INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Cross-linguistic Emotion Perception in Human and TTS Voices

Iona Gessinger[1], *Michelle Cohn*[2], *Benjamin R. Cowan* [1], *Georgia Zellou*[2], *Bernd Möbius*[3]

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

{iona.gessinger|benjamin.cowan}@ucd.ie, {mdcohn|gzellou}@ucdavis.edu,
moebius@lst.uni-saarland.de

Abstract

This study investigates how German listeners perceive changes in the emotional expression of German and American English human voices and Amazon Alexa text-to-speech (TTS) voices, respectively. Participants rated sentences containing emotionally neutral lexico-semantic information that were resynthesized to vary in prosodic emotional expressiveness. Starting from an emotionally neutral production, three levels of increasing 'happiness' were created. Results show that 'happiness' manipulations lead to higher ratings of emotional valence (i.e., more positive) and arousal (i.e., more excited) for German and English voices, with stronger effects for the German voices. In particular, changes in valence were perceived more prominently in German TTS compared to English TTS. Additionally, both TTS voices were rated lower than the respective human voices on scales that reflect anthropomorphism (e.g., human-likeness). We discuss these findings in the context of cross-linguistic emotion accounts.

Index Terms: emotion perception, cross-linguistic, human-computer interaction, voice assistant, German, English

1. Introduction

Millions of individuals now use spoken interaction with voice technology (e.g., Amazon Alexa, Google Assistant, Apple Siri) to complete daily tasks [1]. For some, these interactions can be *cross-linguistic* and/or *cross-cultural*, such as native German speakers interacting with an American English text-to-speech (TTS) voice, e.g., while in the United States. To what extent the same mechanisms of emotion perception operate in these different types of interactions is an understudied question.

The present study tests whether utterances, manipulated with identical phonetic adjustments to approximate increasing 'happiness', are perceived differently based on (1) whether the talker is a real person or a TTS voice, and (2) whether the interaction is in one of the listener's first languages (L1; here German) or second languages (L2; here English).

1.1. Cross-cultural / Cross-linguistic Emotion Perception

In cross-cultural emotion research, it is common to define belonging to the same culture as speaking the same language and having the same country of origin. Both the encoding of emotion in speech and the decoding of emotion from speech depend on the cultural background of the speaker/listener [2, 3]. Regarding the perception side, different theoretical accounts exist. On the one hand, *culture-specific accounts* posit that listeners are more sensitive to emotional distinctions in their native language/culture [4, 5]. On the other hand, *universal accounts* predict that expressions of emotional prosody can be interpreted by all listeners [6]. Results from prior work indicate that universal and culture-specific effects occur jointly [7, 8].

For example, [7] had native English speakers assign an emotion label (anger, joy, fear, sadness) to lexico-semantically neutral sentences produced with a specific intended emotion by native speakers of English, German, Chinese, Japanese, and Tagalog in their respective language. Participants had minimal to no knowledge of the latter four languages. Although the tested languages belong to typologically unrelated families and/or are affected by different cultural contexts, native English speakers recognized the emotions above chance level in all languages. Furthermore, the accuracy was highest for classifying emotions in English, suggesting an in-group advantage for the native English listeners.

Work by [9] took a more gradient approach to emotion classification by having participants rate on a 6-point scale whether lexico-semantically neutral sentences conveyed particular emotions. Native Hebrew and native German speakers were able to correctly identify which emotional categories (anger, joy, fear, sadness) were conveyed in their own and the other language. Participants had no knowledge of the respective other language, i.e., they relied only on the emotional prosody.

In [10], emotional expression was assessed with a fully gradient approach. Native German and American English speakers rated American English stimuli on a 100-point sliding scale, assessing *valence* (positive vs. negative) and *arousal* (calm vs. excited) – without referring to a particular emotion category. The stimuli, produced by a human and a TTS voice, were manipulated to convey subtly increasing levels of 'happiness', which was reflected in overall increasing ratings of valence and arousal by both listener groups. While all participants were either L1 or L2 speakers of English, the fact that the stimuli were lexico-semantically neutral sentences implies that the ratings were solely based on non-verbal properties.

Taken together, these studies demonstrate that culture-specific differences in the perception of emotional prosody are captured by both coarse classification and gradient assessment approaches. The cross-linguistic aspect is usually a by-product of the cross-cultural comparison and cannot be easily distinguished from the latter.

