When speech becomes emotional: cross-cultural vocal emotion recognition in Dutch and Korean

Liang, Y.

Summary

Since Darwin's *The Expression of the Emotions in Man and Animals* (1872; reprint in 1998), emotion research has drawn increasing attention in different areas such as biology, psychology, and linguistics. Darwin proposed that the production and perception of emotions are biologically determined and universal, whereas the social constructivist theory argued that emotions are culturally and linguistically constructed (Harre, 1986). According to the dialect theory, emotion recognition is fundamentally universal, but cultural and linguistic variants in expressive styles can affect recognition accuracy (Elfenbein & Ambady, 2002b). Recently, a growing consensus holds that cross-cultural emotion recognition is the outcome of interactions between universal, cultural, and linguistic factors (Elfenbein, 2013; Elfenbein, Mandal et al., 2002; Mesquita & Frijda, 1992).

While prior studies have shown that vocal emotions can be recognized across cultures, the question still is to what extent the production and perception of emotions are universal or influenced by culture and language. This dissertation addresses this issue by examining how emotions are vocally expressed and perceived in two typologically and culturally distinct languages—Dutch and Korean. Using the Demo/Koremo corpus (Broersma et al., 2025), this dissertation explores cross-language vocal emotion recognition by both discrete and dimensional approaches. There were three perception experiments with listeners and one study based on a comprehensive acoustic analysis of the stimuli, examining 1) cross-cultural and/or cross-linguistic vocal emotion recognition by Dutch and Korean listeners; 2) intensity ratings of vocal emotions by Dutch and Korean listeners; 3) the relative contributions of emotions, speaker language, and gender to acoustic parameters; 4) the impact of culture and prosodic similarity on vocal emotion recognition by American English and French listeners.

The Introductory **Chapter 1** presents an overview of the theoretical and empirical studies on the production and perception of emotions, discussing the discrete and dimensional approaches. It introduces a balanced "two-to-two" design, with speakers and listeners from two typologically different languages—Dutch and Korean. It included four basic emotions (anger, fear, joy, sadness) and four non-basic emotions (irritation, pride, relief, tenderness), balanced in arousal (high arousal/excited vs. low arousal/subdued) and valence (positive/pleasant vs. negative/unpleasant). To avoid semantic cues, all emotions were produced on the basis of a pseudo-sentence /nuto hɔm sɐpik

ɑŋ/, which is pronounceable in Dutch and in Korean but is meaningless in either language. The eight emotions were expressed by eight Dutch and eight Korean voice actors, with the same number of females and males in each language group, allowing us to study gender-related differences in prosodic expression of emotions (Klatt & Klatt, 1990). Each actor produced the same emotions twice, resulting in a total of 256 portrayals (8 emotions × 8 actors × 2 tokens × 2 languages). Finally, this chapter outlines the research questions addressed in each of the following chapters.

Chapter 2 "Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design" examines recognition of vocal Dutch and Korean emotions by Dutch and Korean listeners. This chapter examined 1) whether there is an in-group advantage in cross-cultural emotion recognition; 2) whether there is a difference in recognition accuracy between high-arousal and low-arousal emotions; 3) whether there is a difference in recognition accuracy between positive and negative emotions; 4) whether there is a difference in recognition accuracy between basic and non-basic emotions.

The results revealed that both listener groups recognized vocal emotions above chance, even in the unknown language. Additionally, both listener groups demonstrated an in-group advantage in vocal emotion recognition, such that listeners recognize vocal emotions produced in their native language more accurately than those expressed in the unknown language. These findings support the idea that cross-cultural emotion recognition results from an interaction between universal and culture-/language-specific factors (Elfenbein, 2013; Elfenbein & Ambady, 2002b). Moreover, this study examined the dimensional effects of arousal, valence, and basicness on vocal emotion recognition, within and across cultures. Recognition accuracy proved higher for low-arousal, negative, and basic emotions than for high-arousal, positive, and non-basic emotions, both within and across cultures.

