INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Differences in Gradient Emotion Perception: Human vs. Alexa Voices

Michelle Cohn[1], *Eran Raveh*[2], *Kristin Predeck*[1], *Iona Gessinger*[2], *Bernd Möbius*[2], *Georgia Zellou*[1]

[1]Phonetics Laboratory, Linguistics, UC Davis, Davis, California, USA
[2]Language Science and Technology, Saarland University, Saarbrücken, Germany

{mdcohn, kpredeck, gzellou}@ucdavis.edu, {raveh, gessinger, moebius}@coli.uni-saarland.de

Abstract

The present study compares how individuals perceive gradient acoustic realizations of emotion produced by a human voice versus an Amazon Alexa text-to-speech (TTS) voice. We manipulated semantically neutral sentences spoken by both talkers with identical emotional synthesis methods, using three levels of increasing 'happiness' (0 %, 33 %, 66 % 'happier'). On each trial, listeners (native speakers of American English, n=99) rated a given sentence on two scales to assess dimensions of emotion: valence (negative-positive) and arousal (calm-excited). Participants also rated the Alexa voice on several parameters to assess anthropomorphism (e.g., naturalness, human-likeness, etc.). Results showed that the emotion manipulations led to increases in perceived positive valence and excitement. Yet, the effect differed by interlocutor: increasing 'happiness' manipulations led to larger changes for the human voice than the Alexa voice. Additionally, we observed individual differences in perceived valence/arousal based on participants' anthropomorphism scores. Overall, this line of research can speak to theories of computer personification and elucidate our changing relationship with voice-AI technology.

Index Terms: voice-activated personal assistants, emotion perception, human-computer interaction

1. Introduction

While the primary function of speech is to communicate a message to our interlocutor, the voice also carries other properties, including social details (e.g., region, age, gender) and even our emotional state. Whether we are happy, surprised, sad, or angry might be conveyed on an utterance [1]. Emotional expressiveness has been targeted as a way to make synthetic, text-to-speech (TTS) voices more engaging to human users [2, 3]. These efforts have concentrated primarily on synthesizing basic human emotions, including happiness, sadness, anger, fear, disgust, and surprise [4]; yet perception of such qualities in the synthesis, however, is not always clear to the listeners [4]. The parameters adjusted in emotional synthesis may be contributing to this confusion. Another contributor may be the quality of the synthetic voices; it may be the case that it is difficult for listeners to extract emotion from more robotic-sounding voices. Finally, a third contributor may be the degree to which listeners attribute *human-like emotion* to the systems – which may be due to the human-like characteristics of the system as well as individual differences in personification of the systems. The present paper examines these factors in how listeners perceive emotion in a real human and an Amazon Alexa TTS voice.

1.1. Emotion and modern voice-AI systems

In the last decade, modern voice-activated, artificially intelligent (voice-AI) systems, such as Apple's Siri, Amazon's Alexa, and Google Assistant, have become a common household interlocutor for many human users, particularly in the United States [5]. These systems engage users in a variety of functional and social tasks. For example, users may ask Alexa to "turn off the light", "tell a joke", or even have a conversation [6]. Prior work suggests that humans apply social knowledge to their speech interactions with voice-AI systems, such as gender-related asymmetries [7]. There is some initial evidence that users may also be perceptive to emotional expressiveness in voice-AI systems: speakers vocally align to emotionally expressive productions by the Alexa voice [8] and rate conversations with an Amazon Alexa socialbot higher when the bot uses emotionally expressive interjections [9], but neither of these studies employed a direct human comparison. The present study addresses this gap, examining whether listeners similarly perceive gradient emotion conveyed in TTS and natural human productions.

1.2. Emotion and CASA

Comparing responses to emotion produced by human and voice-AI interlocutors can speak to computer personification theories, such as the *Computers Are Social Actors (CASA)* theoretical framework [10, 11], which holds that humans treat technology as a social actor in interactions and apply social rules and norms from human-human interaction (HHI). This personification of technology is thought to be automatic, subconscious, and driven by the fact that device interaction often involves similar aspects as HHI. For example, participants assigned higher trustworthiness and likeability ratings to a computer system that displayed more empathetic emotion than one that did not [12]. In another study, participants showed different negotiation strategies when haggling with a 'happy' or an 'angry' computer system, in line with emotion-based asymmetries observed in human-human negotiation [13]. Meanwhile, other work has found that negative reactions are triggered by computer behavior in the same ways that a human's actions might engender anger: after a computer system had acted unfairly in a bargaining game, participants in that interaction displayed anger and spiteful behavior toward the device [14].