1.2. Linguistic Background and Emotion

A large body of work has shown that bi- and multilingual individuals experience emotions differently across languages [11, 12]. In [13], for example, Spanish-English bilinguals rated written words in terms of their emotionality on a 7-point scale from *unemotional* to *emotional*. While they did not observe differences in the rating itself, participants tended to remember more emotional words in their L1 compared to their L2 in a

10.21437/Interspeech.2023-711

post-rating word recall task. Similarly, [14] found that Turkish-English bilinguals showed greater skin conductance in response to heard/read taboo words in Turkish (their L1) than for taboo words in English (learned after age 12). Based on these asymmetries, some have proposed that a speaker's L2 might entail a greater *emotional distance* compared to an L1 [15]. While these studies mainly examined lexico-semantically conveyed emotion, the same may be the case when emotion is conveyed via prosody only. In the present study, this could result in greater sensitivity to the emotional expressiveness in the L1 stimuli.

1.3. Emotion in Human-Computer Interaction

Voice technology systems are increasingly imbued with human-like qualities, including emotional expressiveness, to become more engaging conversational partners for human users [16, 17, 18]. Some work has shown that people respond similarly to exaggerated displays of emotional expressiveness in human and TTS voices. For example, speakers mirror emotionally expressive prosody produced by a conversational partner, be it a TTS or human voice [19]. More subtle changes in emotionality have also been found to be perceived in these different talker types. The gradient evaluation of increasing 'happiness' in [10] (see above) revealed that listeners perceived changes in arousal in both human and TTS voices, while changes in perceived valence were found for the human talker only.

Prior work has also shown that there is cross-cultural variation in the degree of acceptance of non-human interlocutors displaying emotion. For example, [20] used the *Negative Attitudes towards Robots Scale* [21] with participants from various countries and found that American respondents had the most positive attitude towards robots showing emotions, while German and Dutch participants held more negative attitudes.

1.4. Present Study

The present study consists of a novel experiment, conducted fully in German by L1 German listeners, as well as a comparison with data from [10], in which German-English bilingual listeners assessed American English utterances (their L2). The stimuli in both experiments were manipulated using the DAVID Emotional Resynthesis platform [22] to convey three levels of subtly increasing 'happiness'. While [10] took a *cross-cultural* approach, comparing the data of the German listeners to those of American listeners rating the same American English utterances, the present study takes a *cross-linguistic* approach.

On the production side, the cross-linguistic aspect is isolated to the extent possible from the speaker's cultural background in this study, since both the German and English emotionally manipulated stimuli were produced in the same way. On the perception side, all listeners are native speakers of German, hence share the same cultural background.

We expect listeners to perceive the increase in 'happiness' in both the human and the TTS voices as has been the case in [10] – reflected at least in the arousal dimension. It remains to be seen whether the lack of an effect in the *valence* dimension for the TTS voice in [10] was due to its non-human nature – and will be the same for the German TTS voice – or whether it was rather due to the specific TTS voice used, namely the default American English Alexa voice.

If the in-group advantage discussed above is mainly due to the culture-specific emotional expressiveness on the production side, we may observe a comparatively smaller or even no difference between the L1 and L2 stimuli in the present study, since they were produced using the same parameters in DAVID. However, it is also possible that increased sensitivity to emotional nuances in the L1 still influences the listeners in the present case and leads to stronger effects for the German stimuli.

2. Material and Methods

2.1. Participants

In total, 89 native German speakers completed the study (46 female, 43 male; mean age 20.6 $\pm$1.2 years, range 18 to 22 years). The participants were recruited through Prolific Academic and reported having moderate prior experience using voice assistants (VAs). While 81 % reported having used such technology before, 44 % of these only infrequently (i.e., "seldom" or "once a month"). This is similar to the distribution in the reference group of native German speakers from [10] (n=111; 71 female, 35 male, 5 other; mean age 21.3 $\pm$3.4 years, range 18 to 33 years), where 79 % have previously used a VA, 39 % of them only infrequently. Roughly equal amounts of participants in both groups have used either Amazon Alexa and other VAs (52.5 %), only Alexa (3.5 %), or only other VAs (44 %).

After the experiment, participants rated their attitude towards Alexa and VAs in general on 5-point scales (1 *very negative* to 5 *very positive*). Results demonstrate that they have a slightly more negative attitude towards Alexa (mean=2.96 $\pm$1) than towards VAs in general (mean=3.24 $\pm$1).[1] Attitudinal data was not available in [10].