Chapter 3 "Interpreting the intensity of vocal emotions across cultures" investigates the intensity ratings by Dutch and Korean listeners collected in Study 1. Intensity is the strength of an emotion perceived by the receiver (Bänziger & Scherer, 2005; Diener et al., 1985; Larsen & Diener, 1987; Sonnemans & Frijda, 1994). People tend to react more strongly to emotions with higher intensity than to those with lower intensity. Further, individuals usually give higher intensity ratings to emotions expressed by members from the same or similar culture/linguistic group than by members from a typologically different group, which is referred to as the in-group intensity bias (Kommattam et al., 2019). This chapter examined 1) whether accurate

trials receive higher intensity ratings than inaccurate ones; 2) whether there is an in-group bias for intensity ratings across cultures; 3) whether the three (dimensional) binary splits by arousal, valence, and basicness can reliably predict intensity ratings.

The findings demonstrated that accurate trials received higher intensity ratings than inaccurate ones from both groups of listeners. Specifically, anger received the highest intensity ratings by either listener group in accurate as well as inaccurate trials. However, no in-group bias was found in the intensity ratings. Moreover, intensity ratings were closely related to arousal, valence, and basicness, such that intensity ratings were higher for high-arousal, negative, and basic emotions than for low-arousal, positive, and non-basic emotions.

Chapter 4 "Classifying emotions from acoustic parameters" analyzed 17 acoustic parameters, which were classified into five categories: 1) pitch-related, 2) amplitude-related, 3) spectrum-related, 4) duration-related, and 5) perturbation-related. This study examined 1) how the acoustic patterns differ across emotions, speaker language, and gender; 2) how accurately the recognition of emotions can be predicted according to acoustic parameters, and the extent to which the classification improves when the materials are split by language; 3) the extent to which a machine learning classifier (Support Vector Machine, SVM) matches the performance of human listeners in emotion recognition.

The results revealed that each emotion exhibited a distinct acoustic profile, which varied across language and gender. Vocal emotions can be reliably differentiated by an SVM) classifier through optimized combinations of acoustic cues. The classification accuracy improved when Dutch and Korean data were analyzed separately. The confusion matrices suggest that there are relevant qualitative differences between the performance of SVM and human listeners that need to be investigated in more detail.

Chapter 5 "Recognizing vocal emotions in unfamiliar languages" extends the investigation to cross-linguistic recognition by American English and French listeners who were unfamiliar with Dutch and Korean. This study examines the relative influence of the Universality hypothesis, Cultural Proximity, Linguistic Proximity, and emotional dimensions (arousal, valence, and basicness) on emotion recognition. In vocal emotion recognition, listeners identify emotions more easily in languages similar to their own. While emotional dimensions such as arousal, valence, and basicness influence recognition, their specific impact on accuracy is not fully understood. This

study examines how universal, cultural, linguistic, and emotional factors affect the perception of vocal emotions. We selected American English and French listeners because English, a stress-timed language, is prosodically similar to Dutch, while French, a syllable-timed language, is similar to Korean. By comparing recognition accuracy between these groups, we aim to clarify the influence of these factors. The study first assessed whether both groups recognized vocal emotions above chance. Next, we compared their accuracy with Dutch recordings. We then evaluated whether French listeners outperformed American English listeners with Korean recordings, given the prosodic similarities. Finally, we analyzed the effects of arousal, valence, and basicness on vocal emotion recognition.

The results show that both American and French listener groups recognized vocal emotions above chance, supporting the Universality hypothesis (Elfenbein, 2013; Elfenbein & Ambady, 2002a). Moreover, the results show that people recognize emotions more accurately when they share a cultural and linguistic background with the speaker, which finding is in line with Elfenbein and Ambady (2003a).

Chapter 6 "Conclusions and discussion" reviews the research questions of each chapter and summarizes the main findings of each study with a discussion and an integration of the results. This chapter addresses the limitations of the empirical studies as well. In conclusion, it highlights the following five novel findings and contributions to the existing literature on emotion research.

