Analysis-by-Synthesis Assessment of Speech Emotion Perception in Different Languages

Jueun Kang[1], Paolo Sani[2]

[1] Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, E-Mail: jueun.kang@iis-extern.fraunhofer.de
[2] Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, E-Mail: paolo.sani@iis.fraunhofer.de

1. Introduction

1.1 Emotion perception in different languages

Emotion conveyed in speech is one of many important means of communication between interlocutors. As the manner of emotional speech is known to be shaped by contextual factors including culture [1], more studies have been devoted to identifying the correlation between the pattern of emotion perception and languages which is closely embodied in culture.

So far, controversial findings have been claimed by researchers, mostly under the theories of "universalism" and "relativism". Based on evolution and basic emotion theories suggested by Charles Darwin [2] and Paul Ekman [3] respectively, universalist researchers claim that basic emotions including anger, happiness, sadness, or disgust are similarly exhibited across different languages or cultural regions, emphasizing that the emotional production is accompanied with specific muscle movements on the face and in articulatory organs [4][5][6][7]. Ekman and Friesen [8], for example, conducted an experiment for facial expression with contempt emotion. In this test, they presented pictures with facial expression of emotion and judges from 10 countries with different cultures such as Germany, United States, and Japan were asked to categorize them. The result indicated that the target emotion was not perceived distinctively by participants with different nationalities.

On the other hand, relativist scholars state that the manner of emotional expression through gestures, speech, or facial expressions is, in fact, highly conditioned by a cultural context a speaker belongs to. According to this theory, therefore, emotions are signaled and perceived distinctively by people from different cultural regions [9][10][11][12]. For example, Lim [13] observes that in East Asian cultures, happiness is more often associated with contentment, as opposed to Western cultures, in which it is more likely to be associated with feelings of excitement. In line with this, Affect Valuation Theory (AVT), proposed by Tsai [14], underlines that people from different cultural regions experience and evaluate the same emotion distinctively, as each culture has ideal affects which members of the community consider desirable and would like to pursue. For instance, participants with a Western culture background indicated preference in high arousal positive state such as physical activities, while those from Eastern culture valued low arousal positive state including calm and relaxed states.

1.2 Present study

While emotion used to be a research area which was studied predominantly by psychologists, researchers in the field of Text-to-Speech (TTS) synthesis have also drawn their attention to this topic, as they aim to produce expressive speech in different languages. Although the recent advance in neural TTS technology can generate very natural and intelligible speech, the question on how listeners perceive emotions in synthesized voice remains open.

In this paper, we present a novel work, to the best of our knowledge, on the perception of artificial emotional speech by including cultural and linguistic factors into the analysis framework. Following the previous work by Gessinger et al. [15], we base our analysis on a similar multi-language scenario, where Korean is added to English and German to provide more linguistic diversity. Furthermore, we expand the emotional states investigated to four, i.e. angry, happy, neutral and sad.

Our research aims are as follows: Firstly, we evaluate our TTS model regarding its expressiveness in different emotions and languages. Moreover, we aim to identify any universal or distinctive pattern of emotion perception across different languages.

To evaluate the synthesized utterances, we conduct a listening test with expert listeners, most of which are native or near-native speakers in either German or English but have little or no proficiency in Korean. This condition allows us to investigate the influence of the linguistic background on the emotion perception. The listening test consists of four experiments, namely naturalness, categorization, valence (how positive or negative) and arousal (how excited or calm) test. The results show that listeners tend to perceive emotions similarly in each language regardless of their cultural and linguistic backgrounds.

2. Data and Methods

2.1 Participants

Each test had a different number of participants: 20 listeners (13 males, 7 females) in naturalness, 16 (10 males, 6 females) in categorization, 20 (11 males, 9 females) in valence, and 15 (10 males, 5 females) in the arousal test. Overall, the subjects ranged in age from 20 to 44 and come from a variety of different countries including Colombia, Germany, India, Iran, Italy, Jordan, Korea, and the UK, with half of them being German.

