Rhythmic Similarity hypothesis for cross-cultural vocal emotion recognition

Yachan Liang[1], Jiyoun Choi[2], Mirjam Broersma[1], Martijn Goudbeek[3], Agnieszka Konopka[4]

[1]Centre for Language Studies, Radboud University Nijmegen, The Netherlands

[2]Sookmyung Women's University, Seoul, Korea

[3]Tilburg University, The Netherlands

[4]University of Aberdeen, UK

ABSTRACT

This study examines the newly formulated Rhythmic Similarity hypothesis, which proposes that cross-cultural vocal emotion recognition is more accurate between users of languages with similar rhythmic structures. To disentangle the close relationship between culture and language, this study tested how well American English and French listeners recognized emotions in recordings produced by native speakers of Dutch and Korean. Since French and Korean share similar rhythmic structures, the prediction was that French listeners would outperform American English listeners in recognizing emotions in Korean. However, this prediction was not supported by our data, disconfirming the Rhythmic Similarity hypothesis. Furthermore, emotion recognition accuracy for both listener groups (American English and French) was higher in Dutch than in Korean, supporting the Cultural Proximity [1] and Language Distance hypotheses [2]. Additionally, recognition accuracy was above-chance for all emotions and was affected by arousal, valence, and basicness in ways consistent with previous findings [3, 4, 5].

Keywords: Rhythmic Similarity hypothesis, vocal emotion recognition, cross-cultural, cross-linguistic

1. INTRODUCTION

Previous research has shown that emotion recognition is a product of interactions between universal and culture-/language-specific factors [4, 6]. Listeners from different cultures and with different native languages recognize emotions above chance even in an unknown language, which shows that emotions are conveyed in a similar way across cultures and languages. At the same time, emotion recognition is more accurate when speakers and listeners share the same culture and native language, a phenomenon known as the in-group advantage [7]. However, the specific contributions of culture and language to emotion recognition remain unknown.

Most studies have addressed emotion recognition predominantly either from a cultural perspective or from a linguistic perspective. Disentangling the effect of culture and language has been notoriously difficult. For this reason, two separate theories address the relative contributions of culture and language to emotion recognition: the Cultural Proximity and Language Distance hypotheses.

The Cultural Proximity hypothesis by Elfenbein and Ambady [1] proposes that listeners with relatively similar cultural backgrounds can identify each other's emotional expressions more accurately than those with very different ones. This theory thus likens cultural variation to "dialects" in languages. For instance, to examine the effect of cultural differences on emotion recognition, Laukka et al. [8] tested emotion recognition in English among listeners from five different English-speaking countries (America, Australia, India, Kenya, and Singapore). They found that listeners were more accurate in recognizing emotions produced by speakers from their own culture than from a different one.

The Language Distance hypothesis introduced by Scherer et al. [2] claims that it is easier for listeners to decode emotions produced in a language that is typologically similar to their native language than in a language that is typologically different. In a pioneering study, Scherer et al. [2] investigated the influence of language distance on emotion recognition. They presented 30 pseudo-utterances expressing five basic emotions (anger, fear, joy, sadness, and neutral) produced in German to listeners from nine different countries whose native languages were German, Dutch, English, French, Italian, Spanish, and Indonesian. While recognition accuracy in each listener group was above chance, German listeners had the highest accuracy, with Dutch and English listeners following next, which is consistent with the Language Distance hypothesis. In contrast, Indonesian listeners, whose native language is typologically the least similar to German, had the lowest accuracy. Similarly, Pell et al. [9] tested the recognition of five basic emotions (anger, disgust, fear, joy, and sadness) in pseudo-utterances produced

in Argentine Spanish, Arabic, German, and English, by monolingual listeners of Argentine Spanish. The results showed that Argentine Spanish listeners had above-chance recognition accuracy in all languages, but with the highest accuracy in their native language.

