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

Iona Gessinger1, 2, Michelle Cohn3, Georgia Zellou3, Bernd Möbius1

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

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 American1 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,
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 condi-
tions.

The work of [12] explores the perception of valence and
arousal in non-verbal affective vocalizations of Canadian ac-
tors by Canadian and Japanese listeners. They observe signif-
icant group/emotion interactions for both dimensions. For ex-
ample, 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 rat-
ings 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 simi-
larly 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 ‘ac-
tive’ 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 emo-
tionally manipulated tokens with the DAVID emotional resyn-
thesis platform [14]2, which adapts pitch and spectral features
to convey major emotions. For example, the ‘happy’ manipu-
lation 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, transforma-
tions only operate on the level of speech cues that can be manip-
ulated without taking the suprasegmental structure of an utter-
ance into account – e.g., intensity, but not speech rate – which
constitutes a simplification of emotional speech. However, re-
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. Sub-
sequently, the participants’ mood ratings changed in the direc-
tion in which the speech was modified, e.g., increased perceived
positivity when hearing ‘happy’ speech feedback.

The features manipulated by DAVID were selected for be-
ing “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 differ-
ences between the groups – mainly driven by the Swedish lis-
teners 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 oc-
cur.

Results from [8] (a within-culture study) show that Ameri-
can 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 in-
creases 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 expres-
siveness to make synthetic voices more appealing to human
users of voice-activated devices [17, 18]. Yet, our understand-
ing 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 emo-
tional resynthesis platform [14], increasing ‘happiness’ by 0 %
(no change), 33 %, and 66 %. This resulted in a total of 90 stim-
uli (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 par-
ticipants 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 us-
age.

2.3. Procedure

The experiment was conducted online via Qualtrics. Before be-
inning 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-

---

2http://cream.ircam.fr
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 (α = 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).

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 (*).

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−17.35</td>
<td>1.22</td>
<td>−14.22</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Group&lt;sub&gt;DE&lt;/sub&gt;</td>
<td>−6.00</td>
<td>0.98</td>
<td>−6.2</td>
<td>0.538</td>
</tr>
<tr>
<td>Happiness&lt;sub&gt;33&lt;/sub&gt;</td>
<td>2.56</td>
<td>0.26</td>
<td>9.66</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Happiness&lt;sub&gt;66&lt;/sub&gt;</td>
<td>3.56</td>
<td>0.26</td>
<td>13.44</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>Voice&lt;sub&gt;Alexa&lt;/sub&gt;</td>
<td>−1.14</td>
<td>0.46</td>
<td>−2.47</td>
<td>0.013*</td>
</tr>
<tr>
<td>G&lt;sub&gt;DE&lt;/sub&gt;·H&lt;sub&gt;33&lt;/sub&gt;</td>
<td>−1.05</td>
<td>0.26</td>
<td>−3.96</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>G&lt;sub&gt;DE&lt;/sub&gt;·H&lt;sub&gt;66&lt;/sub&gt;</td>
<td>−0.98</td>
<td>0.26</td>
<td>−3.70</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>G&lt;sub&gt;DE&lt;/sub&gt;·V&lt;sub&gt;Alexa&lt;/sub&gt;</td>
<td>−1.22</td>
<td>0.46</td>
<td>−2.65</td>
<td>0.008**</td>
</tr>
<tr>
<td>H&lt;sub&gt;33&lt;/sub&gt;·V&lt;sub&gt;Alexa&lt;/sub&gt;</td>
<td>−0.36</td>
<td>0.26</td>
<td>−1.37</td>
<td>0.170</td>
</tr>
<tr>
<td>H&lt;sub&gt;66&lt;/sub&gt;·V&lt;sub&gt;Alexa&lt;/sub&gt;</td>
<td>−0.21</td>
<td>0.26</td>
<td>−0.80</td>
<td>0.425</td>
</tr>
<tr>
<td>G&lt;sub&gt;DE&lt;/sub&gt;·H&lt;sub&gt;33&lt;/sub&gt;·V&lt;sub&gt;Al.&lt;/sub&gt;</td>
<td>0.83</td>
<td>0.26</td>
<td>3.13</td>
<td>0.002**</td>
</tr>
<tr>
<td>G&lt;sub&gt;DE&lt;/sub&gt;·H&lt;sub&gt;66&lt;/sub&gt;·V&lt;sub&gt;Al.&lt;/sub&gt;</td>
<td>0.85</td>
<td>0.26</td>
<td>3.20</td>
<td>0.001**</td>
</tr>
</tbody>
</table>

* *p < 0.001, **p < 0.01, *p < 0.05

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.

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 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 5976/2, 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


