Interspeech 2022
18-22 September 2022, Incheon, Korea

Cross-Cultural Comparison of Gradient Emotion Perception: Human vs. Alexa TTS Voices

Iona Gessinger[1,2], *Michelle Cohn*[3], *Georgia Zellou*[3], *Bernd Möbius*[1]

[1]Language Science and Technology, Saarland University, Saarbrücken, Germany
[2]ADAPT Centre, University College Dublin, Ireland
[3]Phonetics Laboratory, Linguistics, University of California, Davis, USA

{gessinger|moebius}@lst.uni-saarland.de, {mdcohn|gzellou}@ucdavis.edu

Abstract

This study compares how American (US) and German (DE) listeners perceive emotional expressiveness from Amazon Alexa text-to-speech (TTS) and human voices. Participants heard identical stimuli, manipulated from an emotionally 'neutral' production to three levels of increased happiness generated by resynthesis. Results show that, for both groups, 'happiness' manipulations lead to higher ratings of emotional valence (i.e., more positive) for the human voice. Moreover, there was a difference across the groups in their perception of arousal (i.e., excitement): US listeners show higher ratings for human voices with manipulations, while DE listeners perceive the Alexa voice as sounding less 'excited' overall. We discuss these findings in the context of theories of cross-cultural emotion perception and human-computer interaction.

Index Terms: cross-cultural emotion perception, human-computer interaction

1. Introduction

The expression and perception of emotions has both universal and culturally-specific components. On the one hand, the encoding of emotions, for example in facial or vocal expression, can be linked to physical responses of the human body to the given emotional state (*push effects*) [1, 2]. Therefore, there are some aspects of emotional expression that are likely to be similar across cultures and universally decodable (cf. *universality* accounts of emotion expression/perception [3, 4]). On the other hand, there are different conventions and expectations in how emotions are expressed and perceived that exist across individual societies (*pull effects*) [1, 2]. Hence, cross-cultural comprehension of emotional expression may not be equivalent (cf. *culture-specific* accounts of emotion expression/perception [5, 6]).

Furthermore, expectations for how certain types of speakers should convey emotion also vary cross-culturally. For example, [7] found that American respondents had more favorable attitudes about robots showing emotion than German participants. The present study probes *universality* and *culture-specific* accounts of emotion perception, comparing American[1] and German listeners' perception of emotional expressiveness conveyed by human and Amazon Alexa text-to-speech (TTS) voices. This is an extension of prior work [8], which found that American listeners perceive differences in dimensions of emotion (*valence, arousal*) across increasing 'happiness' levels (+0 %, +33 %, +66 %) resynthesized for the human and TTS voices. We hypothesize that cross-cultural differences (if present) might be better detected in these more gradient measures, rather than in the gross classification of the type of emotion (e.g., happy, sad, afraid). At the same time, it is also possible that cultural differences are partly due to differences in the interpretation and classification of emotion categories. This source of variation is eliminated in a dimensional approach to emotion perception, which in turn could reduce the expected cultural differences.

1.1. Cross-cultural Emotion Perception

In a meta-analysis of 37 studies, [9] summarize the state of the art in cross-cultural emotion recognition from voice. Most of these prior studies use speech samples, and some use non-verbal vocalizations, produced with a clearly intended emotion. Overall, recognition accuracy of emotion across studies was found to be above chance, which may serve as evidence for the *universality* accounts of emotion expression/perception. However, recognition accuracy was higher for within-culture conditions than in cross-culture conditions and even decreased with increasing cultural distance, demonstrating a cultural in-group advantage.

These findings support a hybrid *universality/culture-specific* account. For example, the *dialect theory* [10] assumes that emotional expression differs only moderately between cultures, so that cross-cultural recognition is generally possible, yet the differences are sufficiently strong that it is easier for members of the same culture (in-group) to recognize a given emotion than for members of a different culture (out-group). Belonging to the same culture is broadly understood hereafter as speaking the same language and having the same country of origin, as is common in cross-cultural emotion research.