The close relationship between culture and language makes it difficult to disentangle the specific effects of the two on emotion recognition. For example, in Scherer et al.'s [2] study, Dutch and English are not only typologically, but also culturally closer to German than to some of the other languages/cultures in the sample, creating a potential confound. Furthermore, while the Language Distance hypothesis proposes a central role for linguistic similarity, it does not specify which aspects of language typology might play a role. In the present study, we investigate the role of one potentially highly relevant language property, namely rhythmic structure on vocal emotion recognition, with the aim of disentangling the effect of culture and language. We tested emotion recognition in listeners of American English and French exposed to recordings produced by Dutch and Korean speakers. In terms of culture and language, Dutch, American English, and French are related, as they are Indo-European languages spoken by people from Western cultures. Korean, however, is a non-Indo-European language spoken by people from an Asian culture, and is thus unrelated to the other three cultures and languages.

With regard to the rhythmic structure, however, these languages are grouped differently (see Table 1). In Dutch and American English, the foot is the main grouping element in rhythm, and there is lexical stress, i.e., prosodic prominence is used to differentiate word meanings [10, 11]. In French [12] and Korean [13], on the other hand, the phrase is the main rhythmic grouping element, and there is no lexical stress. There are Intonational Phrases (IP) and Accentual Phrases (AP) in French and Korean, and the phrase boundary is typically signaled by a final rising pitch movement and lengthening [12, 14, 15].

Table 1: The relationships between the four languages in terms of culture and language.

	Dutch		Korean	
	Culture	**Language: Overall typology**	**Culture**	**Language: Rhythmic similarity**
American English	Relatively similar	Relatively similar	Dissimilar	Dissimilar
French	Relatively similar	Relatively similar	Dissimilar	Similar

This constellation of languages allows us to investigate whether rhythmic similarities facilitate cross-cultural vocal emotion recognition. Therefore, building upon the Language Distance hypothesis, we propose the Rhythmic Similarity hypothesis, which predicts that cross-linguistic communication of vocal emotion will be more accurate if the languages have similar rhythmic structures than if they have dissimilar rhythmic structures.

This study therefore addresses two questions. First, it examines whether American English and French listeners recognize emotions more accurately in Dutch than in Korean. According to the Cultural Proximity and Language Distance hypotheses, we hypothesize that both groups of listeners will perform better in Dutch than in Korean because American English and French are culturally and linguistically closer to Dutch than Korean (Hypothesis 1). Second, it examines whether French listeners perform better than American English listeners on the Korean recordings. Based on the Rhythmic Similarity hypothesis, we predict that French listeners will outperform American English listeners in Korean since Korean is rhythmically similar to French but not to American English (Hypothesis 2). In addition, we will present a series of analyses comparing recognition accuracy to chance level, and assessing the effects of arousal, valence, and basicness on emotion recognition accuracy.

2. METHOD

2.1. Participants

Participants (referred to as "listeners" below) were twenty-five native American English listeners (19 females, 6 males, age: $M = 20.6$, $SD = 1.94$) who were students at Northwestern University, Chicago, and thirty native French listeners (22 females, 8 males, age: $M = 22.7$, $SD = 4.10$) who were students at the university École Normale Supérieure, Paris. None of the participants reported any speech or hearing problems, or any knowledge of Dutch or Korean. All participants were given either course credits or a small payment as a reward for their participation.

2.2. Materials

As auditory stimuli, we used all the vocal expressions from the Demo/Koremo (Dutch emotion/Korean emotion) corpus collected by Goudbeek and Broersma [16]. To avoid any semantic cues to emotion recognition, the corpus uses a pseudo-phrase [nuto hɔm sɛpikaŋ] which is phonologically legal in both Dutch and Korean, meeting the "stimuli equivalence" requirement proposed by Matsumoto [17]. The pseudo-phrase was produced by professional Dutch and Korean voice actors (referred to as "speakers" below). This corpus includes an equal number of basic and non-basic emotions, which were all used in the present study, contrary to prior studies that have mainly focused on basic emotions like anger, disgust, fear, happiness, sadness, and

surprise [18]. Moreover, the stimuli are balanced on two dimensions that are crucial in the understanding of emotions [19], namely arousal (high-arousal vs. low-arousal) and valence (positive vs. negative; Table 2). The stimuli consisted of 256 portrayals (8 emotions x 2 tokens per emotion x 8 speakers x 2 languages) in total. For more information regarding the corpus and the recording procedure, please refer to Goudbeek and Broersma [16].