In line with CASA [11], there is some evidence for similar perception of emotion produced by a human or computer. In a study examining explicit emotion identification (e.g., happy, sad, surprised, etc.) based on visual and prosodic differences, participants displayed equal responses to the 'human' or 'computer' guise [15]. In a study of facial expression, Noël *et al.* [16] found that subjects' accuracy identifying emotion for a real human face and a digital avatar was equal when context and emotional expressiveness were congruent. Following these studies, one possibility is that individuals will interpret emotional prosody similarly for human and voice-AI speech.

On the other hand, many studies exploring synthesized emotion do not make a direct human versus device comparison. It is possible that while there are similarities in the gross pat-

http://dx.doi.org/10.21437/Interspeech.2020-1938

terns of social responses toward humans and computers/robots, there may be more fine-grained nuances that are missed, particularly in using a between-subjects design [e.g., 8, 9, 12, 13, 14, 17]. For example, participants rated the emotion of synthetic and natural speech similarly when the emotional expression was congruent with the content (e.g., happy prosody with positive content) [17]. Yet, in the incongruent conditions (e.g., sad prosody with positive content), they observed differences for the human and TTS voices in the relative weight listeners gave to the prosody, relative to the content: they rated synthetic speech as 'happier' than human speech when it was produced with sad prosody and happy content. Indeed, there is some evidence that humans respond to voice-AI and human speech differently: participants display less vocal alignment toward Siri and Alexa TTS voices than human voices [18, 19]; this suggests that voice-AI systems may be a distinct type of social actor than another human. Therefore, in the present study, one prediction is that participants may show weaker emotion perception for an Alexa voice, relative to a human voice.

Additionally, many TTS emotion perception studies ask participants to classify very distinct types of stimuli (e.g., basic emotions of happiness or sadness); one unexplored question, to our knowledge, is whether the perceived *magnitude* of emotional expression is similar for human and synthetic voices. In the current study, we examine gradience in emotion perception by adapting neutral speech produced by a human and voice-AI talker (here, Amazon Alexa) at three happiness levels (see Section 2.1.2). One possibility is that listeners will be more sensitive to gradient emotional display by human voices, as human voices are more socially meaningful. Alternatively, another possibility is that listeners will display equal sensitivity for human and voice-AI voices producing multiple levels of an emotion, which would provide support for the CASA account.

1.3. Variation in personification

While the CASA account proposes an automatic mechanism of personification, there is reason to believe that any such response will vary considerably across individuals. For example, participants displayed different patterns of vocal alignment toward voice-AI (Apple Siri) voices based on their cognitive processing style [20]. In another study, individuals interacting with the same robot receptionist communicated differently depending on their attitude towards the virtual interlocutor: as being more 'human-social' or a 'computational-tool' [21]. In the present study, we assess each participant's anthropomorphism of the virtual assistant Alexa across several dimensions, viz. humanness, naturalness, etc., in a pre-experiment survey. We predict that overall anthropomorphism scores will be related to voice-AI emotion perception, i.e., individuals with higher anthropomorphism scores are expected to be more perceptive to emotion by the Alexa voice.

1.4. Current study

In the present study, we examine whether technology personification is gradiently realized in the perception of emotion. In this experiment, we ask two principal questions: 1) Do listeners perceive acoustic variations conveying different levels of emotional state similarly for human and TTS voices?, and 2) Does an individual's gradient perception of TTS voice emotion vary according to the degree to which they personify the system; that is, are listeners better at perceiving emotion for interlocutors they deem as being more 'human-like'? While the general acoustic properties of a recorded human voice and an Alexa voice differ, we used identical parameters for both voices in the emotional synthesis system, DAVID [22]. We selected DAVID given its prior validation: listeners perceive the intended emotions (e.g., happiness, sadness, and fear) in manipulated productions [22]. Additionally, DAVID allows for specification of gradient change toward a given emotion (e.g., 66 % 'happier').