While underexplored, a growing body of work aims to test whether cultures vary in perception of different dimensions of emotion, moving beyond explicit emotion classification. For example, [11] investigate the perception of *arousal* (calm vs. excited) and *valence* (positive vs. negative) in spontaneous Hebrew and German emotional speech by listeners of both cultural backgrounds, who are not proficient in the respective other language. Changes in arousal were perceived relatively consistently by both listener groups, especially in the German stimuli. However, the Hebrew listeners systematically rated the degree of arousal in these stimuli somewhat higher than the German listeners. Moreover, the valence ratings stand out in that they show greater differences between the two groups for both Hebrew and German stimuli: German listeners generally rated the stimuli as having a higher valence (i.e., more positive), while the Hebrew listeners perceived the same stimuli as having a lower valence (i.e., more negative). Furthermore, listeners were more consistent when rating stimuli of their own language. Note, however,

[1] In this paper, we use 'American' to refer to US Americans.

10.21437/Interspeech.2022-146

that the semantic content of the spontaneous speech in that study might not have been emotionally neutral, and thus their results could be partly based on content in the within-culture conditions.

The work of [12] explores the perception of *valence* and *arousal* in non-verbal affective vocalizations of Canadian actors by Canadian and Japanese listeners. They observe significant group/emotion interactions for both dimensions. For example, listener groups did not differ in their ratings of valence for happy/sad stimuli; however, they rated angry/pleased stimuli differently – e.g., angry vocalizations were rated less negatively by Japanese than by Canadian listeners. In contrast, arousal ratings differed only for sad stimuli, with Japanese listeners giving higher ratings than Canadian listeners. They conclude that for both positive and negative emotions, some are perceived similarly across cultures, while others are more culturally specific.

In [6], a group of American and German listeners was tested on pitch and rate manipulated samples of emotionally neutral sentences spoken in different emotional categories (happy, sad, angry, etc.). In addition to identifying the category, listeners rated how 'active' the speaker sounded (from 'passive' to 'active' on a 5-point scale). They found the largest differences for the two listener groups in response to the 'happy' manipulated stimuli, with steeper decreases for the German group.

1.2. Present Study

The present study also examines the emotional dimensions of *valence* and *arousal* cross-culturally. We compare American and German listeners, building on work showing differences in perception of emotional expressiveness in these two groups [13]. By focusing on within-category variation (all for 'happy' speech), we aim to test whether cross-cultural differences are detectable in arousal/valence ratings.

In the present study, following [8], we generated the emotionally manipulated tokens with the DAVID emotional resynthesis platform [14][2], which adapts pitch and spectral features to convey major emotions. For example, the 'happy' manipulation in the present study involves boosting higher frequencies through high-shelf filtering (resulting in a 'brighter' sound) and raising overall pitch.

Since DAVID is designed to work in real-time, transformations only operate on the level of speech cues that can be manipulated without taking the suprasegmental structure of an utterance into account – e.g., intensity, but not speech rate – which constitutes a simplification of emotional speech. However, results in [15] indicate that the modifications convey emotional meaning: DAVID was used to play back to participants their own emotionally modified speech as they were speaking. Subsequently, the participants' mood ratings changed in the direction in which the speech was modified, e.g., increased perceived positivity when hearing 'happy' speech feedback.

The features manipulated by DAVID were selected for being "frequently identified correlate[s] of emotional voices in the literature" [14, p. 326], without reference to language- or culture-specific expression of emotion. A validation study in [14] found that emotionally neutral sentences recorded in four different languages (French, English, Swedish, Japanese) and resynthesized with the DAVID parameters for various emotions (happy, sad, afraid) were accurately identified at a higher than chance level by a separate group of native-listeners for each respective language. However, there were performance differences between the groups – mainly driven by the Swedish listeners scoring lower overall. This demonstrates that DAVID can be applied in different language contexts to successfully convey emotion, while cultural differences in perception may still occur.

Results from [8] (a within-culture study) show that American listeners perceive changes in the happiness level of human speech gradiently, both in terms of valence and arousal. Yet, for TTS (specifically the Amazon Alexa voice), perceived increases due to the happiness manipulation were limited to the arousal dimension.

As mentioned, there are also cross-cultural differences in how emotion is perceived in non-human entities, such as robots [7, 16]. The present study extends that investigation, testing whether German and American listeners differ in how they perceive emotion conveyed by human and TTS voices. More broadly, there is increased interest in adding emotional expressiveness to make synthetic voices more appealing to human users of voice-activated devices [17, 18]. Yet, our understanding of how emotion in TTS voices might be perceived across cultures is still limited.