2.2. German Stimuli

We selected 15 emotionally neutral sentences from the literature [23, 24, 25] that were either German or translated to German – for example, "Ich sehe einen Teppich auf dem Boden." (I see a carpet on the floor.)[2]

The sentences were recorded by a female German native speaker (aged 33 years) and produced by the default German female Amazon Alexa TTS voice. Both speakers produced the sentences with neutral prosody (i.e., without explicitly expressing a particular emotion). The sentences were then manipulated with the DAVID emotional resynthesis platform [22]. We used the default parameters for increasing 'happiness': an upwards pitch shift of 30 cents, inflection, i.e., rapid changes in pitch over periods of 500 ms, and high-shelf filtering with a cut-off frequency of 8 kHz and 3 dB gain per octave. We applied these parameters at the 0 %, 33 %, and 66 % levels. This resulted in a total of 90 stimuli (15 sentences $\times$ 3 happiness levels $\times$ 2 speakers).[3] The English stimuli in [10] were prepared using the same parameters.

2.3. Procedure

The study was approved by the *Human Research Ethics Committee* at *University College Dublin*. It was conducted online via Qualtrics and took ca. 20 min. First, participants rated one emotionally neutral sentence each of the human and the TTS voice in terms of human-likeness, naturalness, comfort, and warmth (referred to as *social ratings* in the following). These stimuli

[1] See Appendix Text A for qualitative data on participants' attitude towards voice assistants:
https://doi.org/10.17605/OSF.IO/FPV3K

[2] We validated the sentences with a state-of-the-art BERT-based sentiment analysis model trained for contemporary German [26], which confirmed that they are overwhelmingly neutral.

[3] See Appendix Figure A (link above) for pitch contours and long-term average spectra of an example manipulation with DAVID that demonstrate the subtlety of the applied changes.

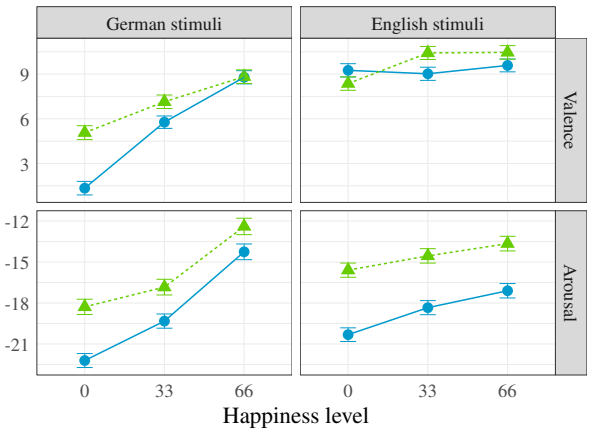

Figure 1: *Mean valence and arousal (scales: −50 to 50) for the three happiness levels as perceived by German listeners in the German and English human voices (▲) and Alexa voices (●), respectively. The standard error is indicated.*

were not used in the subsequent experimental trials and were not manipulated with regard to 'happiness'. The rating was carried out on a sliding scale from 0 to 100 (*machine-like* to *human-like*, *artificial* to *natural*, *eerie* to *comforting*, *cold* to *warm*) with the slider position starting at a neutral position of 50 for each rating. The four dimensions were used to assess participants' attitudes toward the voices they heard (adapted from [27]).

Then, participants proceeded to the experimental trials, where they heard all 90 stimuli (presented randomly within blocks grouped by voice; order of blocks counterbalanced between participants) and rated the *valence* (i.e., how *negative* to *positive* the speaker sounds), and the *arousal* (i.e., how *calm* to *excited* the speaker sounds). Again, the rating was carried out on a sliding scale from 0 to 100 with the slider position starting at 50. With each trial, participants saw a silhouette of either a human or an Amazon Echo, matching the voice type, to ensure that the talker category was clear to them.

3. Analysis and Results

All social ratings (i.e. *human-like*, *natural*, *comfortable*, *warm*) and emotion perception ratings (i.e. *valence*, *arousal*) on the scale from 0 to 100 were centered around zero. Thus, values below 0 indicate a more *machine-like*, *artificial*, *eerie*, or *cold* social rating, a level of positive *valence*, or a level of excitement in the case of *arousal*. Values above 0, in contrast, indicate a more *human-like*, *natural*, *comforting*, or *warm* social rating, a level of negative *valence*, or a level of calmness for *arousal*.

