses towards harassment.

4.1 Limitations

Limitations of the study include that the used voices were less natural than voices used by commercial voice assistants. Second, due to our analytic procedure no a priori power analysis was possible, which might have resulted in an under-powered study. Third, the extent to which the used voices were perceived as stereotypical female or male could have biased the ratings and should be assessed in future studies. Note that 'gender-less' voices are in general not considered to be a possible solution [3, 26]. Fourth, participants were not asked for prior exposure to voice assistants, which might have been a confounding factor. Lastly, the magnitude estimation might have introduced a bias to participants, as the labeling of reference answers with a score of 100 might have evoked the impression of 100 being the highest appropriateness score. This is reflected through the raw data, as ratings below 100 were given more often than above 100. This is potentially problematic, as the reference answers belonged to a medium appropriately response category [9] and were expected to be perceived as less appropriate as responses of the ’polite refusal’ category.

4.2 Future directions

Previous research [9] found participants’ age and the severeness of abuse to affect appropriateness ratings. These variables should therefore be included in follow-up studies. Additionally, following the recommendation of [9], it would be interesting to assess the perceived appropriateness of responses to sexual harassment in live interactions with voice assistants rather than using recordings, since actively being involved in the conversation could potentially change the perception. However, [27] made a first step into this direction asking the subjects to ‘act’ abuse. While this is not only problematic from an ethical point of view (participants did report to feel uncomfortable), it also means that the motivation for abuse was not genuine with a snowball effect on response ratings. In a recent study, [28] report an evaluation with real users from the annual Amazon Alexa Challenge. However, their study does not report on abuse detection accuracy and thus it is hard to know whether users have indeed been abusive. Related research shows that standard methods such as blacklisting words and using off-the-shelf tools (trained on out-of-domain data) show poor results on this task [29, 30].

5 Summary

To conclude, our study is the first study to present evidence that the manipulation of a voice assistant’s gender was associated with changes in the perceived appropriateness of responses to sexual harassment. Further research is needed to understand why, specifically for female voice assistants, the perceived appropriateness of responses differed from our expectations.

6 Acknowledgements

This research received funding from the EPSRC project *‘Designing Conversational Assistants to Reduce Gender Bias’* (EP/T023767/1). During data collection, the first author was affiliated with the PFH Göttingen, Germany.

References

[1] M. West, R. Kraut, and H. Ei Chew, “I’d blush if i could: closing gender divides in digital skills through education,” 2019.

[2] J. P. Cabral, B. R. Cowan, K. Zibrek, and R. McDonnell, “The influence of synthetic voice on the evaluation of a virtual character.” in *INTERSPEECH*, no. 2, 2017, pp. 229–233.

[3] G. Abercrombie, A. C. Curry, M. Pandya, and V. Rieser, “Alexa, Google, Siri: What are your pronouns? Gender and anthropomorphism in the design and perception of conversational assistants.” in *ACL-IJCNLP 2021 3rd Workshop on Gender Bias in Natural Language Processing (GeBNLP 2021)*, 2021.

[4] Legislation.gov.uk. (2019) Equality act 2010. [Online]. Available: https://www.legislation.gov.uk/ukpga/2010/15/section/26

[5] A. C. Curry and V. Rieser, “# metoo alexa: How conversational systems respond to sexual harassment,” in *Proceedings of the second acl workshop on ethics in natural language processing*, no. 7, 2018, pp. 7–14.

[6] A. De Angeli, R. Carpenter *et al.*, “Stupid computer! abuse and social identities,” in *Proc. INTERACT 2005 workshop Abuse: The darker side of Human-Computer Interaction*, no. 4. Citeseer, 2005, pp. 19–25.

[7] B. Reeves and C. Nass, *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, United Kingdom, 1996, no. 5.

[8] L. Fessler. (2017) We tested bots like siri and alexa to see who would stand up to sexual harassment. [Online]. Available: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/

[9] A. C. Curry and V. Rieser, “A crowd-based evaluation of abuse response strategies in

conversational agents," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, no. 8, 2019, pp. 361–366.

[10] L. S. Aloia and D. H. Solomon, "Sex differences in the perceived appropriateness of receiving verbal aggression," *Communication Research Reports*, vol. 34, no. 1, pp. 1–10, 2017.

[11] B. A. Gutek, "Understanding sexual harassment at work," *Notre Dame JL Ethics & Pub. Pol'y*, vol. 6, no. 10, p. 335, 1992.

[12] J. R. Kelly and S. L. Hutson-Comeaux, "The appropriateness of emotional expression in women and men: The double-bind of emotion," *Journal of Social Behavior and Personality*, vol. 15, no. 4, p. 515, 2000.

[13] M. M. Linehan and R. F. Seifert, "Sex and contextual differences in the appropriateness of assertive behavior," *Psychology of Women Quarterly*, vol. 8, no. 1, pp. 79–88, 1983.

[14] C. G. Nelson, J. A. Halpert, and D. F. Cellar, "Organizational responses for preventing and stopping sexual harassment: effective deterrents or continued endurance?" *Sex Roles*, vol. 56, no. 11-12, pp. 811–822, 2007.

[15] P. N. Lewis and C. Gallois, "Disagreements, refusals, or negative feelings: Perception of negatively assertive messages from friends and strangers," *Behavior Therapy*, vol. 15, no. 4, pp. 353–368, 1984.

[16] F. Eyssel and F. Hegel, "(s) he's got the look: Gender stereotyping of robots 1," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[17] F. Eyssel, L. De Ruiter, D. Kuchenbrandt, S. Bobinger, and F. Hegel, "'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, no. 16. IEEE, 2012, pp. 125–126.

[18] C. Nass, Y. Moon, and N. Green, "Are machines gender neutral? gender-stereotypic responses to computers with voices," *Journal of applied social psychology*, vol. 27, no. 10, pp. 864–876, 1997.

[19] J. Payne, A. Szymkowiak, P. Robertson, and G. Johnson, "Gendering the machine: Preferred virtual assistant gender and realism in self-service," in *International Workshop on Intelligent Virtual Agents*, no. 18. Springer, 2013, pp. 106–115.

[20] Prolific. (2019). [Online]. Available: https://www.prolific.co/

[21] Microsoft. (2019). [Online]. Available: https://www.microsoft.com/

[22] J. Novikova, O. Dušek, and V. Rieser, "Rankme: Reliable human ratings for natural language generation," *arXiv preprint arXiv:1803.05928*, no. 21, 2018.

[23] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

[24] S. Sarkar, "Media and women image: A feminist discourse," *Journal of Media and Communication Studies*, vol. 6, no. 3, pp. 48–58, 2014.

[25] E. Camussi and C. Leccardi, "Stereotypes of working women: the power of expectations," *Social science information*, vol. 44, no. 1, pp. 113–140, 2005.

[26] S. J. Sutton, "Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, ser. CUI '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3405755.3406123

[27] H. Chin and M. Y. Yi, "Should an agent be ignoring it? a study of verbal abuse types and conversational agents' response styles," in *Extended Abstracts of the 2019 CHI*

Conference on Human Factors in Computing Systems, no. 23, 2019, pp. 1–6.

[28] H. Li, D. Soylu, and C. Manning, "Large-scale quantitative evaluation of dialogue agents' response strategies against offensive users," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, no. 23. Singapore and Online: Association for Computational Linguistics, July 2021, pp. 556–561. [Online]. Available: https://aclanthology.org/2021.sigdial-1.58

[29] A. Cercas Curry, G. Abercrombie, and V. Rieser, "ConvAbuse: Data, analysis, and benchmarks for nuanced detection in conversational AI," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7388–7403. [Online]. Available: https://aclanthology.org/2021.emnlp-main.587

[30] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser, "Anticipating safety issues in e2e conversational ai: Framework and tooling," 2021.

FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

FRIAS Junior Researcher Conference
Human Perspectives on Spoken Human-Machine Interaction
November 15–17, 2021

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

Aryana Collins Jackson[1], *Elisabetta Bevacqua*[1], *Pierre De Loor*[1], *Ronan Querrec*[1]

[1]ENIB, Lab-STICC UMR 6285 CNRS, 29200, Brest, France

jackson@enib.fr

Abstract

In emergency medical procedures, positive and trusting interaction between followers and leaders are imperative. That relationship 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 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 analysing the agreement among annotators and performing *k*-means clustering and latent Dirichlet allocation, we compile easily-reproducible rules that dictate how speech should appear in each leadership style for use in a virtual agent system.

1 Introduction

During an unexpected medical emergency on a remote site without medical experts nearby, the individuals present must assume the roles of caregivers. Regardless of whether these amateur caregivers have medical experience, a leader is necessary to ensure the procedure is adhered to [1]. We propose a virtual medical assistant agent to guide the caregivers during an emergency in an isolated, remote site. This virtual agent will be fully equipped with knowledge of the humans' capabilities and the medical procedure's tasks and resources.

While the medical procedure is the priority, also of great importance is how the agent interacts with the caregivers in order to create a positive working relationship [1]. To accomplish this goal, we employ situational leadership, enabling the agent to communicate with and guide the caregivers by matching them with an appropriate leadership style [2].

Situational leadership describes four leadership styles composed of high or low levels of task (direction regarding the task) and relationship (socioemotional support) behavior [2]:

1. Directing: high task and low relationship;
2. Coaching: high task and high relationship;
3. Supporting: low task and high relationship;
4. Delegating: low task low relationship.

Despite various studies on the performance of situational leadership [3], no prior work has been completed to discover how leadership style might change vocabulary and syntax, which is what we explore in this paper. Therefore our work provides novel contributions to the fields of human behavior, healthcare, and intelligent virtual agents.

Our SAIBA-compliant agent framework involves text-to-speech, without an emphasis on intonation [4], so leadership style must be determined from text only. We compiled medical leader (coordinator or surgeon) speech into a corpus which were then annotated with leadership style by four people. This annotated corpus was then analysed in order to generate rules regarding agent speech in each of the four leadership styles.

In this paper, we briefly discuss the state of the art, we explain how we built our corpus, and we explain our methods of analysis and results.

2 State of the Art

This work encompasses three main domains: virtual healthcare agents, leadership, and linguistics. Healthcare agents have been used previously for training and coaching [5], questionnaires and diagnostics [6], and patient monitoring [7]. In these

DOI: 10.6094/UNIFR/223815

systems, agents accept spoken input from patients, rather than caregivers, and there are clear and fixed steps in a system in which two-way conversation is encouraged between the agent and the patient. A comprehensive review of ECAs in healthcare is also available from 2018 [8].

A huge amount of research involving ECAs as leaders investigates agents as tutors or teachers, where an agent assumes a role of authority and aims to lead a human through a series of steps [9, 10]. Sometimes, embodied tutors take into account the prior knowledge of the user as well as the actions taken by the user throughout the learning experience [9]. An agent's personalized content and conversations have been found to improve user engagement, improve the quality of speech, provide timely feedback during the interaction, provide adaptive training, and allow for self-reflection [10].

The final component of this research involves linguistics founded in Speech Act Theory (SAT). SAT is a theory of linguistics that explores how words work together to form utterances that perform actions and is based on communicative or speaker's intention and form [11]. While intention and form are not directly correlated, they are related [12]. For example, certain moods (e.g., imperatives, interrogatives, and indicatives) which can explain a speaker's attitude, go hand-in-hand with certain sentence structures. Other work has explored how attitudes manifest in written communication [13].

3 Compiling the Corpus

The corpus contains coordinating nurse or doctor speech from various emergency room simulation videos and some previous literature. The speech was split by complete utterance (294 total), separated by change of speaker and change of situation state (e.g., before a patient receives an IV and after). These utterances were then separated by segment (375 total) designated by a single subject-verb pair [14]. Each utterance, sentence, and segment (referred to as strings from now on) was labeled with its grammatical mood (situational syntactic expression; our corpus contains the imperative, interrogative, and indicative moods), whether the string was direct or indirect (whether its literal meaning differed from its implied meaning) [11], and its speech acts [15].

Four annotators were chosen: one woman and three men, ages 21-29, all native English speakers from the US and Ireland, and each with a minimum education level of some college. None had medical experience, ensuring that the results of our analysis are applicable to novice caregivers.