2. Methods

2.1. Stimuli

Stimuli were taken from [8]: 15 emotionally neutral English sentences recorded by a female human speaker (native US-English speaker) and the female US-English Amazon Alexa default voice. Both speakers produced the sentences in their regular prosody (i.e., not explicitly trying to produce a given emotion). We processed the sentences with the DAVID emotional resynthesis platform [14], increasing 'happiness' by 0 % (no change), 33 %, and 66 %. This resulted in a total of 90 stimuli (15 sentences × 3 happiness levels × 2 speakers).

2.2. Participants

A total of 111 native speakers of German completed the study (71 female, 35 male, 5 other; mean age 21.3 ±3.4 years, range 18 to 33 years), recruited through Prolific Academic. All participants spoke and understood English and had at least 6 years of experience with the language. German participants reported moderate prior usage of voice assistants (e.g., Alexa, Google Assistant, Siri, Cortana): 79 % have used such technology, while 39 % of these only infrequently. Data for the American participants (n=99; 70 female, 29 male; mean age 20.2 ±2.2 years, range 18 to 33 years) come from [8]. The majority (82 %) of the American participants reported prior voice assistant usage.

2.3. Procedure

The experiment was conducted online via Qualtrics. Before beginning the experimental trials, participants completed a rating task of the human and TTS voices. For both voices, they heard a representative recording, i.e., an emotionally neutral sentence not used in the experimental trials that was not manipulated in terms of 'happiness'. Participants were allowed to play the sound file as many times as they needed. The task was to rate the voices in terms of the following four dimensions, each on a sliding scale from 0 to 100: How **machine-like** (0) to **human-like** (100), how **artificial** (0) to **natural** (100), how **eerie** (0) to **comforting** (100), and how **cold** (0) to **warm** (100) do they sound? The slider position started at 50 ('neutral') for each rat-

[2] http://cream.ircam.fr

ing. These dimensions were adapted from [19] to assess participants' attitudes toward the voices they hear.

After that, the participants continued on to the experimental trials. For both listener groups, the design was identical: Participants heard all 90 stimuli (randomly presented within blocks grouped by voice, order of blocks counterbalanced across participants) and rated the *valence*, i.e., how **negative** (0) to **positive** (100) the speaker sounds, and the *arousal*, i.e., how **calm** (0) to **excited** (100) the speaker sounds. The slider position again started at 50 ('neutral') for each rating. On each trial, participants saw a matched-guise image of a human or the Amazon Echo silhouette to ensure that the speaker category was not ambiguous. For the American and German participants, instructions (and response options) were presented in English and German, respectively. In total, the study took roughly 25 minutes.

3. Analysis and Results

We centered all social (i.e., *human-like*, *natural*, *comforting*, *warm*) and emotional (i.e., *valence*, *arousal*) ratings based on the sliding scale from 0 to 100 by subtracting 50 (the 'neutral' value) from all values. For the *human-like* scale, for example, this means that values <0 indicate a more 'machine-like' rating, while values >0 indicate a more 'human-like' rating. For valence, values >0 indicate a degree of positive valence, while values <0 indicate a degree of negative valence. In the case of arousal, values >0 indicate a degree of excitement, while values <0 indicate a degree of calm.

3.1. Emotion Perception

We modeled participants' (centered) valence and arousal ratings in separate linear mixed effects models with the *lme4* R package [20]. Fixed effects included LISTENER GROUP (German, American), HAPPINESS LEVEL (+0 %, +33 %, +66 %), VOICE TYPE (Alexa, Human), and all possible interactions. Random effects included random intercepts for SENTENCE and LISTENER and by-listener random slopes for HAPPINESS LEVEL (due to a convergence issue only for the valence model) and VOICE TYPE. LISTENER GROUP and VOICE TYPE were sum coded, while HAPPINESS LEVEL was treatment coded (relative to +0 %). Participants' mean ratings for *valence* and *arousal* for the human and TTS voices are plotted in Figure 1, while the model outputs are provided in Table 1 for *valence* and Table 2 for *arousal*.

The valence model reveals that German listeners give higher valence ratings overall than American listeners and Alexa's voice receives higher valence ratings than the human voice. The increased happiness levels (+33 %, +66 %) generally receive higher valence ratings than the base level (+0 %). However, this effect is driven by the human voice in both listener groups, while Alexa's voice received a flat rating across happiness levels.