First, our study presents a balanced "two-to-two" cross-over experimental design, with speakers and listeners from two typologically different language and culture. This design allows for a systematic comparison of both within- and between-culture recognition patterns, especially for examining the in-group advantage in emotion recognition. We replicated earlier findings that listeners recognized vocal emotions above chance, even in an unknown language, consistent with the universality hypothesis. Moreover, we found that both listener groups displayed an in-group advantage, such that they recognized vocal emotions more accurately when produced in their native language than in the unknown language.

Second, we employed both discrete and dimensional approaches, including four basic and four non-basic emotions balanced for arousal and valence, providing a more comprehensive understanding of the mechanisms underlying emotion recognition. This study examined not only the recognition accuracy of each of the eight emotions categorically, but also the effects of

three emotional dimensions (arousal, valence, and basicness) on emotion recognition, within and across cultures. We found that recognition accuracy was higher for low-arousal, negative, and basic emotions than for high-arousal, positive, and non-basic emotions, both within and across cultures. To our knowledge, this is the first study that has directly compared the recognition accuracy of low-arousal and high-arousal emotions within and across cultures. This finding bridges the gap in understanding the role of arousal in emotion recognition.

Third, we investigated intensity ratings across cultures. Intensity, as an important dimension of emotions, has received relatively little attention in previous research. We found no in-group bias in intensity ratings. However, the results revealed that intensity ratings were predominantly determined by emotion type rather than by cultural or linguistic factors. Using the dimensional approach, we examined the role of arousal, valence, and basicness in intensity ratings. We found that intensity ratings were generally higher for high-arousal, negative, and basic emotions than for low-arousal, positive, and non-basic emotions.

Fourth, our study provides a comprehensive analysis of 17 acoustic parameters in two typologically different languages, including an under-represented non-Indo-European language—Korean. We found that different vocal emotions display universal and emotion-specific acoustic patterns. Notably, pitch, intensity, and speech rate can successfully differentiate vocal emotions. We also found that laryngeal parameters add nuanced distinctions between vocal emotions. Moreover, vocal emotions exhibit language-dependent acoustic patterns. The articulation rate was remarkably higher in Korean than in Dutch, and there were no speech pauses in Korean. Further, females and males displayed different acoustic characteristics, such that (relative) pitch- and amplitude-related acoustic parameters are generally higher for females than for male actors. Based on the acoustic parameters, we examined whether vocal emotions can be accurately classified by machine learning (SVM models), and compared the performance of machine classifiers for Dutch and Korean in both in-group and out-group conditions. The results revealed that classification rates can be reliably predicted from a combination of acoustic parameters. Notably, like human listeners, machine classifiers performed better in the in-group condition than in the out-group condition for both Dutch and Korean, although the classification rates were above chance in both conditions.

Finally, we examine cross-cultural/linguistic vocal emotion recognition by American English and French listeners, who had no knowledge of Dutch or

Korean. It provides additional evidence to the existing literature on cross-cultural/linguistic emotion recognition by testing the contributions of universality, cultural proximity, prosodic proximity, and emotional dimensions (arousal, valence, and basicness) to cross-cultural vocal emotion recognition. Consistent with the Universality hypothesis, both listener groups generally identified vocal emotions above chance, even in an unknown language. However, we did not find an overall effect of cultural proximity on emotion recognition, since recognition accuracy varied across emotions nor did we find evidence supporting the effect of prosodic proximity on across-cultural/language vocal emotion recognition, as French listeners did not outperform American English listeners in Korean recordings. More importantly, we found that although arousal, valence, and basicness affect vocal emotion recognition, some emotions violated these general patterns. Therefore, while emotional dimensions provide a useful framework to understand recognition accuracy, their explanatory power is insufficient to account for cross-cultural or cross-language variability in emotion recognition.