Given the aim of the study, i.e., to identify crosslinguistic emotion perception, participants were asked at the end of each test to indicate their nationality and language proficiency in German, English and Korean. While the proficiency level in German varied from native to elementary and high-intermediate or native level in English, almost all participants but one reported no proficiency in Korean.

2.2 Emotion dataset and model training

To train the TTS pipeline, we selected three emotional speech corpora that included the four emotional states under investigation, namely angry, happy, neutral and sad. For German, PAVOQUE [16] and three proprietary datasets were used. English speech data was obtained from ESD database[1] [17], and Korean emotional speech data from AI hub[2]. Details for each dataset is shown in Table 1. In each dataset, emotions are labeled and aforementioned emotions were selected, as they are the shared emotions across difference datasets.

Our TTS pipeline [18] comprises: Forward Tacotron[3] as an acoustic model and StyleMelGAN [19] as neural vocoder. Three acoustic models were individually trained according to the number of target languages, keeping the same training settings. To inform the acoustic models on the target emotions, we conditioned them with unique IDs representing each unique speaker-emotion couple during training and inference. The mel-scaled spectrograms generated by the three acoustic models were then converted into time-domain speech signals by the same pre-trained neural vocoder.

Table 1: Training data hours for each language: the number of speakers is specified in the parentheses.

	German	English	Korean
Male	11.04h (2)	4.99h (5)	131.91h (21)
Female	24.07h (2)	4.74h (5)	133.82h (20)
Sum	35.11h	9.73h	265.73h

2.3 Experiment procedure

The listening test was conducted utilizing open-source WebMUSHRA [20]. Listeners were provided with examples at the beginning of categorization, valence, and arousal tests to obtain a general idea on how audio samples should be evaluated. The introductory guidelines also informed the participants that the audio samples were to be evaluated without considering the semantic content of each utterance. In each test, 12 speech samples (four emotions * three stimuli) were provided with durations ranging from two to six seconds. The order was German, English, and Korean.

In naturalness evaluation, two samples with the same stimulus and emotion were given on each page, being respectively the original audio sample and the speech generated by the TTS pipeline. The samples were provided without labels and the order was randomized between pages. This was done to ensure that listeners could not deduce which was the original recording. Listeners were asked to provide a rating between 0 and 100. For the categorization test, each page contained one audio sample and listeners had to select one out of four possible emotions: angry, happy, neutral, and sad. Finally, in the valence and arousal test, participants had to evaluate one sample per page on a scale from "Very negative"/ "very calm" to "very positive"/ "very excited", respectively. It should be noted that while human voice was provided as reference in the naturalness test, other tests included only synthesized speech.

3. Results

Naturalness

Figure 1 shows the score of naturalness between samples of human and synthesized voice, of which human speech always obtained higher scores across three languages. Moreover, between the synthesized voice, German TTS voice was rated as the most natural voice, which is followed by Korean and English. Given that the participants evaluated Korean speech more natural than English but still less than German, which are familiar languages to them, this result confirmed the fact that the amount of training data play a critical role in naturalness perception even when listeners may not understand the speech.

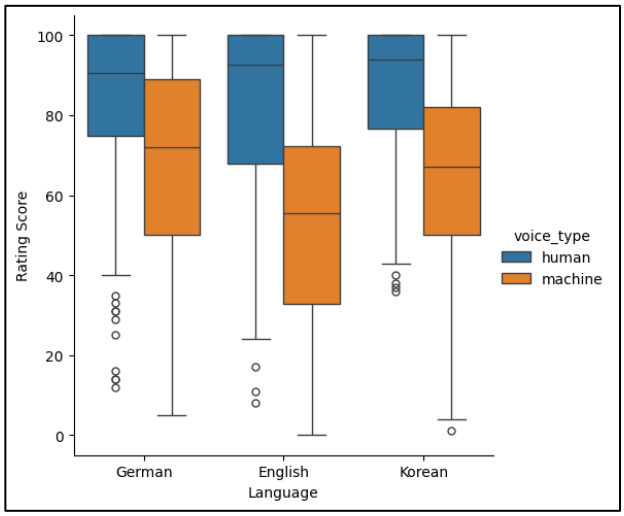

Figure 1: Comparison of naturalness scores between human and synthesized voice: 'machine' indicates synthesized voice