Table 2: The eight emotions included in this study in a valence by arousal grid (reproduced from Goudbeek & Broersma, 2010, p.2212), with basic emotions denoted with *.

		Valence	
		Positive	Negative
Arousal	High	Joy*	Anger*
		Pride	Fear*
	Low	Tenderness	Sadness*
		Relief	Irritation

2.3. Procedure

All participants were tested individually in a sound-attenuated room at their university. On each trial, a computer screen in front of the participant showed an "emotion wheel" (Figure 1), with eight emotions written in the participants' native language (English or French), each with four circles in different sizes indicating different intensities. Participants listened to the recordings via high-quality headphones. There was no time constraint for their responses. The experiment was administered using JATOS [20] on a standard laboratory computer and took around 35-45 minutes.

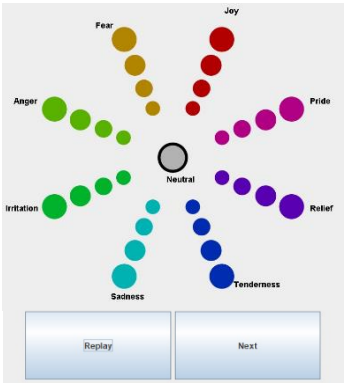

Figure 1: The emotion wheel in English

All participants were given written instructions in their native language, asking them to listen to each recorded stimulus, to choose the emotion they believed the recording expressed from the eight choices shown on the screen, and to indicate the intensity of that emotion. They could also choose Neutral if they felt the recording was not emotional at all. They made their response by clicking on one of the circles on the emotion wheel. They could listen to each recording as often as they preferred. In this study, we only analyzed categorical responses (i.e., recognition accuracy, but not intensity ratings).

The experiment started with a block containing the 128 Korean stimuli, followed by a block containing the 128 Dutch stimuli. Prior to each block, participants were informed about its language, and they performed eight practice trials.

3. RESULTS AND DISCUSSION

Analyses were performed in R [21] with a series of logistic mixed effects models run with the *lme4* package [22]. The models included a combination of the following five factors: Speaker Language (Dutch vs. Korean recordings), Listener Language (American English vs. French listeners), Arousal (high-arousal vs. low-arousal emotions), Valence (positive vs. negative emotions), and Basicness (basic vs. non-basic emotions). We used regression-style contrast coding (i.e., -.5 and .5 contrast codes for the first and second levels of each factor listed above) in all models. The outcome variable was the accuracy of emotion recognition (correct vs. incorrect).

Each model used the maximal random structure justified by the experimental design that allowed convergence. All models initially included random by-participant and by-item intercepts, by-participant slopes for Speaker Language, Arousal, Valence, and Basicness, and by-item slopes for Listener Language. In cases of non-convergence, models were sequentially simplified by dropping the random slope for the variable accounting for the least variance [23] until convergence was reached.

Table 3: Summary of results of the logistic mixed effects model analysis addressing Hypothesis 1 and 2 (only by-item slopes improved model fit).

	Estimates			
	β	Exp(β)	*SE*	*z* value
Model1(Hypothesis 1)				
Intercept	-0.44	0.64	0.10	-4.65***
Speaker Language (*s*)	-0.49	0.61	0.17	-2.83**
Listener Language (*i*)	0.17	1.19	0.10	1.69.
Speaker Language (*s*) x Listener Language (*i*)	-0.03	0.97	0.13	-0.23

"***" $p < .001$, "**" $p < .01$, "*" $p < .05$, "." $.05 < p < .10$

***Notes*: *(s)* and *(i)* indicate the inclusion of by-participant and by-item random slopes respectively.**

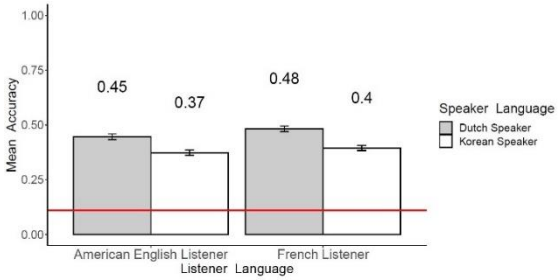

Figure 2: Recognition accuracy for Dutch and Korean recordings by American English and French listeners (by-participant means). Red line indicates chance performance (.11). Error bars are ±1 SE.