The annotators were given the following information: (i) the definitions of task and relationship behavior, (ii) the definitions of each leadership style as explained in the introduction, and (iii) a list of the original descriptors for each leadership style [2]. Annotators were asked to assign a leadership style to each string in the corpus. The order of strings was randomized for each annotator to ensure that it did not have any effect.

4 Pattern Analysis

In order to find the linguistic rules that separate each leadership style from the others, we search for patterns among the annotations. Because this work is not semantic in nature, we do not apply methods such as word embeddings or bag-of-words models [16, 17]. In this section, we discuss the analysis methods we used and the results.

4.1 Agreement Analysis

The Fleiss kappa statistic representing the agreement among annotators on the entire corpus was 0.415 (p-value<0.05), indicating moderate agreement (127 strings total were agreed-upon) [18]. Before understanding what string elements led to agreement, we grouped the annotations by low and high task behavior (directing and coaching together and supporting and delegating together). The Fleiss kappa value then jumped to 0.570 (p-value < 0.05). When grouped by low and high relationship behavior (directing and delegating together and coaching and supporting together), the kappa dropped to 0.362 (p-value < 0.05). These results indicate that annotators agree more on indicators of task behavior than those of relationship behavior and imply that indicators of relationship behavior may be more unique to individual followers.

We analyzed several speech characteristics to understand what elements lead to a consensus of leadership style. Using context from the situations in which speech occurs, we determined whether the string was direct or indirect; an indirect string may have literal and implied meanings that differ while

Table 1: The Fleiss kappa values of strings that only included one mood each, for a total of 328 strings (73 imperatives, 44 without let; 76 interrogatives; 178 indicatives). The overall kappa value is 0.404.

	Imperatives			Interrogatives	Indicatives
	all	*with "let"*	*without "let"*		
Directing	0.187*	0.000	0.274*	-0.008	0.264*
Coaching	0.217*	-0.008	0.326*	0.126*	0.374*
Supporting	-0.043	-0.036	0.256*	0.075	0.310*
Delegating	0.111	0.094	0.010	0.097	0.547*

*p-value < 0.05

direct strings' literal and implied meanings are the same. [11]. The Fleiss kappa for direct strings was 0.377, and the kappa for indirect strings was 0.193. When separated by assigned leadership style, the annotators had far more agreement when it came to direct strings except for when coaching leadership was assigned (kappa = 0.265, p-value < 0.05). This is likely due to the strings that use the interrogative mood yet aim to direct the follower to do something. In cases such as these, the form does not match the intention, and so they are indirect.

We also analysed the agreement in terms of mood (see the Fleiss statistics in Table 1). The annotation results indicate that an imperative containing "let" is often interpreted differently in English than imperatives with other verbs (e.g., "Let's go home" vs "Go home"; the first implies that the speaker is involved whereas the second does not imply involvement by the speaker [11]). Imperatives with "let" are more ambiguous than those without, as shown by the kappa value, implying that leadership speech should generally avoid imperatives using "let".

Generally, interrogative strings were not agreed upon. The strings that annotators most agreed upon were indicatives that were ultimately labeled as containing delegating leadership.

As shown in Table 2, not all kappa values are significant, and some are likely low because the speech acts are not distributed evenly throughout the corpus. Speech acts *offer*, *support*, *request information*, and *respond* do not show up often within the corpus, which indicates that they would not often present themselves during a medical procedure, although there is a possibility that this is due to the size of the corpus. Regardless, it is clear that certain speech acts belong in certain leadership styles by examining the agreement statistics.

4.2 Agreement Between Individuals

We then explored whether there were any patterns in how annotators rated leadership style in terms of age/work experience and gender. The Fleiss kappa statistic for just the male annotators was 0.433 (p-val < 0.001), which is not much higher than the overall kappa statistic of 0.415. The kappa for the three annotators aged 27-29 with significant work experience was 0.387 (p-val < 0.001), indicating that gender and age/work experience had no effect on perceptions of leadership style.

When the annotators' ratings were grouped by task behavior, the agreement among men was 0.536 (p-value < 0.001), and when grouped by relationship behavior, the kappa was 0.397 (p-value < 0.001) - higher than the overall kappa when ratings were grouped by relationship behavior. This might suggest that indicators of relationship behavior could change depending on gender. However, the agreement is still rather low, which again points to relationship behavior being very individual.

When the responses from the older annotators with more work experience were grouped by task behavior, the kappa is 0.56 (p-value < 0.001). When grouped by relationship behavior, the kappa is 0.312 (p-value < 0.001).

More research is needed to understand how individuals perceive relationship behavior and how varying levels of task and relationship behavior influence a follower's performance during a task.

While we gathered some valuable insights from examining the annotated corpus statistically, we performed clustering to discover further patterns between each leadership style.

DOI: 10.6094/UNIFR/223815

Table 2: The Fleiss kappa values of strings containing each speech act. *Totals* refers to the number of strings labeled with that speech act.

	Instruct	**Inform**	**Offer**	**Request information**	**Respond**	**Support**
Directing	0.427*	0.294*		-0.081		-0.031
Coaching	0.531*	0.396*	-0.500	-0.207*		0.593*
Supporting	-0.016	0.099*	-0.500	-0.088	-0.204	0.455*
Delegating	0.180*	0.555*		-0.029	-2.04	-0.138
Totals	*168*	*138*	*1*	*42*	*15*	*11*

*p-value < 0.05

4.3 Clustering

The corpus is first limited to only the strings that were agreed upon by all four annotators in terms of leadership style, leaving 127 strings. Each string was part-of-speech (POS) tagged with Stanford CoreNLP. The POS-tagged strings with the words removed as well as the strings without POS-tags are clustered separately using k-means [17]. The similarity measure used here is cosine similarity which determines the cosine between two vectors. The process involves (i) identifying common sequences of words within a group and (ii) representing each string by a numeric vector composed of 0s and 1s based on the presence of each of those common words or phrases in that particular string [16]. This method is similar to a bag-of-words model in that word order does not matter.

The goal is to identify patterns among the agreed-upon strings and then check whether those patterns are indicative of one leadership style. The sum of squared differences (SSD) is used to determine the number of optimal clusters. The strings are then clustered with k-means into the optimal number of clusters based on the presence of common sequences within each string as explained above. The leadership style present in each cluster and the common sequence(s) that define each cluster then define the linguistic rules for each leadership style.

Common sequences were found by defining the length of the sequence and the number of times that sequence needed to exist among the agreed-upon strings. Clustering with k-means was performed (see Figure 1), and the resulting clusters that contained a single (or nearly a single) leadership style were examined. The common sequences that formed the clusters and were found to be present in only one leadership style are listed in Table 3).

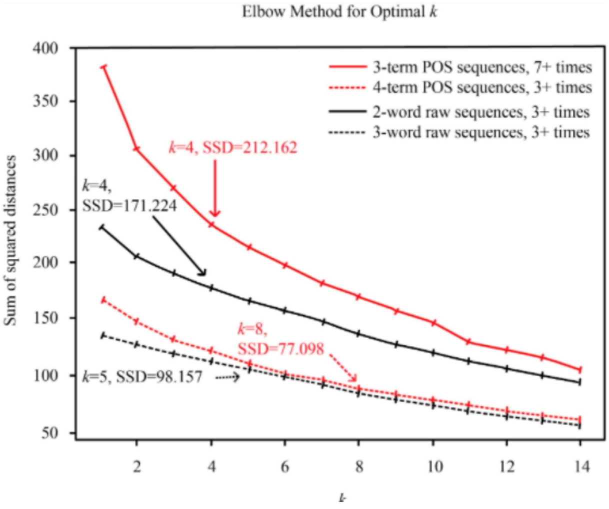


Figure 1: The SSD at optimal k when the raw strings and POS tags only from agreed-upon strings are clustered. The legend gives the number of words or POS terms that form the sequence and the number of times that sequence had to be in the 127 agreed-upon strings for it to be considered a common sequence.

Sometimes, a POS sequence corresponded to a single sequence of raw words; in these cases, the words themselves are in the table instead of the POS tags.

4.4 Analysis of Individual Annotations

Analyzing the agreed-upon strings is useful for finding characteristics of speech that might be universally recognized, but we also must account for differences between the annotators. Using latent Dirichlet allocation (LDA), we explore each annotator's assignment of leadership style [19]. Sequences of raw words did not yield meaningful results, so sequences of three POS tags were used to find important and distinct groups. An initial assessment using LDA on the agreed-upon strings resulted in many of

Table 3: A list of rules generated by clustering on the agreed-upon strings' POS tags. When a sequence of POS tags tended to be a set of specific words, only the specific words were included.

	Directing	**Coaching**	**Supporting**	**Delegating**
Directness	Direct	Direct, Indirect	Direct	Direct
Mood	Imperatives without "let", Indicatives	Interrogatives, Indicatives	Indicatives	Indicatives
Speech acts	instruct	instruct, inform, support	support	inform
Keywords	"We need to, "I want you to", "Carry on with"	"please", "Okay, can someone", "for me, please", "as well, please", "Please, can we", "Can you please", "You can"	"Okay, thank you"	"I see that", "It looks like"
POS tags		MD PRP VB, PRP MD VB		VBZ IN PRP$

the same sequences that were produced by clustering. Only some of our results are discussed here.

The first annotator that we examine is female, age 26, with significant work experience. The most represented POS sequence for strings labeled with directing and coaching leadership was VB DT NN (e.g., "check the pulse"). Strings containing the former were labeled with high-task behavior (directing or coaching) by all annotators, indicating agreement on task behavior when that sequence is used.

Annotator 1 assigned directing leadership to sequence VB JJ PRP (e.g., "make sure you"). Strings containing the former were also labeled with high-task leadership (directing or coaching) by all annotators except for the male annotator aged 21 with less work experience, who labeled them as having delegating leadership.

She assigned coaching to strings with the sequence VB PRP VB, which entirely corresponded to "let's" + verb. Other annotators assigned these strings styles 1-3, which confirms the lack of agreement when "let" is used. If we were tailoring our virtual agent's speech to this annotator in particular, we would use the word "let" to begin utterances with high task and high relationship behavior.

The male annotator aged 21 with limited work experience seemed to assign leadership style that did not match the assignments by the other annotators the most. The most representative sequence of strings he assigned with supporting leadership was PRP VBP DT (e.g., "we have a", "I am a"). The other annotators assigned these strings leadership styles 1-4. This annotator clearly identifies an introductory statement as well as the use of "we" as being an indicator of high relationship behavior, which is not true for the other annotators.

Findings such as these demonstrate how even further personalization of the agent's communication might be necessary to correspond to an individual's definition of task and relationship behavior.

5 Conclusions

Using our annotated corpus of medical leader speech, we have identified linguistic rules for each leadership style. These rules determine what kinds of utterances a leader should make depending on the appropriate leadership style. This work is intended to be used in a dialogue manager for a virtual medical assistant who guides human caregivers during a medical procedure. The agent must communicate in a manner appropriate to the caregiver. By designing the agent's speech according situational leadership rules, we believe that the agent is able to establish a positive working interaction with the caregivers.

6 Acknowledgements

This work has been carried out within the French project VR-MARS which is funded by the National Agency for Research (ANR).

References

[1] T. Manser, "Teamwork and patient safety in dynamic domains of healthcare: a review of the literature," *Acta Anaesthesiol Scand*, vol. 53, no. 2, pp. 143–51, February 2009.

[2] P. Hersey, K. H. Blanchard, and D. E. Johnson, *Management of Organizational Behavior:*

Leading Human Resources, 5th ed. Prentice-Hall, 1988, ch. Situational Leadership, pp. 169–201.

[3] C. Bedford and K. M. Gehlert, "Situational supervision: Applying situational leadership to clinical supervision," *The Clinical Supervisor*, vol. 32, no. 1, pp. 56–69, 2013.

[4] A. Collins Jackson, E. Bevacqua, P. De Loor, and R. Querrec, "Modelling an embodied conversational agent for remote and isolated caregivers on leadership styles," in *Proceedings of the 19th Internation Conference, Intelligent Virtual Agents*. Paris, France: ACM, 2019, pp. 256–259.