Categorization

The categorization test result is illustrated as confusion matrix in Figure 2. Despite the fact that almost no participants indicated any knowledge in Korean, their emotion categorization performance in the language shows, to our surprise, above chance level (0.25) in all emotions. In addition, sadness obtained the highest accuracy rate across three languages. This could be attributed to distinctive features of the emotion which accompanies low pitch and slow speech rate.

Some misclassifications were, however, found especially in angry and happy speech. German angry speech was often classified neutral voice while Korean angry speech is perceived as either neutral or happy voice. As of English, neutral, happy and angry speech was interpreted as neutral speech except for the sad speech although most of participants indicated high proficiency in English. Considering that English TTS voice received the lowest naturalness score, it is possible that the low natural speech could have misled the emotion assessment.

[1] https://github.com/HLTSingapore/Emotional-Speech-Data
[2] https://aihub.or.kr
[3] https://github.com/as- ideas/ForwardTacotron

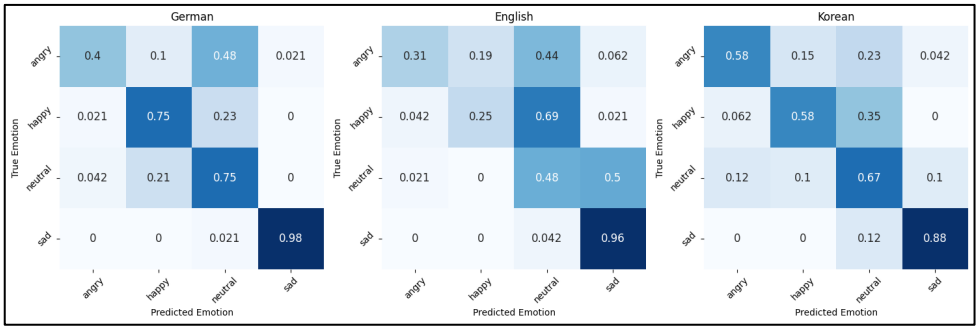

Figure 2: Confusion matrix for categorization (The number of predictions is normalized.)

Valence and arousal

In valence and arousal tests, we evaluated the expressiveness of individual TTS voice and identified a common pattern across different languages. In terms of the valence test, we confirmed in all languages that emotions were distinctively perceived compared to neutral voice, which is illustrated in the upper plot of Figure 3. Note that the y-axis indicates the degree of valence, having neutral speech as a baseline. If y increases, it means the degree of positivity increases, while the decrease of y indicates the stronger negativity. Firstly, German emotional TTS was found to be the most distinctive among three languages. In particular, while angry, happy and sad emotions were perceived to be distinctive compared to the neutral voice in German voice., happy and sad for English and angry and sad for Korean were rated to be distinctive.

The result from the arousal test is also interpreted in a same manner in the second plot of Figure 3. In this test, it was found that both German and English emotional TTS models convey expressiveness in all three emotions, whereas participants only perceived happy and sad emotions distinctively compared to neutral speech in Korean.

To better understand the pattern of emotion perception across different languages, Table 2 summarized results from valence and arousal tests. In this table, we identified that sadness was perceived as a negative and clam state across three languages. Furthermore, happiness was rated as a positive and excited state while anger was evaluated as negative and excited across different languages.

Table 2: Summary of emotion perceptions across different languages: ↑ and ↓ indicate that evaluation level includes the higher and lower level, '+' and '-' mean 'positive' and 'negative', 'E' and 'C' stand for 'excited' and 'calm'.