Hypothesis 1*: Both listener groups are predicted to have higher recognition accuracy in Dutch compared to Korean.* We tested this hypothesis in a model that included the interaction between Speaker Language and Listener Language (Table 3). The model revealed a main effect of Speaker Language, indicating that, consistent with the Cultural Proximity and Language Distance hypotheses, both American English and French listeners recognized emotions more accurately in Dutch than in Korean (a .16 difference; Figure 2). Additionally, there was a marginally significant main effect of Listener Language, as French listeners performed slightly better than American English listeners overall (.06 higher recognition accuracy).

Hypothesis 2*: French listeners are predicted to outperform American English listeners in Korean recordings.* There was no interaction between Speaker and Listener Language: French listeners did not show higher accuracy than American English listeners when responding to Korean recordings (mean accuracy: .40 vs. .37). Thus, the Rhythmic Similarity hypothesis was not supported by our data.

In addition, we tested for the main effects of three emotion variables (Arousal, Valence, and Basicness), and for their interactions with Speaker Language and Listener Language in separate models (Table 4). The results showed a number of similar patterns as in previous studies [3, 4, 5]: higher accuracy for (i) low-arousal, (ii) negative, and (iii) basic emotions than for (i) high-arousal, (ii) positive, and (iii) non-basic emotions (i.e., three main effects). The analyses also showed a significant two-way interaction between Listener Language and Arousal, as well as Listener Language and Basicness (all *zs* < -2.74), but not with Valence. Importantly, no three-way interactions reached significance.

Table 4: Summary of results of the logistic mixed effects model analyses with Arousal, Valence, and Basicness (all random slopes improved model fit, except by-participant slopes for Speaker Language).

	Estimates			
	β	Exp(*β*)	*SE*	*z* value
Analysis with Arousal				
Intercept	-0.45	0.64	0.10	-4.74***
Arousal (*s*)	0.49	1.62	0.19	2.58**
Speaker Language (*s*) x Listener Language (*i*)	0.003	1.00	0.12	0.03
Listener Language (*i*) x Arousal (*s*)	-0.86	0.42	0.19	-4.44***
Speaker Language (*s*) x Listener Language (*i*) x Arousal(*s*)	-0.12	0.89	0.24	-0.49
Analysis with Valence				
Intercept	-0.45	0.64	0.09	-5.20***
Valence (*s*)	-1.34	0.26	0.16	-8.21***
Speaker Language (*s*) x Listener Language (*i*)	-0.02	0.98	0.13	-0.15
Listener Language (*i*) x Valence (*s*)	0.001	1.00	0.18	0.01
Speaker Language (*s*) x Listener Language (*i*) x Valence (*s*)	0.29	1.34	0.26	1.14
Analysis with Basicness				
Intercept	-0.45	0.64	0.09	-4.94***
Basicness (*s*)	-1.06	0.35	0.17	-6.16***
Speaker Language (*s*) x Listener Language (*i*)	-0.02	0.98	0.13	-0.12
Listener Language (*i*) x Basicness (*s*)	-0.48	0.62	0.17	-2.74**
Speaker Language (*s*) x Listener Language (*i*) x Basicness (*s*)	-0.05	0.95	0.25	-0.20

"***" $p < .001$, "**" $p < .01$, "*" $p < .05$, "." $.05 < p < .10$

Notes: *(s)* and *(i)* indicate the inclusion of by-participant and by-item random slopes respectively.

4. CONCLUSION

The primary goal of this study was to assess the role of language in cross-cultural/linguistic emotion recognition by comparing the performance of American English and French listeners on Dutch and Korean recordings. Building upon the Language Distance hypothesis [2], the newly proposed Rhythmic Similarity hypothesis predicted that cross-cultural (or cross-linguistic) communication of vocal emotion would be facilitated in languages with a shared rhythmic structure. However, this hypothesis was not supported by our data. Even though French and Korean share similar rhythmic structures, with prosodic phrases marked by a final rising pitch and lengthening [12], French listeners were not better at recognizing emotions in Korean than American English listeners were.