Critically, we test whether the same gradient manipulations of emotional prosody within a given voice yields similar or different changes in emotion perception across the two speakers (here, human vs. voice-AI). As our aim is to investigate the role of emotional prosody, we conducted a norming study (see Section 2.1.1) of sentences to generate our list of 'emotionally neutral' sentences; accordingly, listeners would primarily respond to the emotional properties conveyed through the voice.

2. Methods

2.1. Stimuli

2.1.1. Norming study: emotionally neutral sentences

We selected sentences that had previously been rated as emotionally 'neutral' (14 from Russ *et al.* [23]; 10 from Ben-David *et al.* [24]; and 2 from Mustafa *et al.* [25]) as well as 94 declarative sentences from the *Speech Perception In Noise (SPIN)* test [26], to a total of 120, for an online emotional valence norming study. The inclusion of the SPIN sentences permits a greater range of perceived valence. The 48 native English speakers (mean age 19.7 ±2.1 years; recruited through the UC Davis subject pool) rated the emotion in all 120 sentences, which were randomly presented on the screen one at a time with no sound. On a given trial, they saw a sentence and used a sliding scale to indicate how negative, positive, or neutral it was; the beginning, middle, and end of the spectrum were labeled with "0 = negative", "50 = neutral", and "100 = positive", respectively. The slider position reset to 50 at the beginning of each trial. The data are available as supplemental material[1].

2.1.2. Synthesizing emotion in human and Alexa voices

We selected the 15 sentences with the ratings closest to 50 (range 48 to 51, mean 49.9) from the norming study (Section 2.1.1), excluding imperatives and sentences with personal pronouns (e.g., "My T.V. has a twelve-inch screen.") that may be incongruous if produced by a voice-AI system. We also excluded two sentences with negative words (e.g., "garbage" and "shipwrecked"). The remaining 15 sentences had 4 to 8 words (mean 5.9 ±1.2). We recorded a native English female speaker producing the 15 target sentences in citation format. We generated the same 15 sentences with default US-English female Alexa voice using the Alexa Skills Kit. Recordings had a sampling rate of 44.1 kHz and were amplitude normalized[2] based on mean intensity measurements in Praat [27].

Next, we generated three 'happiness' levels (at 0 % (no change), 33 %, and 66 % happier) with the DAVID emotional synthesis platform [22] in the Max programming language [28]. We used the DAVID default values for 'happiness', including a fundamental frequency (f_0) increase of 30 cents[3], and high shelf filter (8 kHz, gain 3 dB). We passed all sentences through

[1] http://dx.doi.org/10.17632/tm2scpw8mg.1

[2] 65 dB for human, 64 dB for Alexa voices; as the Alexa samples were generated in a systematically different manner than the human recordings (i.e., not through air transmission), this normalization was relative and adjusted (by ear) by the first author.

[3] A cent is a logarithmic unit of pitch (1 octave = 100 cents)

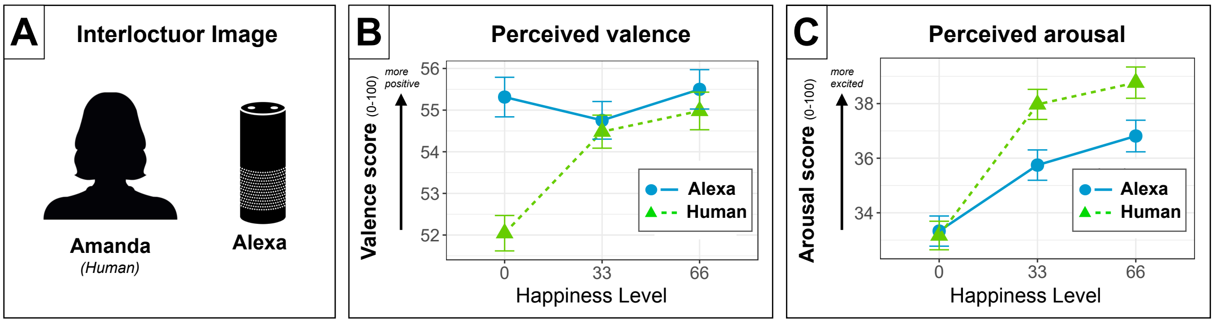

Figure 1: *(A) Human and synthetic speakers' silhouettes. The corresponding silhouette appeared on the screen for all trials within a speaker block in the human ('Amanda') or device ('Alexa') condition. (B-C) Summary of valence (B) and arousal (C) results. The blue dots and green triangles indicate the mean scores for Alexa and the human voices, respectively. Error bars show the standard error.*

the DAVID re-synthesizer at 0 %, 33 %, and 66 % of the 'happiness' parameters (e.g., 33 % increase in f_0 toward 30 cents: increase of 9.9 cents). This resulted in a total of 90 stimuli[4]. (15 sentences × 3 happiness levels × 2 interlocutors).