3.1. Emotion Perception Ratings

Figure 1 shows the mean (centered) ratings of *valence* and *arousal* for the German and English human voices and Alexa voices, respectively. The ratings were modelled in separate linear mixed-effects models for *valence* and *arousal* using the *lme4* package [28] in R [29]. The fixed effects included STIMULUS LANGUAGE (English, German), HAPPINESS LEVEL (0 %, 33 %, 66 %), VOICE TYPE (Human, Alexa), and all possible interactions. Random effects included random intercepts for SENTENCE and LISTENER and by-listener random slopes for VOICE TYPE. Due to convergence issues, by-listener random slopes for HAPPINESS LEVEL could not be included. STIMULUS LANGUAGE and VOICE TYPE were sum coded, while HAPPINESS LEVEL was treatment coded (relative to 0 %). The model outputs are provided in Table 1.

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

Valence	Coef.	SE	t	p
(Intercept)	6.00	0.92	6.53	<0.001***
Stimulus_G	−2.81	0.92	−3.06	0.002**
Happiness_{33}	2.08	0.23	9.09	<0.001***
Happiness_{66}	3.41	0.23	14.87	<0.001***
Voice_{Alexa}	−0.71	0.42	−1.67	0.096
S_G*H_{33}	1.17	0.23	5.11	<0.001***
S_G*H_{66}	2.20	0.23	9.59	<0.001***
S_G*V_{Alexa}	−1.15	0.42	−2.71	0.007**
H_{33}*V_{Alexa}	0.02	0.23	0.07	0.941
H_{33}*V_{Alexa}	0.48	0.23	2.08	0.038*
S_G*H_{33}*V_{Alexa}	1.16	0.23	5.06	<0.001***
S_G*H_{66}*V_{Alexa}	1.36	0.23	5.93	<0.001***
Arousal	Coef.	SE	t	p
(Intercept)	−19.11	1.12	−17.01	<0.001***
Stimulus_G	−1.15	1.12	1.02	0.305
Happiness_{33}	1.85	0.27	6.91	<0.001***
Happiness_{66}	4.76	0.27	17.83	<0.001***
Voice_{Alexa}	−2.17	0.42	−5.15	<0.001***
S_G*H_{33}	0.33	0.27	−1.25	0.210
S_G*H_{66}	2.19	0.27	−8.18	<0.001***
S_G*V_A	0.19	0.42	−0.44	0.658
H_{33}*V_{Alexa}	0.60	0.27	2.26	0.024*
H_{66}*V_{Alexa}	0.85	0.27	3.19	0.001**
S_G*H_{33}*V_{Alexa}	0.14	0.27	−0.51	0.608
S_G*H_{66}*V_{Alexa}	0.22	0.27	−0.81	0.415

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

The valence model indicates that German listeners give higher ratings to the English stimuli than to the German stimuli overall. The increased happiness levels (33 %, 66 %) generally receive higher valence ratings than the base level (0 %), with the increase being more pronounced for the German stimuli, especially the German Alexa voice.

The arousal model shows no difference between ratings of the German and English stimuli overall, but reveals that Alexa's voices generally receive lower arousal values. The increased happiness levels (33 %, 66 %) receive higher arousal ratings than the base level (0 %). The increase in perceived arousal is more pronounced for the Alexa voices and, regarding the third manipulation level (66 %), perceived more strongly in the German stimuli for both the human and the Alexa voice.

3.2. Social Ratings

Figure 2 shows the (centered) social ratings. In all four dimensions, the German and English Alexa voices received lower scores than the respective human voices, as assessed with unpaired two-sample t-tests ($\alpha = 0.05$; p-values corrected for multiple comparisons; all $p < 0.001$).

The ratings of the German vs. English version of a voice differed significantly only in the case of the human-likeness of the human voices, where the German voice was rated somewhat more machine-like ($t(168.1) = -3.45$, $p < 0.01$).

4. Discussion

This study investigated how German listeners perceive subtle, gradient changes in the emotional expressiveness of German and English human and TTS voices produced by resynthesis. The aim was to investigate whether perception depends on the talker type or the language, i.e., the listeners' L1/L2. We found both shared and distinct patterns across talkers and languages.