A second aim was to investigate the effects of Cultural Proximity and Language Distance on emotion recognition. The results that both groups of listeners recognized vocal emotions more accurately in Dutch, which is both culturally and linguistically closer to American English and French than to Korean, are in line with previous findings that cultural and linguistic similarities improve emotion recognition cross-culturally/linguistically [1, 2, 7].

Additionally, we replicated earlier findings that listeners are capable of recognizing discrete emotions above chance even when they are produced by a speaker from a different culture and language [4]. Furthermore, we found the effects of arousal, valence, and basicness on the recognition of emotion consistent with previous research in this domain [3, 4, 5].

To conclude, our findings corroborate earlier work on the effects of culture and language on vocal communication of emotion. However, since we found no evidence for the Rhythmic Similarity hypothesis, the precise mechanisms by which linguistic similarity contributes to emotion recognition remain elusive.

5. REFERENCES

[1] H. A. Elfenbein and N. Ambady, "Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition," *J. Cross. Cult. Psychol.*, vol. 34, no. 1, pp. 92–110, 2003, doi: 10.1177/0022022102239157.

[2] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *J. Cross. Cult. Psychol.*, vol. 32, no. 1, pp. 76–92, 2001, doi: 10.1177/0022022101032001009.

[3] Y. Liang, M. Goudbeek, A. Konopka, J. Choi, and M. Broersma, "Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design," *Lang. Speech*.

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

[5] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 23, p. E3086, 2015, doi: 10.1073/pnas.1508604112.

[6] P. Ekman, "Universals and cultural differences in facial expressions of emotion," *J.K. Cole (Eds.), Nebraska Symposium on Motivation*, vol. 19. University of Nebraska Press, Lincoln, pp. 207–283, 1972.

[7] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychol. Bull.*, vol. 128, no. 2, pp. 203–235, 2002, doi: 10.1037/0033-2909.128.2.203.

[8] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *J. Pers. Soc. Psychol.*, vol. 111, no. 5, pp. 686–705, 2016, doi: 10.1037/pspi0000066.

[9] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *J. Nonverbal Behav.*, vol. 33, no. 2, pp. 107–120, 2009, doi: 10.1007/s10919-008-0065-7.

[10] C. Gussenhoven, "Transcription of Dutch intonation," *Sun-Ah Jun (Eds.), The phonology of intonation and phrasing*, Oxford: Oxford University Press, 2005, pp. 118–145. doi: 10.1093/acprof.

[11] P. A. Bertan, "Prosodic Typology: On the Dichotomy between Stress-Timed and Syllable-Timed Languages A," *Lang. Des.*, vol. 2, pp. 103–130, 1999.

[12] S. Jun and C. Fougeron, "Realizations of accentual phrase in French intonation," vol. 14, no. 2002, pp. 147–172.

[13] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *J. Phon.*, vol. 40, no. 3, pp. 351–373, 2012, doi: 10.1016/j.wocn.2012.02.003.

[14] S.-A. Jun, "Intonational phonology of Seoul Korean revisited," *Japanese/Korean Linguist. 14*, pp. 15–26, 2006.

[15] J. Kim, C. Davis, and A. Cutler, "Similarity : II . Syllable Rhythm," vol. 51, no. 4, pp. 343–359, 2008.

[16] M. Goudbeek and M. Broersma, "The Demo/Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*., 2010, pp. 2211–2215.

[17] D. Matsumoto, "Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence," *Psychol. Bull.*, vol. 128, no. 2, pp. 236–242, 2002, doi: 10.1037/0033-2909.128.2.236.

[18] P. Ekman, "Are there basic emotions?," *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992, doi: 10.1037/0033-295X.99.3.550.

[19] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[20] K. Lange, S. Kühn, and E. Filevich, "'Just another tool for online studies' (JATOS): An easy solution for setup and management of web servers supporting online studies," *PLoS One*, vol. 10, no. 6, pp. 1–14, 2015, doi: 10.1371/journal.pone.0130834.

[21] R Core Team, "R: A language and environment for statistical computing," in *R Foundation for Statistical Computing*, Vienna, 2022. [Online]. Available: https://www.r-project.org

[22] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[23] D. . Barr, R. Levy, C. Scheepers, and H. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Mem. Lang.*, vol. 68, no. 3, pp. 255–278, 2013, doi: 10.1016/j.jml.2012.11.001.