2.2. Participants

Participants (none of whom completed the norming task) consisted of 99 native speakers of American English, recruited from the UC Davis Psychology subject pool (70 females, 29 males; mean age 20.2 ± 2.2 years); 81 of them reported some experience using a voice-AI system.

2.3. Procedure

Subjects completed the experiment online, via Qualtrics. First, they provided basic demographic information, as well as their voice-AI usage. Next, participants completed an audio calibration step to ensure that the stimuli were audible and understandable via their computer's audio device: they heard one sentence (not used in the experimental trials) produced by each interlocutor (human and Alexa) and were asked to select what they heard out of a set of options; if their response was correct, they continued to the experimental trials; if not, they were taken to a screen that indicated that they needed to increase the volume. Participants could not continue to the experimental trials until they answered correctly.

Then, they completed a voice-AI anthropomorphism survey, adapted from Ho and MacDorman [29]. Using sliding scale response (0-100), participants heard a single sentence produced by the Amazon Alexa voice (note that the sentence was not manipulated in terms of emotion) and rated to what degree they thought the voice was machine-like/human-like, artificial/natural, eerie/comforting, and cold/warm.

In the experimental trials, participants were told that they would hear sentences produced by either an Amazon Alexa or a real person ('Amanda'), rate the sentences, and answer a few randomly presented listening comprehension questions. Participants were told that they would only hear each sentence once, and to respond as quickly and accurately as possible. Speaker condition (voice-AI/human) was divided into blocks (order counterbalanced across subjects). During all trials of each block, participants saw the corresponding Alexa/human silhouette on the screen (see Figure 1.A). On each trial, subjects heard an emotionally neutral sentence in one of the three happiness levels and rated it on two dimensions of emotion using a sliding scale: valence (0 = negative, 50 = neutral, 100 = positive) and arousal (0 = calm, 50 = neutral, 100 = excited). At the beginning of each trial, the slider position reset to 50. The sentences were only presented aurally and were randomized by happiness level. Each participant rated all 90 stimuli. Additionally, listeners heard a listening comprehension question after the experimental trials for each speaker: they heard a semantically anomalous sentence produced by the speaker (either human or Alexa) and identified the sentence from a multiple choice list. Participants needed to answer correctly to receive credit for the study. In total, the experiment took roughly 30 minutes.

2.4. Analysis

We analyzed participants' valence and arousal scores for the sentences with separate linear mixed models (LMMs), using the lme4 R package [30]. In both models, the fixed effects included HAPPINESS LEVEL (3 levels: 0 %, 33 %, and 66 % happier), INTERLOCUTOR (2 levels: human, device), and all possible interactions. Random effects included by-SUBJECT random intercepts, with by-SUBJECT random slopes for INTERLOCUTOR. The linear mixed models (sLMMs) were fit by REML t-tests and used Satterthwaite approximations to determine the degrees of freedom. The p-values were derived from the output of these fits with the lmerTest package [31].

For the anthropomorphism analysis, we calculated a composite anthropomorphism score, summing the totals for each of the responses (human-like, natural, comforting, warm) for the voice-AI; a higher score indicates greater personification. On the subset of data for the Alexa talker, we modeled valence and arousal scores in separate linear mixed models (LMMs). Main effects included ANTHROPOMORPHISM SCORE (continuous) and HAPPINESS LEVEL, their interaction, as well as by-SUBJECT random intercepts.