[5] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders," *PLoS ONE*, vol. 8, no. 12, 08 2017.

[6] P. Philip, J.-A. M. Franchi, P. Sagaspe, E. de Sevin, J. Olive, S. Bioulac, and A. Sauteraud, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," *Scientific Reports*, no. 1, 02 2017.

[7] L. Black, M. F. Mctear, N. D. Black, R. Harper, and M. Lemon, "Appraisal of a conversational artefact and its utility in remote patient monitoring." 18th IEEE Symposium on Computer-Based Medical Systems, 07 2005, p. 506–8.

[8] L. Laranjo, A. G. Dunn, H. L. Tong, and A. B. Kocaballi, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, p. 1248–1258, 07 2018.

[9] J. Taoum, A. Raison, E. Bevacqua, and R. Querrec, "An adaptive tutor to promote learners' skills acquisition during procedural learning." ITS Workshops, 2018.

[10] A. B. Kocaballi, S. Berkovsky, J. C. Quiroz, and L. Laranjo, "The personalization of conversational agents in health care: Systematic review," *Journal of Medical Internet Research*, vol. 11, no. 21, 11 2019.

[11] J. R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, 1979.

[12] R. Ferreira, R. Lins, S. Simske, F. Freitas, and M. Riss, "Assessing sentence similarity through lexical, syntactic and semantic analysis," *Computer Speech and Language*, vol. 39, 02 2016.

[13] M. Hansen, S. Fabriz, and S. Stehle, "Cultural Cues in Students' Computer-Mediated Communication: Influences on E-mail Style, Perception of the Sender, and Willingness to Help," *Journal of Computer-Mediated Communication*, vol. 20, no. 3, pp. 278–294, 01 2015.

[14] M. Weisser, *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. John Benjamins Publishing Company, 2018.

[15] H. Bunt, "The DIT++ taxanomy for functional dialogue markup," in *Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS. Budapest, Hungary: ACM, January 2009.

[16] J. Oliva, J. I. Serrano, M. del Castillo, and A. Iglesias, "Symss: A syntax-based measure for short-text semantic similarity," *Data Knowledge Engineering*, vol. 70, pp. 390–405, 04 2011.

[17] R. Khoury, "Sentence clustering using parts-of-speech," *International Journal of Information Engineering and Electronic Business*, vol. 4, 02 2012.

[18] J. Landis and G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33 1, pp. 159–74, 1977.

[19] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey," vol. 78, no. 11, 2019.

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

{dattas1|heidi.christensen}@sheffield.ac.uk

Abstract

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 patient's engagement and communicative behaviours especially concerning their progress. These behaviours are closely related to the patient's 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 PWP's and patient's 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).

1 Introduction

Low intensity psychological interventions, such as Cognitive Behavioural Guided self-help (CBT-GSH) involve focusing on healthy and unhealthy thought patterns. Other interventions such as Cognitive Analytic Guided Self-help (CAT-GSH) involve thinking about reciprocal roles and their impact on our relationships. In both cases, the role of emotions is important: the patient's emotions are considered a dynamic factor that could establish a positive therapeutic alliance in the sessions. This involves agreeing on tasks, treatment goals, as well as establishing a genuine human relationship. The PWP can acknowledge and motivate the patient's positive feelings through talking therapy. Furthermore, the PWP should be empathetic with the patient's emotions and evaluate any flawed thinking that has impacted the patient's mood [1]. Assessing the competence includes evaluating the PWP's use, knowledge, and implementation of the treatment [2]. Predicting patient emotion automatically using specific dimensions could help capture the small variations in emotions which could help the PWP determine the appropriate treatment plan.

The common techniques used in Speech Emotion Recognition (SER) are the discrete and dimensional models. Discrete emotion theory defines six categories of basic emotions: sadness, happiness, fear, disgust, anger, and surprise [3]. The dimensional emotional model uses several controlled dimensions to represent emotions in a continuous manner such as arousal, valence, control, and power [4]. These dimensions are categorical and universal aspects of emotion. One of the most adopted dimensional models in SER is a two-dimensional model consisting of *arousal* versus *valence*. The *arousal* dimension describes the strength of the felt emotion. It may range from excited to apathetic. The *valence* dimension illustrates whether an emotion is positive such as happy and calm or negative such as anger and depressed [5].

In this study we have explored automatic methods of continuously predicting labels of emotions for arousal and valence in guided self-help sessions. The guided self-help sessions dataset used is not labelled with emotional labels and hence the system is trained on a benchmark database and then this model is used to predict emotions in the guided self-help sessions dataset. Considering that the emotions in those session recordings are natural, it is essential to preserve that in the data to match the study scope. For that reason, the chosen benchmark database for training is a natural speech emotion database called Remote Collaborative and Affective Interactions (RECOLA) [6].

2 Guided Self-help Sessions Dataset

The recordings of the sessions used in this study were collected for a study aimed at comparing efficiency and clinical durability of anxiety disorders manualised treatments, the Cognitive Analytic Guided Self-help (CAT-GSH) and Cognitive Behavioural Guided Self-help (CBT-GSH) [7]. In total, 54 session recordings has been included in this study, where each includes a conversation between a PWP and a patient as part of their therapy treatment. The PWPs deliver low-intensity interventions for mild to moderate anxiety. They *guide* the patients through treatment in contrast to traditional therapists [8]. The sessions vary in length between 30 and 40 minutes totalling 27 hours and 14 minutes. Some sessions were conducted over the mobile phone due to the Covid-19 pandemic. The 54 sessions are split into 20 mobile phone sessions and 34 face-to-face sessions. Table 1 describes the full dataset. Each session is labelled with patient depression and anxiety outcome scores. Those scores could assist in determining the patient emotional state as patients with a high level of anxiety or depression are expected to reveal more negative emotions [9]. Also, each session has a PWP's competence rating score. The PWP's ability to understand, manage, and handle the patient's emotions is closely related to the PWP competence ratings [10]. Furthermore, each session had time-stamped speaker turn labels allowing for speaker-specific analysis of the acoustic features.

Table 1: Patient demographics and dataset info

Patient demographics	Total	Average	Min	Max
Number of patients	54	-	-	-
Female	39 %	-	-	-
Age	-	39	16	74
In-person sessions	34 %	-	-	-

The benchmark database used in the study was designed in a collaboration by a team in informatics and psychology at the Université de Fribourg, Switzerland [6]. The RECOLA database was recorded to study the socio-affective behaviours from multimodel data. The spontaneous and naturalistic interactions were recorded during the resolution of a collaborative task performed in dyads and remotely through video conference. The recordings have emotionally annotated time continuously for the dimensions (arousal and valence) using six annotators for every 0.4 seconds.

3 Dimensional Emotion Recognition Baseline

The Audio-Visual Emotion recognition Challenge (AVEC) is one of the well-known in the field of continuous emotion recognition. One of the AVEC 2018 sub-challenges related to emotion recognition is the Gold-standard Emotion Sub-challenge (GES) that focus on generating dimensional emotion labels by fusing continuous annotations of dimensional emotions rated by several annotators [11]. Then, the fused annotation is used in the challenge to train and evaluate a baseline emotion recognition system using the RECOLA dataset. The emotion labels are by nature highly variable and subjective [12] and the GES challenge followed the dominant approach in the literature referred to as *gold standard* that is to combine the annotations for each recording across time using the Evaluator Weighted Estimator (EWE) based approach [13].

The GES uses different supervision levels in feature extraction: at the supervised level, features depend directly on the expert's knowledge-based representations. The emotions are continuously recognised and then summarised using low-level descriptor (LLDs) features with a set of functionals

DOI: 10.6094/UNIFR/223820

computed over a fixed duration sliding window. The LLDs usually contains spectral, cepstral, prosodic, and voice quality features. The Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was used as the baseline feature set. It comprises 88 features covering the formerly mentioned acoustic features. In addition, 13 Mel-frequency cepstral coefficients (MFCCs), and their deltas and delta deltas were computed using a set of acoustic LLDs. The features were extracted using the openSMILE toolkit [14]. The semi-supervised level used the technique of Bag of Audio Words (BoAW) which represents the distribution of LLDs based on a dictionary learned from them [15]. The MFCCs were used in BoAW as a front end to compute the acoustic features. The BoAW was extracted using the open-source toolkit openXBOW [16].

Several dimensional regressors were used in the challenge such as Support Vector Machines (SVM) from the liblinear toolkit [17], and Generalised Linear Models (GLMs) such as Ridge regression, Elastic Net, Lasso from the scikit-learn toolbox [18]. Furthermore, multi-task formulation of the Elastic Net and Lasso algorithms have been implemented to make use of the correlations between the dimensions. The challenge achieves the best results using the BoAW in recognising the dimensional emotions under the audio modality. The valence dimension results are quite challenging comparable with the arousal which is commonly agreed in the literature [11].

4 Experimental setup

The AVEC 2018 challenge was used as a baseline for the study for the dimensional emotion recognition system. The GES Sub-challenge was selected due to their compatibility with the experiment requirements and the availability of the gold standard as labels for the emotions assuring the efficiency of the emotional labels. As an initial phase, the baseline was implemented using the same baseline training and development sets to train the system for further experiments. The features selected to be used for the guided self-help sessions dataset are eGeMAPS and BoAW because BoAW features gained the best correlation coefficients in the baseline for the audio modality [11]. The eGeMAPS used at the beginning to build and train the SER system using the guided self-help sessions dataset. The eGeMAPS functionals computed are the arithmetic mean and the coefficient of variation on all 42 LLDs. Furthermore, the functionals applied to the pitch and loudness features were: percentiles 20, 50 and 80, the range of percentiles 20 to 80, and the mean and standard deviation of the slope of rising/falling signal parts. The functionals related to pitch, jitter, shimmer and all formant related LLDs were only calculated for the voiced regions. In addition, some temporal features were calculated, such as the rate of loudness peaks per second, average length and standard deviation of continuous voiced and unvoiced segments, and the rate of voiced segments per second.

The BoAW are audio representations formed by bagging acoustic LLDs such that each frame-level LLD vector is allocated to an audio word from a codebook retained from the training data [19]. A fixed-length histogram representation of an audio recording is generated by counting the number of assignments for each audio word. The extracted features from the training set are then concatenated and normalised for the classification phase. The classifiers used in the recognition system are SVM and Generalised Linear Models (ridge regression, lasso, multi-task lasso, elastic net, and multi-task elastic net).

5 Results

The study explored two main objectives: the effect of using several feature sets in the SER system on the prediction of the arousal and valence dimensions from a quantitative perspective, including changes in the patient only segments or the whole sessions (PWP and patient). The other objective is to conduct a qualitative study to understand the subtle changes in the patient's emotions and emotions within the interaction between the patient and the PWP during several periods in the therapy treatment.

To explore the differences between the results gained using the eGeMAPS and the BoAW features, the averaged predictions of the dimensions arousal and valence were plotted for the patient only speaking turns as shown in Figure 1 and 2. Figure 3 and 4 presents the same latter measures for the full sessions. The Figures show that predictions of both features in the arousal dimension are close to each other, while the BoAW gained more reasonable results than the eGeMAPS in the valence dimension. The results gained confirms the baseline results such that the valence is a bit challenging to predict, comparable with the arousal. Furthermore, the BoAW enhances the valence predictions, while both features predictions are approximately analogous to each other in the arousal.