Test	**Language**	**A** vs. N	**H** vs. N	**S** vs. N
Valence	German	✓ (-↓)	✓ (+↑)	✓ (-)
	English	x	✓ (+)	✓ (-↑)
	Korean	✓ (-)	x	✓ (-)
Arousal	German	✓ (E↑)	✓ (E)	✓ (C ↓)
	English	✓ (E)	✓ (E)	✓ (C ↓)
	Korean	x	✓ (E↑)	✓ (C ↓)

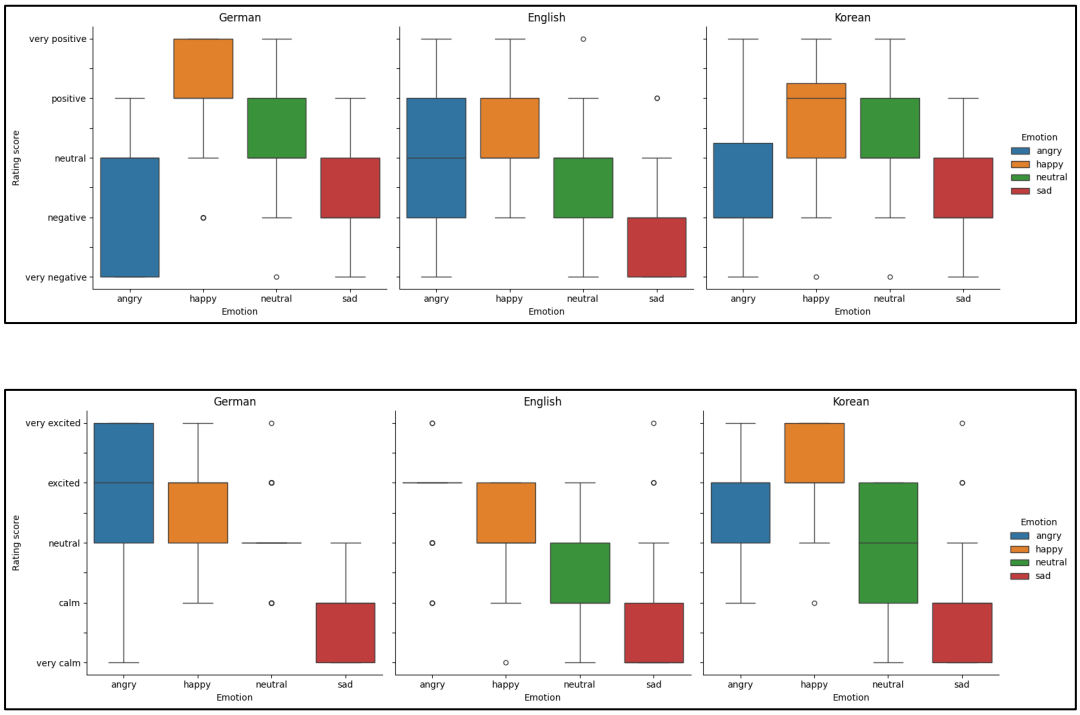

Figure 3: Valence (upper) and arousal (below) evaluation

4. Discussion and Conclusion

In this study, emotional speech synthesized with three different languages was perceptually evaluated by listeners from different language and culture backgrounds.

The experiment was conducted with respect to four tests: naturalness, categorization, valence and arousal. In the naturalness test, it is found that synthesized speech in German was rated as the most natural voice while emotional speech in English received the lowest naturalness score. The poor naturalness in English is presumably due to the considerable small duration of training data. While this underlines the importance of the amount of training data, it is interesting to note that despite the largest amount of training data, Korean TTS voice was rated as the second natural speech among three languages. One of potential explanations for this could be the absence of the linguistic knowledge affecting naturalness evaluation.

According to our research aims, we could evaluate the expressiveness of our TTS models in categorization, valence and arousal tests. Particularly, it was confirmed in valence and arousal tests that German emotional TTS was the most expressive among three different language models. Emotional voice in English and Korean also showed expressiveness to some extent, but it depended on the type of emotions, i.e., high arousal emotions. For example, angry voice in English and happy voice in Korean were not distinctive compared to neutral speech in respective languages. This is likely due to the low naturalness in English speech, and, at the same time, it may present the limitation of our TTS in terms of conveying features of high arousal emotions e.g., high pitch.