Both non-human talkers (German and English) were rated lower on scales that reflect anthropomorphism (i.e., human-likeness, naturalness, comfort, warmth) and the ratings were overall very similar between talkers of the same type – with the exception that the German human voice was rated less human-like than the English human voice. However, these differences were not reflected in the subsequent emotion ratings. Note that the identity of the system had not been explicitly stated at this point in the experiment, i.e., participants would have only known that it was an artificial voice if they had recognized Alexa. Therefore, the differences in social ratings are only based on what was heard. The talker type was then explicitly stated during the emotional rating task so that the participants knew whether they were hearing a human or a TTS voice when rating valence and arousal dimensions. Consequently, participants' attitudes towards VAs may have influenced these ratings.

For both their L1 and L2, listeners did perceive the changes resulting from the increase of 'happiness' by 0 %, 33 %, and 66 %. This was reflected in an increase of ratings for how excited (arousal) and how positive (valence) the stimuli sounded to them overall, consistent with *universal accounts* of emotion perception [6]. At the same time, the observed effects were in fact more pronounced for the German stimuli (i.e., their L1), supporting *culture-specific accounts*, such as those that predict greater sensitivity to emotion in one's first language(s) than for languages learned later in life [11, 15, 12].

Most strikingly, the third 'happiness' manipulation step (i.e., 66 %) was perceived as significantly more excited in both German voices (human and TTS) and the German Alexa voice elicited a considerable increase in perceived positivity – in contrast to the American English Alexa voice for which valence did not increase across manipulation levels.

The overall stronger effects for the L1 stimuli are somewhat surprising, since identical manipulations were used to increase 'happiness' with the DAVID Emotional Resynthesis platform for all stimuli included in this comparison. This should eliminate the cultural influence on emotion expression (i.e., on the production side) in the stimuli. However, we still compared four different voices (2 human and 2 TTS). Hence, emotion expression was not decoupled from the talker who provided a certain 'neutral' baseline for the manipulations in each case. It is possible that the 'neutral' baseline of American English Alexa differs from that of German Alexa in such a way that the specific manipulations carried out with DAVID would be less perceptible in the former. In other words, the same acoustic changes may contribute to emotion perception differently depending on the baseline they are applied to. We thus suggest that the lack of an effect on the *valence* dimension for the TTS voice in [10] was not due to its non-human nature, but rather to the specific voice used in the experiment, the default American English Alexa voice.

In contrast, the prominent increase in perceived arousal from the second to the third manipulation step occurs similarly for both German voices and is not observed for the two English voices. Since the underlying acoustic changes are again the same in both cases, the language per se seems to play a role here. Perhaps the language brings back the cultural aspect through expectations that listeners have of German speakers (from Germany) and English speakers (from the USA). If this is the case, our results would suggest that the same acoustic changes indicate a greater change in arousal when produced by German speakers compared to American English speakers. Further research is needed to verify if this is indeed the case.

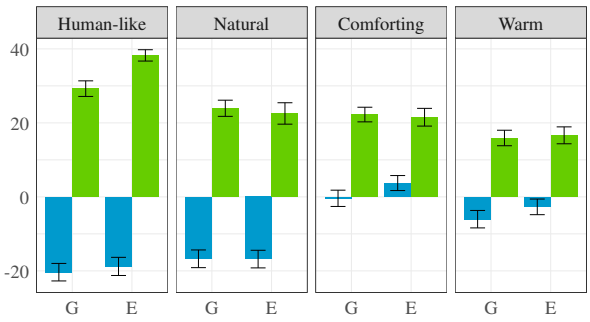

Figure 2: *Mean social ratings (scales: −50 to 50) as perceived by German listeners in the German (G) and English (E) human voices (■) and Alexa voices (■), respectively. The standard error is indicated.*

The present study has several limitations that can serve as avenues for future research. First, the experiment used a between-participants design. While our modeling accounts for inter-listener differences, it is possible that we would see larger differences in a within-participants design, such as with German-English bilinguals providing ratings for both languages. Additionally, the present study uses languages participants are proficient in. Future studies can test languages that are completely unfamiliar to listeners, which would require them to rely solely on the acoustic-prosodic features of the stimuli.