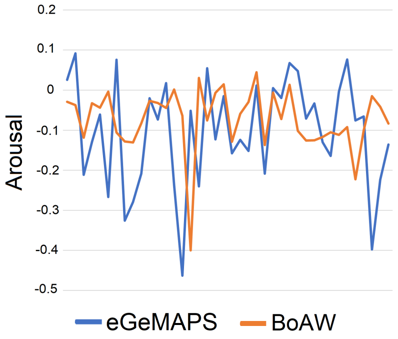


Figure 1: Arousal prediction for patient turns

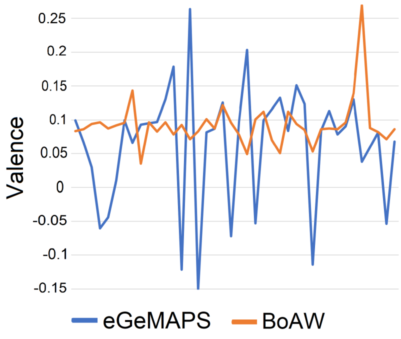


Figure 2: Valence predictions for patient turns

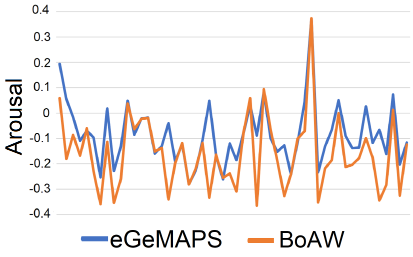


Figure 3: Arousal predictions across a full session

The predicted arousal and valence values as well as the interactional patters were qualitatively analysed across the sessions and several observations made. There were more patient speaking turns during the sessions recorded at the start of their guided self-help plan comparable to the later sessions in therapy. This was especially true in sessions with high PWP competence ratings. This is likely because the patient in those earlier sessions are invited to spend time describing their issues and how these affects their normal life. In the later sessions, there are more PWP speaking turns, which could relate to the PWP trying to resolve the patient's problems and assist them in dealing with those mental difficulties. In the high competence rating sessions, the PWP tries to sync the patient's emotions and shows various emotions as a sign of approval and empathy toward the patient. The low competence rating sessions show that the PWP does not reveal a variation in emotions or show natural emotions along with the session.

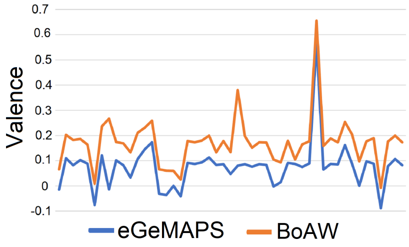


Figure 4: Valence predictions across a full session

6 Conclusions

In this study, we investigated the efficacy of predicting continuous emotional labels on guided self-help sessions. The AVEC 2018 were used as a baseline for building and training the SER system. Due to the unavailability of the emotions labels in the guided self-help dataset, the RECOLA database was used as a benchmark database for training the system. The trained system was tested using the guided self-help sessions dataset. The features set used for training the system are eGeMAPS and BoAW. Several classifiers have been used for predicting the continuous emotional labels, such as SVM, Ridge regression, Elastic Net, and Lasso. The BoAW feature results approved the baseline results by improving the outcomes of the valence dimension. Furthermore, several remarks were reported in the qualitative study relating to the numbers of the speakers speaking turns and PWP competence in several stages in the therapy treatment.

References

[1] J. S. Beck, *Cognitive behavior therapy [electronic resource] : basics and beyond*, 2nd ed. New York ; London: Guilford, 2011.

[2] J. P. Barber, B. A. Sharpless, S. Klostermann, and K. S. McCarthy, "Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature." *Professional Psychology: Research and Practice*, vol. 38, no. 5, p. 493, 2007.

[3] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013, vol. 11.

[4] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[5] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[6] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[7] S. Kellett, C. Bee, V. Aadahl, E. Headley, and J. Delgadillo, "A pragmatic patient preference trial of cognitive behavioural versus cognitive analytic guided self-help for anxiety disorders," *Behavioural and Cognitive Psychotherapy*, p. 1–8, 2020.

[8] N. Firth, M. Barkham, S. Kellett, and D. Saxon, "Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis," *Behaviour Research and Therapy*, vol. 69, pp. 54–62, 2015.

[9] W. Riley, F. Treiber, and M. Woods, "Anger and hostility in depression," *The Journal of nervous and mental disease*, vol. 177, pp. 668–74, 12 1989.

[10] J. McKenna and J.-a. Mellson, "Emotional intelligence and the occupational therapist," *British Journal of Occupational Therapy*, vol. 76, no. 9, pp. 427–430, 2013.

[11] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[12] A. Tversky, "Intransitivity of preferences." *Psychological review*, vol. 76, no. 1, p. 31, 1969.

[13] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 381–385.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010.

[15] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt, and B. Schuller, "Multimodal bag-of-words for cross domains sentiment analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4954–4958.

[16] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," 2017.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, 2011.

[19] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484.

FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

FRIAS Junior Researcher Conference
Human Perspectives on Spoken Human-Machine Interaction
November 15–17, 2021

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

ndpevy1@sheffield.ac.uk; m.reuber@sheffield.ac.uk

Abstract

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 an experience of transient loss of consciousness 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.

1 Introduction

Advances in speech processing technology have allowed researchers to explore the relationship between properties of speech and other group characteristics. One particularly prominent area of research involves the identification of individuals with a health condition based upon these speech properties, for example psychiatric disorders [1] and Alzheimer's disease [2]. Although previous research aiming to identify a health condition using speech have demonstrated promising results, these studies are typically hindered by the scarcity of medical speech data that is available to train models [3]. Therefore, researchers may be limited by the data that is available, which could lead to variability between the context of each dataset, for example whether the participant is communicating using a telephone [4, 5], in person [6], or whether they are communicating directly to a computer [7]. One way to collect data reliably is to use speech data from human-computer interactions, but this method requires consideration regarding whether features that were effective predictors of the diagnosis in human-human interactions can maintain a high level of performance for recordings of human-computer interactions.

Research into human-computer interactions has explored differences in how people speak while interacting with a computer. It has been shown that people speak more concisely but take considerably longer because of long pauses associated with turn transitions while speaking with a computer compared to a person [8]. Research has also found that people speak louder and [9] and have a slower rate of speaking [10] when speaking to a computer compared to a person. The differences in how people speak with a computer can be influenced by the speaking style of the computer, such as changes in loudness and speaking rate in response to the same changes in the computer generated speech [11, 12]. These findings demonstrate that people speak differently when interacting with a computer.

However, many of these studies have explored speech properties in human-computer interactions where the computer is an active responder during the interaction, but this may not be applicable for the applications that are used in medical interviews because the focus is on the patient to provide information rather than the computer to respond to the patient. Therefore, these applications do not need to be and are not always responsive [7]. Research exploring speech differences in patient narratives without the presence of a responsive counterpart would increase our knowledge of how a person's speech may vary between human-human and human-computer interactions without the interference of communication problems or differences that are caused by the computer processing the speech input and responding. Furthermore, this would increase our knowledge of human-computer interaction differences that are present in narratives, which are a prominent part of medical encounters.

The objective of this study was to investigate whether there are differences in a select number of speech properties that are frequently used in speech processing applications in the medical field [1, 2] between narratives that are delivered to a human co-participant or an unresponsive virtual avatar. The features included speech rate, word count, pause-to-word ratio, and the total amount of time that people paused.

2 Method

2.1 Data

The data used for this analysis was taken from two independent studies exploring differences in how people who have experienced transient loss of consciousness (TLOC) describe what happened. TLOC is a term used to describe a short period of unconsciousness that often involves amnesia for the unconscious period, abnormal motor control, and a loss of responsiveness [13]. The three most common causes of TLOC are epilepsy, nonepileptic seizures, or syncope (fainting) [14].

The first dataset consisted of people talking about their experience of TLOC to a neurologist at the Royal Hallamshire Hospital in Sheffield, UK [15, 16]. There were 19 recordings from people who had experienced epileptic (n=6) or nonepileptic seizures (n=13). The neurologist was given strict instructions to encourage the patient to talk about their seizures from their own perspective; therefore, they did not interrupt the patient and waited for the patient to finish talking before asking a follow-up question. Patients were asked to describe their first, worst, and last seizure in turn. Only descriptions of the last seizure were used in this analysis, which were prompted by an utterance following the format ”tell me about your last attack”. There were 13 women and 6 men with a median age of 33. The average length of the audio recordings was 79.1 seconds (SD = 48.4).

The second dataset comes from an on-going study exploring the feasibility of predicting the cause of TLOC using interactions with a virtual avatar. The virtual avatar consisted of a series of videos where an animated head asked questions that were pre-recorded by a human. Participants are asked to provide information about their experience of TLOC by completing a closed attack history questionnaire and verbally describing what happened during their most recent attack to the virtual avatar. Participant's responses were recorded automatically once each video stopped playing. This analysis focuses on the first utterance that was “please tell me in as much detail as possible what happened during your most recent attack that caused you to lose consciousness”. Only responses where participants responded with a narrative were included in the analysis because there were some instances where participants resisted describing what happened in favour of making a complete negation (e.g. “I can't recall anything”), which is not uncommon in seizure consultations [15, 16]. There were 23 recordings from people with a diagnosis of epilepsy (14), nonepileptic seizures (3), or syncope (6). 14 recordings were from women and 9 were from men with a median age of 36. The average length of the audio recordings was 69.4 seconds (SD = 82.5).

2.2 Analysis

All of the recordings were manually transcribed. For the human-human interactions, a subsection was extracted from the whole recording that started immediately after the neurologist requested a description of the last seizure and ended when the neurologist started speaking again. A pause was defined as a silent period for more than 30 milliseconds, and

DOI: 10.6094/UNIFR/223821

the duration of each pause was manually calculated using Praat (version 6.1.34, 1992-2020, produced by Paul Boersma and David Weenink). The closure pauses of plosives were not recorded as pauses. Pauses immediately after the neurologist's turn or at the end of the narrative could be caused by different reasons depending on the group, for example unfamiliarity with the application or the reluctance of the neurologist to speak too soon to ensure the patient had finished their description ; therefore, these pauses were removed from the analysis. The remaining audio file, transcript, and recorded pauses were used to calculate the following features:

- **Speech rate** - the number of words spoken per minute
- **Word count** - the total number of words spoken during the narrative
- **Percentage of time spent pausing** - The total pause time divided by the overall audio length and multiplied by 100
- **Pause-to-word ratio** - The absolute number of pauses divided by the total number of words and multiplied by 100. Pause-to-word ratio was used instead of pause frequency to account for variations in the size of the narrative.

The normality of each measurement was tested using the Shapiro Wilk test, the homogeneity of variance was tested using Barlett's test, and group differences were tested using either an independent T-Test or Mann Whitney U test depending on the outcome of the first two tests. The alpha value was set at 0.05.

3 Results

A group comparison of the univariate data for each feature is displayed in Figure 1.

3.1 Speech rate

An Independent T-Test found a significant difference in the speech rate between people speaking to another person (mean=154.25. SD=40.5) compared to people speaking to the virtual avatar (mean=125.62, SD=38.5), t(41)=2.285, p<0.05. People speaking with a human spoke faster than those speaking with the virtual avatar.

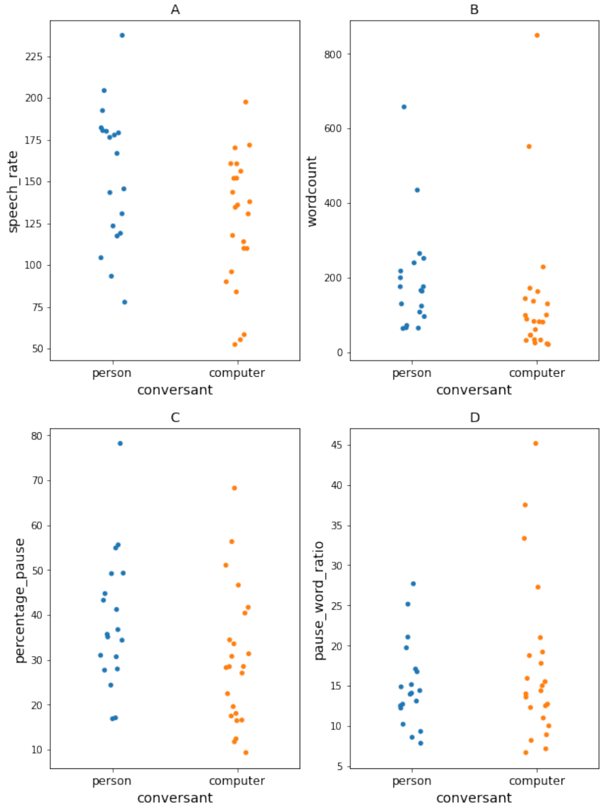


Figure 1: a strip plot comparing the scores for each measurement between people conversing with a person and a computer. A) speech rate. B) total number of words. C) percentage of time spent pausing. D) pause-to-word ratio.