Meanwhile, we identified a common pattern of emotion perception in line with the universalism theory. Specifically, sad emotion displayed a common pattern in listeners' perception, which is confirmed categorization, valence and arousal tests. Consequently, this draws us to the idea that lack of linguistic background may not hinder the basic emotion perception in synthesized voice although this should be further investigated by enlarging the number of participants in each cultural and linguistic background, as well as comparing the synthesized and human voice to see how listeners compare them.

As a main contribution, this study presents an approach to evaluate emotional TTS in different languages by applying several assessment criteria. In this study, we not only confirmed the expressiveness of emotions in each language, but also identified a common pattern in emotion perception.

Bibliography

[1] Hofstede, G. Cultures and Organizations Software of the Mind; Intercultural Cooperation and its importance for Survival. New York u. a: McGraw-Hill, 2010, p.5.

[2] Darwin, C. The Expression of the Emotions in Man and Animals. Cambridge University Press, 1872.

[3] Ekman, P.: An argument for basic emotions. Cognition & Emotion, 6 (1992), 169-200.

[4] Trojan, F. Biophonetik. Mannheim, West Germany: Bibliographisches Institut, 1975.

[5] Laver, J. The phonetic description of voice quality. Cambridge, England: Cambridge University Press, 1980.

[6] Tartter, V. C.: Happy talk: Perceptual and acoustic effects of smiling on speech. Perception and Psychophysics, 27 (1980), 24-27.

[7] Scherer, K. R.: Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99(2) (1986), 143–165.

[8] Ekman, P., & Friesen, W.V.: A New Pan-Cultural Facial Expression of Emotion. Motivation and Emotion, 10 (1986), 159-168.

[9] Matsumoto, D., & Takeuchi, S.: Emotions and intercultural communication. Ibunka communication kenkyu (Intercultural Communication Research, Kanda University of International Studies Intercultural Communication Institute), 11 (1998), 1-32.

[10] Matsumoto, D., Leroux, J., & Yoo, S. H.: Emotion and intercultural communication. 関西学院大学社会学部紀要 (The Faculty of Sociology, Kwansei Gakuin University), 99 (2005), 15-38.

[11] Dewaele, J. Culture and Emotional Language. In Farzad Sharifian (Ed.). The Routledge Handbook of Language and Culture. Oxford: Routledge, 2015, p. 357-370.

[12] Hareli, S., Kafetsios, K., & Hess, U.: A cross-cultural study on emotion expression and the learning of social norms. Frontiers in Psychology, 6 (2015).

[13] Lim, N.: Cultural differences in emotion: Differences in emotional arousal level between the East and the West. Integrative Medicine Research, 5(2) (2016), 105-109.

[14] Tsai, J. L.: Ideal affect: Cultural causes and behavioral consequences. Perspectives on Psychological Science, 2(3) (2007), 242–259.

[15] Gessinger, I., Cohn, M., Cowan, B.R., & Zellou, G., Möbius, B.: Cross-linguistic Emotion Perception in Human and TTS Voices. Proc. INTERSPEECH 2023 (2023) 5222-5226.

[16] Steiner, I., Schröder, M., & Klepp, A.: The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech. Phonetik & Phonologie 9 (2013), 83–84.

[17] Zhou, K., Sisman, B., Liu, R., & Li, H.: Emotional Voice Conversion: Theory, Databases and ESD. Speech Communication, 137 (2021), 1-18.

[18] Zalkow, F., Sani, P., Fast, M., Bauer, J., Joshaghani, M., Lakshminarayana, K.K., Habets, E.A.P., Dittmar, C.: The AudioLabs System for the Blizzard Challenge 2023. Proc. 18th Blizzard Challenge Workshop, (2023) 63-68.

[19] Mustafa, A., Pia, N., & Fuchs, G.: StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2020) 6034-6038.

[20] Schoeffler, M. et al.: webMUSHRA — A Comprehensive Framework for Web-based Listening Tests. Journal of Open Research Software. 6(1) (2018), p.8.