The arousal model shows that Alexa's voice receives lower arousal ratings than the human voice – in particular for the German listeners. The increased happiness levels (+33 %, +66 %) generally receive higher arousal ratings than the base level (+0 %) – this effect is stronger for American listeners than for German listeners and, in the case of American listeners, particularly affects the human voice.

3.2. Social Ratings

Participants' (centered) social ratings are plotted in Figure 2. Unpaired two-sample t-tests ($\alpha = 0.05$; p-values corrected for multiple comparisons) confirm differences in the ratings of Alexa's voice and the human voice for both listener groups across all four dimensions, with Alexa always scoring lower than the human voice (all $p < 0.001$). Differences between the ratings of the two listener groups for the same voice are only found in the case of Alexa's naturalness, with German listeners considering the TTS voice less *artificial* sounding ($t(207.15) = 3.05$, $p < 0.01$), and for the level of comfort of both the human and the TTS voice, which German listeners consider less *eerie* sounding (human: $t(197.19) = 3.29$, $p < 0.01$; TTS: $t(204.31) = 3.68, p < 0.01$).

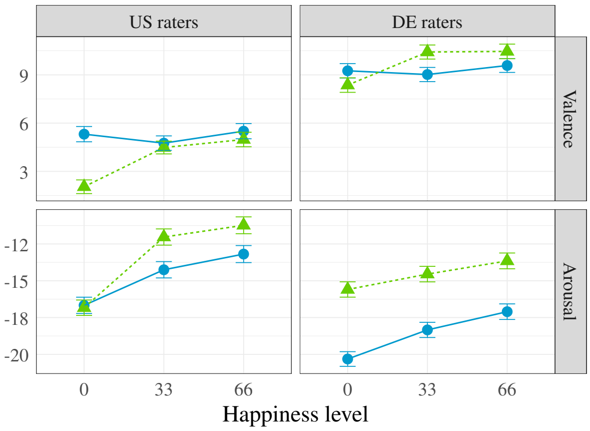

Figure 1: *Mean valence and arousal (scales: −50 to 50) for the three happiness levels as perceived in Alexa's voice (●) and the human voice (▲) by US raters and DE raters. The standard error is indicated.*

Table 1: *Perceived **valence** – parameter estimates (coefficients with standard error, t-statistic, and p-value) for the factors* LISTENER GROUP *(DE 1, US −1)* HAPPINESS LEVEL *(base level 0 % vs. 33 %, 66 %),* VOICE TYPE *(Alexa 1, Human −1), and their interactions (*).*

	Coef.	SE	t	p
(Intercept)	6.24	0.99	6.32	<0.001***
Group$_{DE}$	2.56	0.77	3.34	0.001**
Happiness$_{33}$	0.93	0.34	2.74	0.006**
Happiness$_{66}$	1.39	0.42	3.27	0.001**
Voice$_{Alexa}$	1.04	0.42	2.50	0.012*
G$_{DE}$*H$_{33}$	−0.01	0.34	−0.04	0.967
G$_{DE}$*H$_{66}$	−0.17	0.42	−0.41	0.679
G$_{DE}$*V$_{Alexa}$	−0.60	0.42	−1.43	0.152
H$_{33}$*V$_{Alexa}$	−1.32	0.23	−5.82	<0.001***
H$_{66}$*V$_{Alexa}$	−1.13	0.23	−4.98	<0.001***
G$_{DE}$*H$_{33}$*V$_{Al.}$	0.18	0.23	0.78	0.434
G$_{DE}$*H$_{66}$*V$_{Al.}$	0.25	0.23	1.09	0.278

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

4. Discussion

We compared American and German listeners' perception of gradient changes in emotional expressiveness in a human and a TTS voice. Our goal was to explore what is culturally-universal and what is culturally-specific in perceiving small differences in vocal expressiveness across different speaker types. We find

Table 2: *Perceived* ***arousal*** *– parameter estimates (coefficients with standard error, t-statistic, and p-value) for the factors* LISTENER GROUP *(DE 1, US −1)* HAPPINESS LEVEL *(base level* 0 % *vs.* 33 %, 66 %*),* VOICE TYPE *(Alexa 1, Human −1), and their interactions (*).*