3.2 Word count

A Mann Whitney U test demonstrated a significant difference (U=118.5, p<0.01) between the number of words said by people speaking with a human (median=165) compared to people speaking with the virtual avatar (median=82). People speaking with another human said more during their narrative than those speaking with the virtual avatar.

3.3 Percentage of time spent pausing

An independent T-Test showed no significant difference between the percentage of time spent pausing for people speaking with another human (mean=38.6, SD=14.4) compared to people speaking with the virtual avatar (mean=30.02, SD=14.8), t(41)=1.845, p=0.072). Therefore, neither group was silent for more time during the total narrative.

3.4 Pause-to-word ratio

A Mann Whitney U test indicated no significant difference (U=208.5, p=0.405) between the pause-to-word ratio for people speaking with another human (median = 14.07) and those speaking with the virtual avatar (median=14.37). Therefore, neither group paused more frequently than the other when pause frequency was standardised by the number of words spoken.

4 Discussion

The research project explored whether there are differences in the properties of speech during narratives about a seizure between human-human or human-computer interactions. We observed that there was a significant difference in the speech rate and word count between the two contexts: people were more likely to say more and speak faster while conversing with another human compared to the virtual avatar. However, there was no difference in the frequency of pauses or the total time spent pausing throughout the narrative when these measures were standardised by the amount of words spoken and the total narration time, respectively.

These findings suggest that the presence of another person may encourage more speech during a narrative telling about what happened during a recent loss of consciousness, which could have implications for medical speech technology that relies on these properties of speech. Furthermore, the finding that there were no differences in the frequency and duration of pauses suggests that these measures may remain consistent between the two contexts. Differences in how people pause can assist the differential diagnosis between people with epilepsy and nonepileptic seizures [17], and these findings suggest that people should not pause differently when they are being interviewed by the virtual avatar.

The finding that people speak slower while speaking with the virtual avatar supports previous research that has reported a slower rate of speech during human-computer interactions [10], although our findings demonstrate that this difference is still present in the absence of spoken responses from the computer. One potential reason that people speak slower to the virtual avatar is that people accommodate their speech because they perceive that the computer will have difficulty understanding them [18]. A second potential explanation is that people were changing their speech rate to match the speaking style of the virtual avatar, commonly known as entrainment [19]. Given that variations in the prosodic features of computer generated speech can influence the users trust of the application [20] and that robots can be more persuasive and favourable depending on their speaking style [21], future research should explore whether changing the design features of the virtual avatar and how it communicates can impact on how much information people provide.

Although there is evidence that people say less during interactions with a computer, people may still convey all the relevant information for the task. Walker et al. [22] found that people gave qualitatively similar answers to questions while speaking to a virtual agent when compared to interactions with a neurologist. Therefore, although the responses are shorter, they may still adhere to the sequential requirements of the previous turn.

There are multiple potential reasons why people say more while speaking with a person. Firstly, in our data, the neurologist does not take a turn until they are sure that the participant has finished speaking [15, 16]. Whereas this allows the participant to complete their narrative uninterrupted and is similar to the virtual avatar which cannot respond to what the participant has said, hesitating to take a turn could be interpreted by the participant as a sign that the neurologist is waiting for more information. Secondly, the normative structure of conversation may constrain participants to use explicit closing statements to indicate that they have finished their narrative. Closing statements are one reason why people produced more verbose responses in human-human interactions in previous research [8]. The participant can move on swiftly during interactions with the virtual avatar by pressing a button, which may allow them to produce a shorter response without the requirement of this response being accepted as complete by the other co-participant. Thirdly, a human co-participant can encourage another to produce more elaborate responses using verbal encouragement, for example “mm”, and nonverbal encouragement, for example eye contact and nodding, whereas the virtual agent used in our study was not programmed to respond actively to patient responses and therefore could not encourage further talk.

4.1 Limitations

There are multiple limitations to this analysis. Firstly, there are differences in the recording quality between the two datasets because the data was collected in different contexts and the participants were at different distances from the recording device, which made it difficult to make comparisons for other paralinguistic features, for example differences in pitch or loudness. Secondly, the analysis only used recordings of people talking about their experience of TLOC. Although this allowed for a direct comparison between the two groups, making comparisons of speech recordings for other health conditions across the two contexts would result in more generalisable finding. Therefore, future research should explore human-human versus human-computer interaction differences for other health conditions. Finally, a different sample was used for the human-human and human-computer interactions. There is a possibility that the group differences are caused by individual differences between participants. Future research should explore these group differences using a within-participant design to investigate whether individuals change how much they say between the two contexts.

5 Conclusions

The objective of this analysis was to compare properties of speech between human-human and human-computer interactions during medical interviews. We observed that people speak faster and say more during interactions with a person compared to interactions with a computer, but there was no difference in the frequency and duration of pauses. These findings concur with previous research showing that people speak differently while interacting with a computer and demonstrate the importance of considering how properties of speech may change when transitioning from human led to computer led interviews. Future research should continue to explore how properties of speech change between these two types of interactions for other health conditions to allow researchers to make inferences about how the performance of a medical diagnostic system that relies on speech will change if there is a transition between these two methods of collecting the speech data.

6 Acknowledgements

This research is funding by Epilepsy Research UK in the form of a PhD scholarship.

References

[1] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[2] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, “Alzheimer’s disease and automatic speech analysis: a review,” *Expert systems with applications*, vol. 150, p. 113213, 2020.

[3] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis, “Speech technology for healthcare: Opportunities, challenges, and state of the art,” *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 342–356, 2020.

[4] W. Pan, J. Wang, T. Liu, X. Liu, M. Liu, B. Hu, and T. Zhu, “Depression recognition based on speech analysis,” *Chinese Science Bulletin*, vol. 63, no. 20, pp. 2081–2092, 2018.

[5] B. Wang, Y. Wu, N. Taylor, T. Lyons, M. Liakata, A. J. Nevado-Holgado, and K. E. Saunders, “Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews,” *Age (years)*, vol. 44, no. 17, pp. 34–21, 2020.

[6] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen, “Diagnosing people with dementia using automatic conversation analysis,” in *Proceedings of interspeech*. ISCA, 2016, pp. 1220–1224.

[7] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, “An avatar-based system for identifying individuals likely to develop dementia,” in *Interspeech 2017*. ISCA, 2017, pp. 3147–3151.

[8] A. Johnstone, U. Berry, T. Nguyen, and A. Asper, “There was a long pause: influencing turn-taking behaviour in human-human and human-computer spoken dialogues,” *International Journal of Human-Computer Studies*, vol. 42, no. 4, pp. 383–411, 1995.

[9] E. Raveh, I. Steiner, I. Siegert, I. Gessinger, and B. Möbius, “Comparing phonetic changes in computer-directed and human-directed speech,” *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 42–49, 2019.

[10] M. Cohn, K.-H. Liang, M. Sarian, G. Zellou, and Z. Yu, “Speech rate adjustments in conversations with an amazon alexa socialbot,” *Frontiers in Communication*, vol. 6, 2021.

[11] L. Bell, J. Gustafson, and M. Heldner, “Prosodic adaptation in human-computer interaction,” in *Proceedings of ICPHS*, vol. 3. Citeseer, 2003, pp. 833–836.

[12] N. Suzuki and Y. Katagiri, “Prosodic alignment in human–computer interaction,” *Connection Science*, vol. 19, no. 2, pp. 131–141, 2007.

[13] M. Brignole, “Management of transient loss of consciousness in emergency department: what’s new from 2018 european society of cardiology guidelines?” *Emergency Care Journal*, vol. 14, no. 1, 2018.

[14] I. A. Kotsopoulos, M. C. de Krom, F. G. Kessels, J. Lodder, J. Troost, M. Twellaar, T. van Merode, and A. J. Knottnerus, “The diagnosis of epileptic and non-epileptic seizures,” *Epilepsy research*, vol. 57, no. 1, pp. 59–67, 2003.

[15] M. Schwabe, S. J. Howell, and M. Reuber, “Differential diagnosis of seizure disorders: a conversation analytic approach,” *Social science & medicine*, vol. 65, no. 4, pp. 712–724, 2007.

[16] M. Schwabe, M. Reuber, M. Schondienst, and E. Gulich, “Listening to people with seizures: how can linguistic analysis help in the differential diagnosis of seizure disorders?” *Communication & medicine*, vol. 5, no. 1, p. 59, 2008.

[17] N. Pevy, H. Christensen, T. Walker, and M. Reuber, “Feasibility of using an automated analysis of formulation effort in patients’ spoken seizure descriptions in the differential diagnosis of epileptic and nonepileptic seizures,” *Seizure*, 2021.

[18] R. Smiljanić and A. R. Bradlow, “Speaking and hearing clearly: Talker and listener factors in speaking style changes,” *Language and linguistics compass*, vol. 3, no. 1, pp. 236–264, 2009.

[19] Š. Beňuš, “Social aspects of entrainment in spoken interaction,” *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[20] Š. Beňuš, M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, and R. Levitan, “Prosodic entrainment and trust in human-computer interaction,” in *Proceedings of the 9th International Conference on Speech Prosody*. International Speech Communication Association Baixas, France, 2018, pp. 220–224.

[21] K. Fischer, O. Niebuhr, L. C. Jensen, and L. Bodenhagen, “Speech melody matters—how robots profit from using charismatic speech,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–21, 2019.

[22] T. Walker, H. Christensen, B. Mirheidari, T. Swainston, C. Rutten, I. Mayer, D. Blackburn, and M. Reuber, “Developing an intelligent virtual agent to stratify people with cognitive complaints: A comparison of human–patient and intelligent virtual agent–patient interaction,” *Dementia*, vol. 19, no. 4, pp. 1173–1188, 2020.

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

Hilla-Marja Honkalammi[1], *Outi Veivo*[2], *Marjut Johansson*[3]

Department of French Studies, University of Turku, Finland

[1]himhon@utu.fi
[2]outi.veivo@utu.fi
[3]marjut.johansson@utu.fi

Abstract

Social robots bring new possibilities to education. This paper presents an analysis of young learners' interactions in robot-assisted language learning (RALL) and seeks to describe how they give each other advice in foreign language speaking situations. Especially when a problem arises due to insufficient languages skills, the learners engage in problem-solving negotiations. The data consist of eight video-recorded learning situations where eight pairs of 10 to 12 years old children interact in English as a foreign language (EFL) with a robot. This paper presents microanalyses on advice-giving situations where the learners help each other and succeed in their common task of answering the robot's questions correctly. These microanalyses show that the learners give each other normative and epistemic advice. The results of the present study suggest that interaction problems in RALL situations lead to fruitful problem-solving interactions between the learners.

1 Introduction

Over the last few years, social robots have become increasingly common, and in this development, they have also been introduced as a tool for foreign language teaching. Although social robots have been largely studied, there are relatively few studies on robot-assisted language learning (RALL) [1], [2]. In this paper, we examine advice-giving situations in RALL as part of the research project RoboLang of the University of Turku [3].

We are interested in RALL situations where primary school children interact with a robot during English as a foreign language (EFL) classes. In different phases of the learning situation, the learners often encounter interaction problems with the robot, and the interaction breaks down [4]. The learners therefore find themselves in problem-solving situations. In this paper, we focus on the advice that the learners give each other to solve problematic situations in the child–robot interaction (CRI). Repair and advice have been studied especially in conversation analysis [5]–[7] but have not yet been examined in the context of learner interaction in RALL.

In this paper, our main aim is to analyse the ways in which the learners help each other in CRI. Namely, we examine what kind of advice young learners give one another in RALL. We will start by outlining the literature on social robots in language learning and on advice-giving. We will proceed by describing the data and analysis. Finally, we will go on to discuss the ways young learners help each other to succeed in RALL tasks.