Furthermore, participants may have different attitudes towards specific voice assistants. For example, we found that participants in the present study had a slightly more negative attitude towards Alexa than towards other VAs. This may influence the assessment of emotional expressiveness, among other aspects. Given the increasing popularity of speech technology systems and the associated evolution of user opinion, it becomes crucial to distinguish between findings about VAs in general and specific systems in particular.

Finally, the present study is based on comparing four different talker voices, each providing different 'neutral' baselines. Future work holding the voice constant (e.g., using a bilingual speaker or a multilingual TTS voice) can shed further light on speaker-specific and language/culture-specific effects.

5. Conclusion

The present study contributes to our understanding of how humans perceive emotional expressiveness in non-human speakers. We show that artificially produced emotionality may be perceived differently in speakers of the same type (e.g., German vs. English TTS), as well as similarly in speakers of different categories (e.g., human vs. non-human).

6. Acknowledgements

We would like to thank Yuri Bizzoni for the sentiment analysis. This work was funded by Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at UCD and an Amazon Research Grant to GZ.

7. References

[1] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, search, and IoT: How people (really) use voice assistants," *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 3, pp. 17–1, 2019.

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

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

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

[5] 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.

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

[7] W. F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 158, no. 1/4, pp. 407–424, 2006.

[8] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing emotions in a foreign language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009.

[9] V. Shakuf, B. Ben-David, T. G. Wegner, P. B. Wesseling, M. Mentzel, S. Defren, S. E. Allen, and T. Lachmann, "Processing emotional prosody in a foreign language: the case of German and Hebrew," *Journal of Cultural Cognitive Science*, vol. 6, no. 3, pp. 251–268, 2022.

[10] I. Gessinger, M. Cohn, G. Zellou, and B. Möbius, "Cross-cultural comparison of gradient emotion perception: Human vs. Alexa TTS voices," in *Interspeech*, Incheon, Korea, 2022, pp. 4970–4974.

[11] A. Pavlenko, "Affective processing in bilingual speakers: Disembodied cognition?" *International Journal of Psychology*, vol. 47, no. 6, pp. 405–428, 2012.

[12] Y. Yao, K. Connell, and S. Politzer-Ahles, "Hearing emotion in two languages: A pupillometry study of Cantonese-Mandarin bilinguals' perception of affective cognates in L1 and L2," *Bilingualism: Language and Cognition*, pp. 1–14, 2023.

[13] L. J. Anooshian and P. T. Hertel, "Emotionality in free recall: Language specificity in bilingual memory," *Cognition & Emotion*, vol. 8, no. 6, pp. 503–514, 1994.

[14] C. L. Harris, A. Ayçiçeği, and J. B. Gleason, "Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language," *Applied Psycholinguistics*, vol. 24, no. 4, pp. 561–579, 2003.

[15] B. Keysar, S. L. Hayakawa, and S. G. An, "The foreign-language effect: Thinking in a foreign tongue reduces decision biases," *Psychological science*, vol. 23, no. 6, pp. 661–668, 2012.

[16] 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.

[17] A. R. F. Rebordao, M. A. M. Shaikh, K. Hirose, and N. Minematsu, "How to improve TTS systems for emotional expressivity," in *Interspeech*, Brighton, UK, 2009, pp. 524–527.

[18] M. P. Aylett, L. Clark, B. R. Cowan, and I. Torre, "Building and designing expressive speech synthesis," in *The Handbook on Socially Interactive Agents: Volume 1*. New York, NY, USA: Association for Computing Machinery, 2021, p. 173–212.

[19] M. Cohn, K. Predeck, M. Sarian, and G. Zellou, "Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers," *Speech Communication*, vol. 135, pp. 66–75, 2021.

[20] 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.

[21] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, "Measurement of negative attitudes toward robots," *Interaction Studies*, vol. 7, no. 3, pp. 437–454, 2006.

[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] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[24] B. M. Ben-David, P. H. H. M. 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.

[25] S. Defren, P. Wesseling, S. Allen, V. Shakuf, B. M. Ben-David, and T. Lachmann, "Emotional Speech Perception: A set of semantically validated German neutral and emotionally affective sentences," in *Speech Prosody*, Poznań, Poland, 2018, pp. 714–718.

[26] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Training a broad-coverage German sentiment classification model for dialog systems," in *Language Resources and Evaluation Conference (LREC)*. Marseille, France: European Language Resources Association, 2020, pp. 1627–1632. [Online]. Available: https://aclanthology.org/2020.lrec-1.202

[27] 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.

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

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/