3. Results

Figure 1 shows the mean scores of valence and arousal for the Alexa and human voices over the three levels of happiness. The outcomes of the LMM fits (see Section 2.4) for valence and arousal are summarized in Tables 1 and 2. Valence scores were overall lower for the human speaker relative to Alexa. There was also an interaction between HAPPINESS LEVEL and INTERLOCUTOR: there was a larger increase in valence for the human talker at the higher happiness levels (33 % and 66 %) (see Figure 1.B). While the score difference between the human

[4] http://www.coli.uni-saarland.de/~raveh/Interspeech_2020/stimuli/

Table 1: *Summary of fixed effects in valence scores.*

	Coef	SE	df	t	p
(Intercept)	55.31	1.31	108	42.06	≪0.001 ***
Happ.33	−0.56	0.49	8708	−1.14	0.250
Happ.66	0.18	0.49	8708	0.38	0.710
Int.Human	−3.27	1.12	129	−2.93	0.004 **
Happ.33:Int	3.00	0.70	8708	4.32	≪0.001 ***
Happ.66:Int	2.75	0.69	8708	3.97	≪0.001 ***

Table 2: *Summary of fixed effects in arousal scores.*

	Coef	SE	df	t	p
(Intercept)	33.33	1.67	105	19.89	≪0.001 ***
Happ.33	2.42	0.55	8708	4.38	<0.001 ***
Happ.66	3.48	0.55	8708	6.31	≪0.001 ***
Int.Human	−0.16	1.39	122	−0.12	0.740
Happ.33:Int	2.39	0.78	8708	3.05	0.002 **
Happ.66:Int	2.12	0.78	8708	2.71	0.007 **

speaker and Alexa is large for non-manipulated speech, this gap is closed in the 33 % happiness level, and the scores are virtually identical in the 66 % happiness level. Moreover, the scores for Alexa are relatively stable, whereas the human scores rise sharply in the 33 % happiness level. Arousal ratings (see Figure 1.C) show a different pattern: while excitement for the two voices is equal for the non-manipulated speech (0 % happiness level), the scores for *both* speakers show an increase from 0 % to 33 % and 66 % happiness levels. This increase is larger for the human, relative to the Alexa, voice.

As for the anthropomorphism scores, we observed variation across participants (mean 131.9 ±71.2, range 0 - 300). In both mixed effects models, there were interactions between ANTHROPOMORPHISM SCORE and HAPPINESS LEVEL. Figure 2 illustrates the anthropomorphism scores in the different conditions. In the valence model, participants with higher anthropomorphism scores rated the Alexa voice as sounding more positive at the baseline happiness level, 0 % [*Coef* = 0.02, *SE* = 5.2e-03, t = 3.1, p < 0.01]. No other interactions were observed. In the arousal model, a higher anthropomorphism score was associated with less perceived excitement at the highest happiness level, 66 % [*Coef* = -0.02, *SE* = 6.0e-03, t = -3.2, p < 0.01]; no other interactions were significant for the arousal model.

4. Discussion and Conclusion

Overall, we found that listeners perceive emotion gradiently in both human and voice-AI (here, Amazon's Alexa) voices. However, this was limited to arousal ratings for the Alexa voice, while both valence and arousal ratings for the human voice rose with the increasing 'happiness' manipulations. This finding is broadly in line with the CASA theoretical framework [11, 10], as the subjects were hearing different levels of 'excitement' in both a human and an Alexa voice that were manipulated identically. Yet, these findings also illuminate a possibly limited aspect of technology personification: listeners heard variation in valence in the human voice, but not for the Alexa voice. This might indicate that users still do not expect – or are not used to – TTS voices that show this dimension of emotion. Another factor may be the nature of the task. Listening tasks are somewhat passive comparing to the typical use of voice-AI personal assistants. It is possible that the range of listeners' emotion judgments would be wider in more naturalistic, dyadic interactions. Future work exploring emotion perception across different types of interactions (e.g., more functional, more social) are needed to further explore this effect.

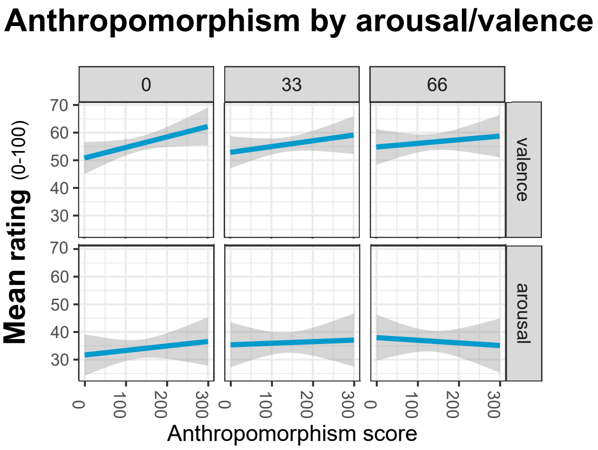

Figure 2: *Effect of anthropomorphism scores on perceived Happiness Level (0, 33, 66 %) on valence (top panel) and arousal (bottom panel) ratings of the Alexa (blue solid line).*