2 Background

2.1 Social robots in language learning

According to recent research, robots provide new opportunities for learning [1], [2]. Robots can have a positive effect on learning outcomes [1], [8]. Their effectiveness may emerge from their social behaviour, especially their multimodal interaction where embodiment and gestures enrich the interaction [1]. Robots can also have a positive impact on the learners' motivation [2], [9]. Robots can arouse learners' curiosity, engage them in realistic dialogue practice and have features that can support learning [8]–[10]. For example, repeatability is helpful for comprehension and pronunciation [10]. The anthropomorphism also encourages speakers to treat robots as real speakers

DOI: 10.6094/UNIFR/223816

while proposing a situation where the learners do not have to worry that the robots might tire, laugh, or scold them [10]. These elements have been found to reduce foreign language anxiety, which can be a crucial obstacle to using the foreign language [1], [9], [11].

2.2 Advice-giving in language learning

Language learning can be perceived as a socio-cognitive process and a social practice [12], [13]. Studying classroom interactions, for example, negotiations of meaning, repair, and advice, is therefore a means of understanding learning [14].

Advice is something that "describes, recommends, or otherwise forwards a preferred course of future action" [5, p. 368]. Advice can be defined as having two dimensions: normative and epistemic [5], [7]. The normative dimension refers to the prescriptive aspect of advice that highlights the preferred course of action [5], [7]. For instance, a health advisor may advise a new mother with an imperative: *No, always be very very quiet at night.* [5, p. 387]. The epistemic dimension, on the other hand, refers to the knowledge asymmetry between the giver and the recipient of the advice, that is, the advisor knows more than the advisee [5], [7]. For example, advice may be given by sharing knowledge: *The hospital recommend that she shouldn't start solids until she's four months.* [5, p. 387].

Further, advice can be divided by its initiator [5]. Advice may be self-initiated, that is, the person experiencing the problem requests advice and starts the situations [5]. For instance, a new mother may ask: *Shall I let her tell me when she's hungry?* and receive advice: *Well yes, that's sensible.* [5, p. 371]. Other-initiated advice is initiated by the advice giver, who may inquire about the need for advice, for example, with a question *Are you doing your exercises?* and after a problem-indicative response continue with an assessment *I think it's quite important to [...]* [5, p. 383].

To our knowledge, advice-giving in RALL has not been previously researched, but previous studies note that the robot may initiate cooperation and collaborative learning between the human participants [8], [15]. It has also been observed that communication breakdowns are common in RALL as there are two major factors contributing to them: the learners' linguistic and the robot's technical limitations [4]. Therefore, advice and collaboration are important for the success of the human–robot interaction. In this study, our research question is: "What kind of advice do young learners give each other in RALL?" More specifically, we set out to explore to what extent learners use self- and other-initiated advice in RALL and how do the forms of advice vary in their normative and epistemic dimensions.

3 Data and methods

The data for this study were collected in the University of Turku's research project RoboLang in 2019. The data analysed in this study consist of 8 learning situations – approximately 17 minutes each, for a total of 2 hours and 19 minutes – where two learners interact with the robot. The data were collected during EFL lessons in a Swedish-speaking school in Finland. Before the data collection, informed consent forms were obtained from the local school authorities and from the participants' parents. The learning situations were recorded with two cameras and one audio recorder. There were 16 participants, comprising 7 girls and 9 boys. They formed 5 same-sex and 3 opposite-sex pairs. They were 10–12 years old, aged 11.25 on average.

In these learning situations, the learners met the humanoid robot NAO 6 (by SoftBank Robotics) for the first time. The learning application used on NAO was the Elias Robot application by Utelias Technologies. The application is used via a laptop showing images related to the speaking exercises. By using the laptop, the learners can also ask the robot to repeat a word or a sentence and move on to the next exercise. In the pre-programmed lessons, the robot proposes simple repetition tasks using individual words and question–response phrases to teach vocabulary and conversational structures. At the end of the learning situation, the learners have the possibility to discuss the lesson with the robot by asking general questions or questions related to the lesson.

The participants take turns interacting with the robot. The teacher who is present in the RALL situation manages the learning situation, for example, by mediating between the robot and the children and providing advice in some interaction breakdown situations. The space is organised so that the robot is in the middle of the teacher and the learners.

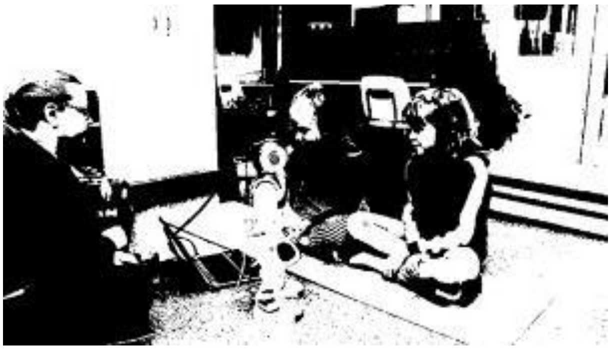


Figure 1: Spatial organization of the RALL situations.

The recorded videos were transcribed using a simplified version of the multimodal transcription conventions by Jefferson [16]. The conventions were simplified to focus on the content of the interaction (see section 8 *Transcription conventions*). The speakers were marked by the following codes: T for the teacher, R for the robot and C for the children. From these transcriptions we searched for advice sequences based on Heritage and Sefi's definition [5]. After identifying the advice-giving situations, we effected sequence analysis on them. Next, we will present you the results of our analysis.

4 Results

4.1 Overview

The analysis of the data reveals that the learners encounter interaction problems with the robot. These interaction problems often arise from the pre-programmed structure of the RALL dialogues: learners need to utter the answers expected by the robot in order to proceed in the lesson [4]. Further, these problems arise due to the learners' limited linguistic resources [4]. For example, problems emerge from pronunciation errors or forgetting the target word. These problems emerge either before or during the interaction with the robot. When these kinds of problems occur in human–robot interaction and when the learner talking with the robot does not know how to proceed, the children may help each other. That is, one learner may start to give advice to the other. Advice may be necessary before talking to the robot, for example, when the learner does not remember what he or she needs to say, or after talking to the robot when the robot does not validate the learner's response.

In our data, we identified 320 advice-giving situations. In 71 % of these situations (n = 226), advise was given by the teacher. In 29 % of the cases (n = 94), learners were giving advice to each other. Out of these 94 situations, 53 % were epistemic and 47 % normative pieces of advice. The epistemic pieces of advice include translations, clues, and correct answers, while the normative ones include, for example, imperatives and verbs of obligation. Also, when analysing the sequences, we took note of who initiated the advice. Out of these 94 advice-giving situations between learners, 56 % were other initiated and 44 % self-initiated.

4.2 Microanalyses

Next, we will present three examples of the analysed advice-giving situations. Example 1 shows an other-initiated epistemic advice that takes the form of a simple question-response sequence. In this example, the learners have already repeated the phrases once and are trying to remember the phrase in question from a picture shown on the laptop. The phrase that the robot expects in this situation is a negative response to the question, *Do you have any brothers?*. The negotiation occurs before the interaction with the robot, due to the learner's limited vocabulary.

Example 1

```
1     C1: *presses "next" on the laptop*
2     T:  *sits down*
3     C1: {C2's name} *beckons C2*
4 →   C2: *°vad säger jag°* What do I say?
          *looks at T*
5 →   C1: *°no, I don't°*
          *looks at C2*
6     C2: no, I don't
7     R:  no. I don't
8     C1: *presses "next" on the laptop*
```

```
(Video 1)
```

In Example 1, the advice-requesting and advice-giving take the form of a simple open question and the response to it. On line 3, C1 initiates the advice by mentioning C2's name and beckoning with a hand gesture. After this initiation, C2 requests advice by asking quietly, *vad säger jag* (*what do I say*). He directs the question to the teacher with a glance (line 4), but it is the other learner present in the situation, C1, who answers it. In answering the question, C1 gives C2 an epistemic advice by telling C2 the phrase that the robot expects to hear. In this situation, the recipient of the advice recognises his ignorance by asking an open question.

However, the recipients of the advice often try to avoid recognising their ignorance and the knowledge asymmetry in the situation, for example, by displaying knowledge themselves. An illustration of this is given in Example 2, where the learner tries to recall a target word.

```
Example 2
1      C11: *presses "next" on the laptop*
2 →    C11: *det var mor var det* It was mom,
            wasn't it?
            *looks at C12*
3 →    C12: *nods* juu Yeah.
4 →    C11: °jag tror det° I think so.
5      C11: mother
6      R:   mother
7      C11: *presses "next" on the laptop*
(Video 6)
```

In Example 2, the request for advice, in this example also the initiation of the advice sequence, is not an open but a closed question. The request is formulated as an assertion, *det var mor* (*it was mom*), and an added tag question, *var det* (*wasn't it*). C12 confirms the proposition twice – with a nod and an affirmation, *juu* (*yeah*). This type of advice is also epistemic, but the recipient has demonstrated his own knowledge by providing a proposition in the request. C11 only asks for confirmation of knowledge, and therefore, the epistemic asymmetry between the advisor and the advisee is not as substantial as in Example 1. Further, C11 affirms his knowledge again in his response to the advice with *jag tror det* (*I think so*).

In Example 3, the learning situation is starting, and the learners are repeating target words after the robot. The robot provides two alternatives for the word, but the learner only needs to repeat one of them.

```
Example 3
1      C15: *presses "next" on the laptop*
2      R:   mother. mom
3      C16: moder mom
4      C16: moder
5 →    C15: °säg på nytt bara° Just say again.
6      C16: moder
7      R:   *nods*
8      C16: moder
9 →    C15: *°säg bara mom°* Just say mom.
            *looks at C16*
10 →   C16: mom
11     R:   mother. mom
12     T:   jes Yep.
13     C15: *presses "next" on the laptop*
(Video 8)
```

In Example 3, C16 begins by repeating both alternatives and then chooses the first one, *mother*. However, C16 pronounces the word incorrectly; the unfamiliar EFL phoneme /ð/ is replaced by the familiar phoneme /d/. On line 5, C15 initiates advice by encouraging C16 to repeat the word. This advice takes the form of a strong normative advice – an imperative *säg på nytt bara* (*just say again*). C16 acts on the advice but the robot still does not validate the answer and only nods its head. This is a sign that it is detecting voice without understanding it. When the situation does not resolve itself by simple repetition of the word, C15 offers advice a second time. This time, the normative advice form is joint to a proposition to use the other, simpler alternative *mom* that does not include the problematic phoneme /ð/. The advice is therefore an imperative, *säg bara mom* (*just say mom*) where the advisor provides an alternative course of action. When C16 accepts the advice by conforming to it, the robot immediately accepts the word.

5 Discussion

Combining the study of interaction with social robots, our research shows that it is most often the teacher who gives advice in RALL, but that the young learners give each other advice as well. When a learner is experiencing a possible problem in the CRI, they or their peer learner may initiate advice sequences which can include both normative and epistemic forms of advice. In learner-learner situations, there were no significant differences in the proportions of who initiated the advice (44 % self-initiated and 56 % other-initiated) or which dimension of advice was used (47 % normative and 53 % epistemic).

It seems that working with the robot creates learning situations which allow peer interaction and collaboration between young learners. This may be because the learners treat the learning situation with a robot as a shared task and not an individual one, even though they take turns in answering the robot. Consequently, the learners work together to solve the conversational breakdowns in CRI, give each other advice and engage in cooperative learning. As the scope of this study does not allow comparison between RALL and non-RALL foreign language learning situations, further research is needed to determine the robot's influence on the peer interaction.

In this study, we found that social robots can provide practice for foreign language interaction itself but also for problem-solving and for peer interaction (see also [8]). Our results suggest that social robots may enable meaningful learning interaction even when the interaction with the robot breaks down; the interaction problems in RALL can lead to fruitful problem-solving interactions between the learners. RALL situations may therefore be profitable even when the interaction seems cumbersome. Consequently, the robots' limitations should not deter from using social robots as tools for language learning. This study provides an introductory glimpse into peer interaction between learners in RALL – a topic that could help us gain a better understanding of the possibilities of social robots and collaborative language learning.

7 Acknowledgements

We would like to thank Anni Nieminen for contributing to data collection and transcription, and the research group RoboLang for their useful comments.

8 Transcription conventions

The recorded videos were transcribed with simplified conventions of Jefferson [16].