	Coef.	SE	t	p
(Intercept)	−17.35	1.22	−14.22	$<0.001^{***}$
Group$_{DE}$	−0.60	0.98	−0.62	0.538
Happiness$_{33}$	2.56	0.26	9.66	$<0.001^{***}$
Happiness$_{66}$	3.56	0.26	13.44	$<0.001^{***}$
Voice$_{Alexa}$	−1.14	0.46	−2.47	0.013^{*}
G$_{DE}$*H$_{33}$	−1.05	0.26	−3.96	$<0.001^{***}$
G$_{DE}$*H$_{66}$	−0.98	0.26	−3.70	$<0.001^{***}$
G$_{DE}$*V$_{Alexa}$	−1.22	0.46	−2.65	0.008^{**}
H$_{33}$*V$_{Alexa}$	−0.36	0.26	−1.37	0.170
H$_{66}$*V$_{Alexa}$	−0.21	0.26	−0.80	0.425
G$_{DE}$*H$_{33}$*V$_{Al.}$	0.83	0.26	3.13	0.002^{**}
G$_{DE}$*H$_{66}$*V$_{Al.}$	0.85	0.26	3.20	0.001^{**}

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

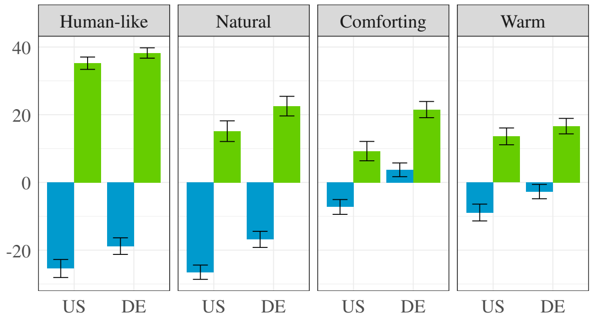

Figure 2: *Mean social ratings (scales:* −50 *to* 50*) as perceived in Alexa's voice (*🟦*) and the human voice (*🟩*) by US raters and DE raters. The standard error is indicated.*

evidence for both shared and distinct perceptions of gradient changes in emotion across German and American listeners.

In both groups, increasing the happiness expression conveyed in the voices (from +0% to +33% and +66%) increases both valence and arousal ratings, providing evidence that listeners from different cultural backgrounds both perceive these increases in emotionality, in line with *universality* accounts [3, 4].

Yet, examining different types of talkers – a human and a TTS voice – revealed some differences across the listener groups in perceived increases in 'happiness', supporting *culture-specific* accounts [5, 6]. In particular, the change in arousal from the initial (i.e., 'neutral') happiness level to both the +33% and +66% levels in the human voice is perceived more strongly by American than by German listeners. Since the voices were producing American English, this result is consistent with the *dialect theory* of emotion perception: in-group members show better emotion recognition than out-group members [10].

Furthermore, we see some more global differences in emotion perception across the groups. German listeners provide higher valence ratings overall (i.e., the productions sound more 'positive'). At the same time, German listeners rate the Alexa voice as sounding less excited (i.e., lower arousal rating) than American listeners. These differences appear to be cultural; yet, future work is needed to tease apart the contributions of language and type (human vs. TTS) as we used just one voice in each category and all stimuli were produced in American English.

Still, we observe some gradient perception of emotional expressiveness for human (valence and arousal) and TTS (arousal only) voices, which supports theories of computer personification [21]: People are applying their knowledge of emotion from human-human interaction to TTS. Future work examining other emotional categories can test the extent of TTS personification, and possible uncanniness response, as seen for emotionally expressive avatars [22].

The current study used female voices for both the human and TTS condition, based on the fact that Alexa still defaults to a female-sounding voice in both the American and German versions. The majority of the participants in the study were also female-identifying. The interplay of emotion decoder, encoder, and category is discussed in the literature [23]. Future work can address this aspect for the present scenario by including male voices and factoring in the influence of listener gender.

Social ratings for the (non-emotional) human and TTS voices show similar patterns across the groups: Both groups rated the human voice to be more human-like, natural, comforting, and warm, compared to the Alexa TTS voice. There were three small differences across groups, though: German listeners rated the TTS voice as sounding less artificial and less eerie than the US listeners, suggesting that German listeners might display a stronger anthropomorphism of the Alexa TTS voice. German listeners also rated the human voice as more comforting (i.e., less eerie). It is possible that social ratings might differ in a second language (here: L2 English). Future work can address this limitation by including productions by the same set of talkers in both languages, ideally by native bilingual speakers.