Additionally, we found evidence that individual variation in anthropomorphism of voice-AI mediates emotion perception of the Alexa voice: participants who displayed greater personification of the Alexa voice rated it as being more positive at baseline, while also rating the voice as sounding *less* excited at the 66 % happiness level. Our valence anthropomorphism findings, (i.e., participants who personify Alexa more tend to also rate the voice as sounding happier) are in line with research suggesting greater generalization of positive attitudes (here, more human-like qualities) to other domains [32]. While the decrease in arousal ratings was unexpected, the lack of correspondence between the valence and arousal results are consistent with prior work showing their separable effects, which are further affected by patterns of individual variation (e.g., personality, cultural background; cf. [33]). One limitation in this study is that the participants were not balanced by gender, with far more female than male raters. While we made no a priori hypotheses about how individuals might respond differently according to their gender, this may be a source of variation [34]. Future work examining different types of emotion, as well as comparing individuals of different linguistic/cultural backgrounds, genders, and even ages can further our understanding of sources of variation in the relationship between voice-AI/human emotion perception and anthropomorphism.

Overall, our findings suggest that the way humans engage with voice-AI systems is similar in some ways to humans – in perceiving increases in 'arousal' – but perception of emotion multidimensionality (i.e., both valence and arousal) appears to be limited to natural human productions.

5. Acknowledgments

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No.1911855 to MC. Funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID MO 597/6-2 and STE 2363/1-2.

6. References

[1] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion”, *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[2] C. Creed and R. Beale, “Emotional intelligence: Giving computers effective emotional skills to aid interaction”, in *Computational Intelligence: A Compendium*, Springer, 2008, pp. 185–230.

[3] A. R. F. Rebordao, M. A. M. Shaikh, K. Hirose, and N. Minematsu, “How to improve TTS systems for emotional expressivity”, in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[4] J. E. Cahn, “The generation of affect in synthesized speech”, *Journal of the American Voice I/O Society*, vol. 8, no. 1, pp. 1–1, 1990.

[5] F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, and D. Lottridge, “Understanding the long-term use of smart speaker assistants”, *Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–24, 2018.

[6] H. Fang, H. Cheng, E. Clark, A. Holtzman, M. Sap, M. Ostendorf, Y. Choi, and N. A. Smith, “Sounding board: University of Washington’s Alexa Prize submission”, *Alexa Prize*, 2017.

[7] H. Bergen *et al.*, “I’d blush if i could: Digital assistants, disembodied cyborgs and the problem of gender”, *Word and Text, A Journal of Literary Studies and Linguistics*, vol. 6, no. 01, pp. 95–113, 2016.

[8] M. Cohn and G. Zellou, “Expressiveness influences human vocal alignment toward voice-AI”, in *Proc. Interspeech 2019*, Sep. 2019, pp. 41–45. DOI: 10.21437/Interspeech.2019-1825.

[9] M. Cohn, C.-Y. Chen, and Z. Yu, “A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog”, in *SIGdial*, 2019, pp. 293–306. DOI: 10.18653/v1/W19-5935.

[10] C. Nass and Y. Moon, “Machines and mindlessness: Social responses to computers”, *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000. DOI: 10.1111/0022-4537.00153.

[11] C. Nass, Y. Moon, J. Morkes, E.-Y. Kim, and B. Fogg, “Computers are social actors: A review of current research”, *Human values and the design of computer technology*, vol. 72, pp. 137–162, 1997.

[12] S. Brave, C. Nass, and K. Hutchinson, “Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent”, *International journal of human-computer studies*, vol. 62, no. 2, pp. 161–178, 2005. DOI: 10.1016/j.ijhcs.2004.11.002.

[13] C. M. de Melo, P. Carnevale, and J. Gratch, “The effect of expression of anger and happiness in computer agents on negotiations with humans”, in *International Conference on Autonomous Agents and Multiagent Systems – Volume 3*, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 937–944.