C1	Child 1
T	Teacher
R	Robot
**	Gesture marked with the corresponding stretch of talk
.	Micropause
°°	Low volume
{C2's name}	Anonymised name
What do I say?	Translation from Swedish
→	Point of interest

References

[1] T. Belpaeme *et al.*, "Guidelines for Designing Social Robots as Second Language Tutors," *International Journal of Social Robotics*, vol. 10, no. 3, pp. 325–341, 2018, doi: 10.1007/s12369-018-0467-6.

[2] R. van den Berghe, J. Verhagen, O. Oudgenoeg-Paz, S. van der Ven, and P. Leseman, "Social Robots for Language Learning: A Review," *Review of Educational Research*, vol. 89, no. 2, pp. 259–295, 2019, doi: 10.3102/0034654318821286.

[3] "RoboLang | RoboLang," *University of Turku*. https://sites.utu.fi/robolang/ (accessed Aug. 02, 2021).

[4] O. Veivo and M. Mutta, "Communication breakdowns in child robot interaction in RALL," forthcoming.

[5] J. Heritage and S. Sefi, "Dilemmas of advice: Aspects of the delivery and reception of advice in interactions between health visitors and first-time mothers," in *Talk at work: Interaction in institutional settings*, P. Drew and J. Heritage, Eds. Cambridge: Cambridge University Press, 1992, pp. 359–417.

[6] E. A. Schegloff, G. Jefferson, and H. Sacks, "The Preference for Self-Correction in the Organization of Repair in Conversation," *Language*, vol. 53, no. 2, pp. 361–382, 1977, doi: 10.2307/413107.

[7] C. Shaw and A. Hepburn, "Managing the Moral Implications of Advice in Informal Interaction,"

Research on Language and Social Interaction, vol. 46, no. 4, pp. 344–362, 2013, doi: 10.1080/08351813.2013.839095.

[8] O. Engwall and J. Lopes, "Interaction and collaboration in robot-assisted language learning for adults," *Computer Assisted Language Learning*, 2020, doi: 10.1080/09588221.2020.1799821.

[9] N. Randall, "A Survey of Robot-Assisted Language Learning (RALL)," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 1, pp. 1–36, Jan. 2020, doi: 10.1145/3345506.

[10] C.-W. Chang, J.-H. Lee, P.-Y. Chao, C.-Y. Wang, and G.-D. Chen, "Exploring the Possibility of Using Humanoid Robots as Instructional Tools for Teaching a Second Language in Primary School," *Journal of Educational Technology & Society*, vol. 13, no. 2, pp. 13–24, 2010, [Online]. Available: https://www.jstor.org/stable/jeductechsoci.13.2.13

[11] M. Alemi, A. Meghdari, and M. Ghazisaedy, "The Impact of Social Robotics on L2 Learners' Anxiety and Attitude in English Vocabulary Acquisition," *International Journal of Social Robotics*, vol. 7, no. 4, pp. 523–535, 2015, doi: 10.1007/s12369-015-0286-y.

[12] G. Kasper, "Beyond Repair: Conversation Analysis as an Approach to SLA," *AILA Review*, vol. 19, no. 1, pp. 83–99, 2006, doi: 10.1075/aila.19.07kas.

[13] L. Mondada and S. Pekarek Doehler, "Interaction sociale et cognition située: quels modèles pour la recherche sur l'acquisition des langues?," *Acquisition et interaction en langue étrangère*, vol. 12, 2000.

[14] O. Sert, *Social Interaction and L2 Classroom Discourse*. Edinburgh: Edinburgh University Press, 2015. doi: 10.3366/j.ctt1g09vt3.

[15] A. Ahtinen and K. Kaipainen, "Learning and Teaching Experiences with a Persuasive Social Robot in Primary School – Findings and Implications from a 4-Month Field Study," in *Persuasive Technology. Designing for Future Change*, vol. 12064, S. B. Gram-Hansen, T. S. Jonasen, and C. Midden, Eds. Springer, Cham, 2020, pp. 73–84.

[16] G. Jefferson, "Glossary of Transcript Symbols with an Introduction," in *Conversation Analysis: Studies from the First Generation*, G. H. Lerner, Ed. Amsterdam: John Benjamins, 2004, pp. 13–31.

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

Xinyi Chen[1], *Andreas Maria Liesenfeld*[12], *Shiyue Li*[1], *Yao Yao*[1]

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

y.yao@polyu.edu.hk

Abstract

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 remarkably 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. The results showed that disfluent speech (surrounded by "um"/"uh") boosted memory retention only in pre-recorded speech for a retelling task but not in synthesized speech. We discuss the implications of current findings and possible directions of future work.

1 Introduction

With the increasing use of smart technology, more and more people become experienced with interacting with voice-AI systems, such as Siri and Amazon Alexa. While the effectiveness of such systems is impressive, probably few people would agree that talking to a machine is exactly the same as talking to a real person. There are two perspectives one can take when considering the differences between human-computer and human-human conversations. The first one regards how human-like the existing voice-AI systems are. Intelligibility aside, the speech produced by a voice-AI is mostly fluent, standard, and void of emotions, whereas real humans are often disfluent, accented, and sometimes emotional in spontaneous conversations. Such discrepancies have led to calls for a more diverse and realistic design of AI voice [1] and attempts to build voice-AI systems that are not completely fluent [2, 3, 4, 5, 6, 7].

The second perspective regards whether human talkers would show the same speech and processing behavior when interacting with a voice-AI as with a human interlocutor. Cohn, Zellou and colleagues [8, 9] reported evidence of phonetic imitation and speech rate entrainment by human talkers in human-computer interaction, similar to what has been observed in human-human interaction. However, a recent study by [10] found that when working on a picture naming task with a robot partner, human talkers did not show the partner-elicited inhibitory effects reported for the same task when speakers were paired with a human partner.

Building on previous research, the current study aims to explore the effects of disfluency in human-computer conversations. We implemented a speech robot that could produce disfluent speech, and investigated how human talkers would process disfluent robot speech. Instead of a voice-AI, we employed Furhat (https://furhatrobotics.com/), a humanoid robo with the ability to produce not only speech but also facial expressions, head and lip movement, and eye contact, which is purportedly more realistic and engaging than a voice-AI [11, 12].

The experiment we report in this paper focuses on the effects of disfluency on memory recall. Among the existing linguistic literature on disfluency in natural speech and its impact on comprehension [13, 14], a number of studies [15, 16] reported that information presented with disfluency markers (e.g. "um"/"uh") tends to be better remembered and recalled by the listener compared to information presented in fluent speech, presumably due to heightened attention induced by disfluency markers. However, [17] failed to find the same facilitatory effect of disfluency in their web-based replication study. In the current study, we attempt to replicate the memory recall experiment with human-computer conversations. We hypothesize that the same memory advantage associated with disfluency would be observed when a robot conversational partner produces disfluent speech. Since the existing text-to-speech (TTS) systems may not be able to produce natural-sounding disfluent speech, we included both pre-recorded speech produced by a human reader and synthesized speech generated by a TTS system.

2 Methods

The main task was a human-computer dialogue in Mandarin Chinese, during which the human participant would be asked to recall the information that Furhat had presented earlier. The experiment adopted a 2×2 design, varying the type of Furhat speech (pre-recorded natural speech vs. synthesized speech) and the presence of disfluency (fluent vs. disfluent). Natural speech was recorded by a female Mandarin speaker in her 20s (i.e., the reader), while synthesized speech was generated using the Amazon Polly TTS system with the Zhiyu Chinese female voice. Disfluent speech was featured by the insertion of fillers “um” and “uh” at utterance-initial positions every two or three sentences. The inserted fillers’ average duration is 0.42s in the synthesized voice condition and 0.63s in the pre-recorded voice condition. There is no added surrounding silence, consistent with the findings in [18] of fillers’ environment in Mandarin Chinese. The story duration by experiment condition is in table 1. Each participant was randomly assigned to one of the four conditions. Mean memory accuracies were compared across conditions. In the space below, we report more details of the experimental methods.

Table 1: Story duration(seconds) by experiment condition

Experiment condition		Story 1	Story 2	Story 3
Pre-recorded speech	Fluent	79	79	121
	Disfluent	82	81	124
Synthesized speech	Fluent	91	84	123
	Disfluent	93	89	126

2.1 Participants

All the participants are native Mandarin speakers from Mainland China recruited from a local university in Hong Kong. The study was approved by the ethics committee of the Hong Kong Polytechnic University, and all the participants gave written consent prior to the experiment. None of the participants studied linguistics, psychology, or computer science; in general, the participants had little or no prior experience interacting with a humanoid robot, although some had used a voice AI such as Siri, Google Assistant, or Amazon Alexa before. In this paper, we report the results with 57 participants (22M, 35F; 18-36 y.o., mean = 24.36, sd = 4.51), 15 for the pre-recorded disfluent condition, and 14 for the other three conditions.

2.2 Materials

The critical materials of this experiment are three stories for the memory test: a short story about the *Little Prince* (Story 1), a fantasy story constructed by the authors (Story 2), and a short story from *Alice in the Wonderland* (Story 3) that was translated from one of the stories used in [15]. Each story had a few hundred Chinese characters (Story 1: 293; Story 2: 324; Story 3: 499). In the disfluent version, “um”/“uh” was inserted at utterance-initial positions every two or three sentences with equal chance between the two markers. Both Story 1 and Story 2 were followed by two multiple-choice questions (each with four choices) regarding some factual detail presented in the story (e.g., “What is the nationality of the author of *Little Prince*? American, British, German, or French?”). Sentences that contained answers to the questions all appeared with an utterance-initial “um”/“uh” in the disfluent version of the story. After hearing Story 3, the participant would be asked to retell the story with as much detail as they could remember. Six (out of 14) plot points in Story 3 occurred with an utterance-initial “um”/“uh” in the disfluent version (crucial), while the remaining 8 plot points did not vary between the fluent and disfluent versions (control).

The natural speech stimuli were recorded in a soundproof booth, using an AKG C520 head-mounted microphone connected to a UR22MKII interface. Before the recording, the reader listened to a sample of the synthesized speech so that she could match in speech rate and style in her own production. To create natural disfluent stimuli, we elicited naturally produced tokens of “um” and “uh” by asking the reader to retell the stories from memory, and then inserted the clearest disfluency tokens to the designated locations in the fluent productions.

The TTS system was overall successful in generating fluent synthesized speech, but the synthesis of disfluent speech proved to be challenging. This is not surprising, given the lack of attention to the modeling of naturally produced disfluencies in TTS systems. When conventional spellings of disfluency markers (“um”/“uh”/“恩”/“啊”/“額”/“哦”) were inserted into the Chinese text, the resulting synthesized speech was so unnatural that intelligibility was greatly compromised. We ended up using two rare characters “㝗” (for “um”) and “馬我” (for “uh”), which gave the best synthesis results among all the characters with similar pronunciations.

2.3 Procedure

The human-computer dialogue task took place in a soundproof booth, with Furhat (410 mm (H)×270 mm (W)×240 mm (D)) placed on a table against the wall and the participant seated in a chair about 85 cm away, facing Furhat and roughly at the same eye level. The session was recorded by two video cameras, placed on nearby tables and directed at Furhat and the participant, respectively. The participant’s verbal responses were also recorded by Furhat’s built-in microphone. After the participant sat down, upon detecting the presence of the participant, Furhat would “wake up” from the sleep state and initiate the dialogue routine. The routine consisted of five sections, all led by Furhat: (1) greeting and self introduction (e.g., “My name is Furhat. We will play a game today.”), (2) small talk (e.g., “Have you spoken to a robot before?”), (3) practice multiple-choice questions (e.g., “Which one of the following is a Chinese musical instrument?”), (4) story telling and memory test (e.g., “Next, I will tell you a story and then ask you some questions.”), (5) ending (e.g., “Thank you. The task has completed. You can leave the room now.”).

A complete session lasted about 15 minutes. The critical section for analysis is (4), which contains all three stories. The preceding sections (i.e., (1)-(3)) serve to familiarize the participant with the interaction with Furhat. Throughout the conversation, Furhat’s speech was accompanied by constant lip movements and occasional facial expressions (e.g., smiling, eyebrow movement). For comprehension, Furhat used the Google Cloud speech-to-text system; when recognition failed, we designed subroutines for Furhat to ask for maximally two repetitions from the participant for each response. Based on the participant’s responses to multiple-choice questions in sections (3) and (4), Furhat would keep track of and report the participant’s cumulative point after each answer.