5. Conclusion

Overall, this study adds to the growing body of work examining cross-cultural variation in emotion and highlights the importance of comparing different *types* of interlocutors, such as human and TTS voices, when examining emotion perception. As TTS voices become even more commonplace and technological advances allow developers to more readily tune emotional expressiveness (e.g., Amazon: emotion [24]), examining human perception of this emotional expressiveness is critical for our scientific understanding of human-computer interaction, particularly in applications of cross-cultural communication.

6. Acknowledgements

We would like to thank Kristin Predeck and Eran Raveh for their assistance in earlier aspects of the project. This work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project-ID MO 597/6-2, and Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. The material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC. The work was partially supported by an Amazon Faculty Research Award to GZ.

7. References

[1] U. Scherer, H. Helfrich, and K. R. Scherer, "Paralinguistic behaviour: internal push or external pull?" in *Language*, 1980, pp. 279–282.

[2] K. R. Scherer, "Vocal affect signaling: a comparative approach," in *Advances in the Study of Behavior*, 1985, vol. 15, pp. 189–244.

[3] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, "Universal recognition of three basic emotions in music," *Current Biology*, vol. 19, no. 7, pp. 573–576, 2009.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[5] M. H. Bond, "Emotions and their expression in Chinese culture," *Journal of Nonverbal Behavior*, vol. 17, no. 4, pp. 245–262, 1993.

[6] C. Breitenstein, D. Van Lancker, and I. Daum, "The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample," *Cognition and Emotion*, vol. 15, no. 1, pp. 57–79, 2001.

[7] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Cultural differences in attitudes towards robots," in *AISB Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction*, University of Hertfordshire, Hatfield, UK, 2005, pp. 1–4.

[8] M. Cohn, E. Raveh, K. Predeck, I. Gessinger, B. Möbius, and G. Zellou, "Differences in gradient emotion perception: Human vs. Alexa voices," in *Interspeech*, Shanghai, China, 2020, pp. 1818–1822.

[9] P. Laukka and H. A. Elfenbein, "Cross-cultural emotion recognition and in-group advantage in vocal expression: a meta-analysis," *Emotion Review*, vol. 13, no. 1, pp. 3–11, 2021.

[10] H. A. Elfenbein and N. Ambady, "Universals and cultural differences in recognizing emotions," *Current Directions in Psychological Science*, vol. 12, no. 5, pp. 159–164, 2003.

[11] H. R. Pfitzinger, N. Amir, H. Mixdorff, and J. Bösel, "Cross-language perception of Hebrew and German authentic emotional speech," in *ICPhS*, 2011, pp. 1586–1589.

[12] M. Koeda, P. Belin, T. Hama, T. Masuda, M. Matsuura, and Y. Okubo, "Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners," *Frontiers in Psychology*, vol. 4, p. 105, 2013.

[13] S. Sommers and C. Kosmitzki, "Emotion and social context: An American-German comparison," *British Journal of Social Psychology*, vol. 27, no. 1, pp. 35–49, 1988.

[14] 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.

[15] J.-J. Aucouturier, P. Johansson, L. Hall, R. Segnini, L. Mercadié, and K. Watanabe, "Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 948–953, 2016.

[16] G. Trovato, T. Kishi, N. Endo, M. Zecca, K. Hashimoto, and A. Takanishi, "Cross-cultural perspectives on emotion expressive humanoid robotic head: recognition of facial expressions and symbols," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 515–527, 2013.

[17] 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.

[18] 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.

[19] 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.

[20] D. Bates, "Fitting linear mixed models in R," *R News*, vol. 5, no. 1, pp. 27–30, 2005.

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

[22] A. Tinwell, M. Grimshaw, D. A. Nabi, and A. Williams, "Facial expression of emotion and perception of the Uncanny Valley in virtual characters," *Computers in Human Behavior*, vol. 27, no. 2, pp. 741–749, 2011.

[23] A. Lausen and A. Schacht, "Gender differences in the recognition of vocal emotions," *Frontiers in Psychology*, vol. 9, p. 882, 2018.

[24] R. Singh and S. Baloni Ray, "It's not just what you say: studying user's affective response to Alexa voice interface stimuli," in *India HCI 2021*, 2021, pp. 99–104.