[14] R. E. Ferdig and P. Mishra, “Emotional responses to computers: Experiences in unfairness, anger, and spite”, *Journal of Educational Multimedia and Hypermedia*, vol. 13, no. 2, pp. 143–161, 2004.

[15] C. Bartneck, “Affective expressions of machines”, in *CHI’01 extended abstracts on Human factors in computing systems*, ACM, 2001, pp. 189–190.

[16] S. Noël, S. Dumoulin, and G. Lindgaard, “Interpreting human and avatar facial expressions”, in *IFIP Conference on Human-Computer Interaction*, Springer, 2009, pp. 98–110.

[17] C. Nass, U. Foehr, S. Brave, and M. Somoza, “The effects of emotion of voice in synthesized and recorded speech”, in *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition*, AAAI North Falmouth, MA, 2001.

[18] M. Cohn, B. F. Segedin, and G. Zellou, “Imiating Siri: Socially-mediated vocal alignment to human and device voices”, in *ICPhS*, Aug. 2019, pp. 1813–1817. DOI: 10.21437/Interspeech.2019-1825.

[19] 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”, *Interspeech 2019*, pp. 4005–4009, 2019.

[20] C. Snyder, M. Cohn, and G. Zellou, “Individual variation in cognitive processing style predicts differences in phonetic imitation of device and human voices”, *Proc. Interspeech 2019*, pp. 116–120, 2019.

[21] M. K. Lee, S. Kiesler, and J. Forlizzi, “Receptionist or information kiosk: How do people talk with a robot?”, in *Proceedings of the 2010 ACM conference on computer supported cooperative work*, 2010, pp. 31–40.

[22] L. Rachman, M. Liuni, P. Arias, A. Lind, P. Johansson, L. Hall, D. Richardson, K. Watanabe, S. Dubal, and J.-J. Aucouturier, “David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech”, *Behavior research methods*, vol. 50, no. 1, pp. 323–343, 2018.

[23] J. B. Russ, R. C. Gur, and W. B. Bilker, “Validation of affective and neutral sentence content for prosodic testing”, *Behavior research methods*, vol. 40, no. 4, pp. 935–939, 2008. DOI: 10.3758/BRM.40.4.935.

[24] B. M. Ben-David, P. H. van Lieshout, and T. Leszcz, “A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury”, *Brain injury*, vol. 25, no. 2, pp. 206–220, 2011. DOI: 10.3109/02699052.2010.536197.

[25] M. B. Mustafa, R. N. Ainon, and R. Zainuddin, “EM-HTS: Real-time HMM-based Malay emotional speech synthesis”, in *ISCA Workshop on Speech Synthesis*, 2010. DOI: 10.22452/mjcs.vol23no2.3.

[26] L. L. Elliott, “Verbal auditory closure and the speech perception in noise (spin) test”, *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 6, pp. 1363–1376, 1995. DOI: 10.1044/jshr.3806.1363.

[27] P. Boersma and D. Weenink, “Praat”, *Doing phonetics by computer (Version 5.1)*, 2005.

[28] M. Puckette, *Max/msp (version 7): Cycling’74*, 2014.

[29] C.-C. Ho and K. F. MacDorman, “Revisiting the Uncanny Valley theory: Developing and validating an alternative to the Godspeed indices”, *Computers in Human Behavior*, vol. 26, no. 6, pp. 1508–1518, 2010.

[30] 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. DOI: 10.18637/jss.v067.i01.

[31] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “Lmertest package: Tests in linear mixed effects models”, *Journal of statistical software*, vol. 82, no. 13, 2017.

[32] R. H. Fazio, E. S. Pietri, M. D. Rocklage, and N. J. Shook, “Positive versus negative valence: Asymmetries in attitude formation and generalization as fundamental individual differences”, in *Advances in experimental social psychology*, vol. 51, Elsevier, 2015, pp. 97–146.

[33] P. Kuppens, F. Tuerlinckx, M. Yik, P. Koval, J. Coosemans, K. J. Zeng, and J. A. Russell, “The relation between valence and arousal in subjective experience varies with personality and culture”, *Journal of personality*, vol. 85, no. 4, pp. 530–542, 2017.

[34] A. Lausen and A. Schacht, “Gender differences in the recognition of vocal emotions”, *Frontiers in Psychology*, vol. 9, p. 882, 2018. DOI: 10.3389/fpsyg.2018.00882.