After the dialogue task completed, the participant would leave the booth and complete a post-study interview with a human researcher, where the participant would evaluate their experience of interacting with Furhat and provide ratings of naturalness and friendliness for Furhat’s speech.

3 Results

3.1 Memory and recall accuracy

We analyzed the accuracy of 57 participants’ verbal responses in the memory recall test. For the multiple-choice questions after Stories 1 and 2, a correct answer gets 1 point and a wrong answer gets 0. If the participant answered wrong first and then changed to the correct answer, they would get a half point. For Story 3, we followed the grading rubrics in [15], the participant gets 1 point for the correctly remembered plot and 0 otherwise. The total points were divided by the number of questions (or plot points) to derive accuracy scores in the range of 0-100%. We built generalized mixed effects models (using the lme4 package [19]) to examine the effects of disfluency (Fluent vs. Disfluent) and type of speech(Pre-recorded vs. Synthesised) and their interaction on response accuracy, with by-participant and by-question/plot random effects. Naturalness and friendliness ratings are modeled separately in ordinal logistic regression models (using the MASS package [20]) with similar structure as the accuracy models. Figure 1 plots the mean accuracy scores by experimental condition by question category.

As shown in Figure 1, overall, participants’ response accuracy for the multiple-choice questions in stories 1 and 2 (mean accuracy = 75%, baseline = 25%) is much higher than the retelling in story 3 (mean accuracy = 49%). In addition, we observe sim-

DOI: 10.6094/UNIFR/223818

ilar patterns across stories: (1) when Furhat produces pre-recorded natural speech, disfluent speech tends to elicit higher memory accuracy than fluent speech; (2) when Furhat produces synthesized speech, disfluent speech seems to elicit lower accuracy rates; (3) in retelling task of story 3, the general accuracy rate is higher in crucial conditions than in control conditions; (4) fluent synthesized speech tends to elicit higher accuracy than fluent pre-recorded speech. Only the last pattern is significant for Stories 1 and 2 ($\beta = 1.15, p = 0.03$), while the other comparisons did not reach statistical significance (ps > 0.05). Interestingly, the pattern also holds for the control portions of Story 3, which were not affected by the disfluency manipulation. Whether this reflects an overall disfluency advantage that goes beyond local utterances or simply individual differences in baseline memory and recall performance awaits further investigation.

3.2 Post-study interview results

Figure 2 shows the naturalness and friendliness scores obtained from the post-study interviews. The naturalness rating didn't differ much across conditions except for the lowest rating for disfluent synthesized speech, suggesting that the participants were indeed sensitive to the unnaturalness of synthesized disfluent speech. The only statistically significant difference lies in comparing disfluent synthesized speech ($\beta = -1.54, p = 0.03$) and fluent synthesized speech. Meanwhile, for friendliness rating, fluent pre-recorded speech tends to be more preferred over disfluent pre-recorded speech. In contrast, the opposite trend seems present for synthesized speech, suggesting a possible compensation for the phonetic awkwardness of disfluent synthesized speech, although neither trend reaches statistical significance.

4 Discussion

This study aims to investigate whether disfluency in machine speech may influence human listeners' memory retention of linguistic information. Our results showed better memory retention in disfluent conditions only for pre-recorded speech and for retelling in story 3, but not for synthesized speech. The possible reason is that the inserted disfluency token may not sound natural enough. First, the lowest naturalness rating was found in the disfluent synthesized speech.

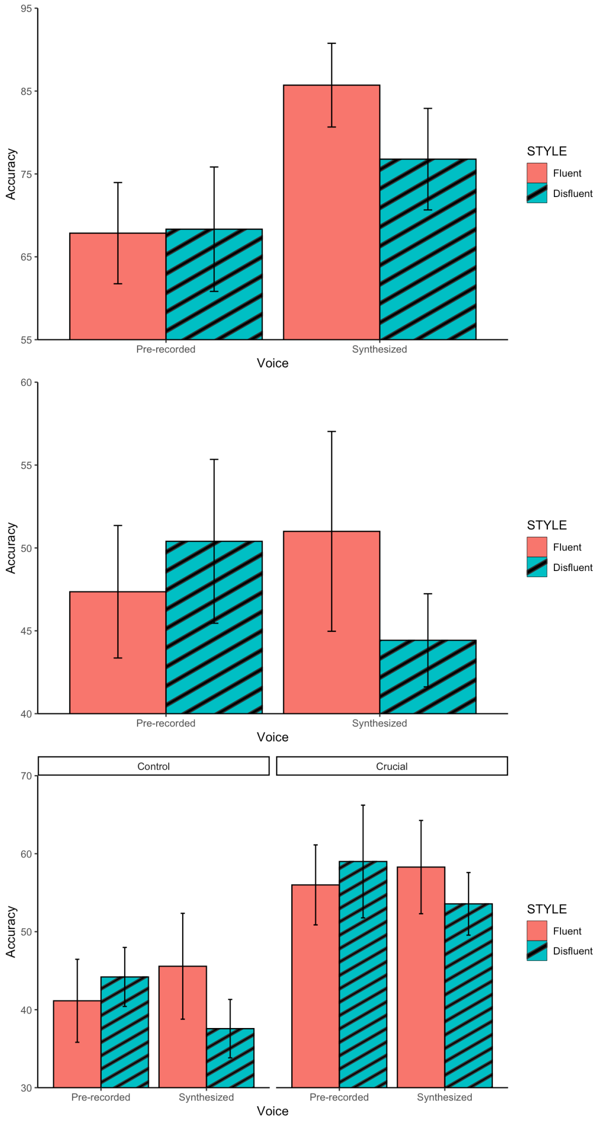


Figure 1: Mean accuracy scores (%) by experimental condition and question category.

Second, as indicated in [4], using a distinct phone for filler in the synthesized speech was preferred. Nevertheless, we didn't have this separate phone in our synthesized speech. Third, the location of inserted fillers may have an influence, in [7] , there was a higher preference of synthesizer-predicted filler pause types from location-only annotation than a more precise annotation, we can differentiate the fillers' location in future studies. Fourth, the disfluency token duration is shorter than the ones in [15], which are more than 1 second as pointed out in [17]. These being considered, the higher accuracy for crucial plots in story 3 may also not be attributed to the disfluency's facil-

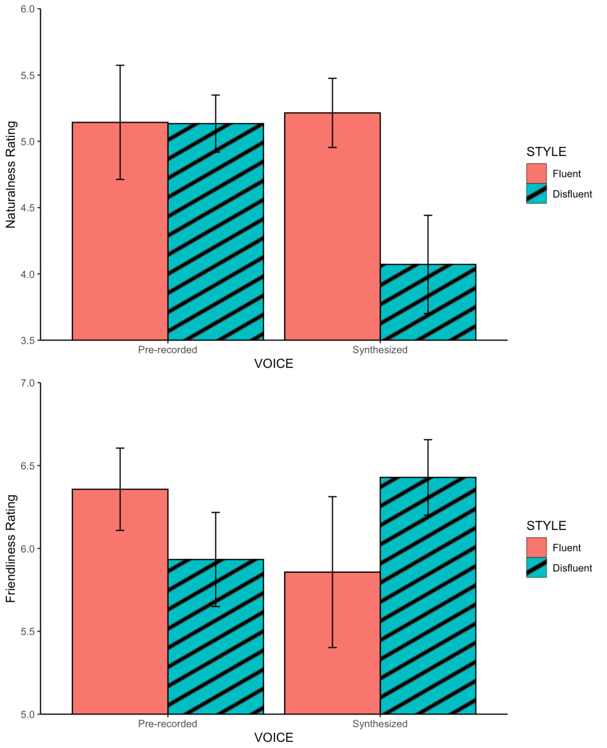


Figure 2: Mean naturalness and friendliness scores by experimental condition.

itatory effect but may be due to the plots' intrinsic difficulty to remember. The task difference may need to be taken into consideration. An unexpected pattern emerged that synthesized fluent robot speech elicited better listener memory recall than pre-recorded fluent speech across all the stories. It may be attributable to the congruence of visual and audio impression in this condition, also the highest consistency between lip movement and speech sound.

In our future work, we will continue to investigate the implementation of disfluency in machine speech and the impact of a disfluent conversational AI on the human interlocutor's comprehension and processing.

5 Acknowledgements

We thank Albert Chau, Sarah Chen, Yitian Hong, and Xiaofu Zhang for their assistance with the experiment.

References

[1] S. J. Sutton, P. Foulkes, D. Kirk, and S. Lawson, "Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–14, 2019.

[2] J. Adell, D. Escudero, and A. Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," *Speech Communication*, vol. 54, no. 3, pp. 459–476, Mar. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639311001580

[3] S. Betz, P. Wagner, and D. Schlangen, "Micro-structure of disfluencies: basics for conversational speech synthesis," in *Interspeech 2015*. ISCA, Sep. 2015, pp. 2222–2226. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2015/betz15_interspeech.html

[4] R. Dall, M. Tomalin, and M. Wester, "Synthesising Filled Pauses: Representation and Datamixing," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 7–13.

[5] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive Hesitation Synthesis: Modelling and Evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, p. 9, Mar. 2018. [Online]. Available: http://www.mdpi.com/2414-4088/2/1/9

[6] B. Carlmeyer, S. Betz, P. Wagner, B. Wrede, and D. Schlangen, "The hesitating robot - implementation and first impressions," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 77–78. [Online]. Available: https://doi.org/10.1145/3173386.3176992

[7] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019.

[8] M. Cohn, K.-H. Liang, M. Sarian, G. Zellou, and Z. Yu, "Speech rate adjustments in conversations with an Amazon Alexa socialbot," *Frontiers in Communication*, vol. 6, no. May, pp. 1–8, 2021.

[9] G. Zellou, M. Cohn, and T. Kline, "The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors," *Language, Cognition and Neuroscience*, vol. 0, no. 0, pp. 1–15, 2021. [Online]. Available: https://doi.org/10.1080/23273798.2021.1931372

[10] O. A. Wudarczyk, M. Kirtay, D. Pischedda, V. V. Hafner, J.-D. Haynes, A. K. Kuhlen, and R. Abdel Rahman, "Robots facilitate human language production," *Scientific Reports*, vol. 11, no. 1, p. 16737, Aug. 2021.

[11] L. Cominelli, F. Feri, R. Garofalo, C. Giannetti, M. A. Meléndez-Jiménez, A. Greco, M. Nardelli, E. P. Scilingo, and O. Kirchkamp, "Promises and trust in human–robot interaction," *Scientific Reports*, vol. 11, no. 1, p. 9687, May 2021.

[12] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, vol. 67, p. 101178, May 2021.

[13] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–169, 2001.

[14] J. E. Arnold and M. K. Tanenhaus, "Disfluency effects in comprehension: how new information can become accessible," in *The processing and acquisition of reference*, E. Gibson and N. J. Pearlmutter, Eds. Cambridge, Massachusetts: MIT Press, 2011, no. December, pp. 197–218.

[15] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of Memory and Language*, vol. 65, no. 2, pp. 161–175, 2011.

[16] H. R. Bosker, J. Tjiong, H. Quené, T. J. M. Sanders, and N. H. De Jong, "Both native and non-native disfluencies trigger listeners' attention," *Proceedings of Disfluency in Spontaneous Speech 2015*, pp. 1–4, 2015.

[17] B. Muhlack, M. Elmers, H. Drenhaus, M. van Os, R. Werner, M. Ryzhova, and B. Möbius, "Revisiting recall effects of filler particles in german and english," in *Proceedings of Interspeech 2021*. Brno, Czechia: Interspeech, 2021, inproceedings.

[18] Y. Zhao and D. Jurafsky, "A preliminary study of Mandarin filled pauses," p. 4, 2005.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4,' *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[20] W. N. Venables, B. D. Ripley, and W. N. Venables, *Modern applied statistics with S*, 4th ed., ser. Statistics and computing. New York: Springer, 2002, oCLC: ocm49312402.