

**When speech becomes emotional.
Cross-cultural vocal emotion recognition
in Dutch and Korean**

Published by
LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
<http://www.lotschool.nl>

Cover illustration: By the author.

ISBN: 978-94-6093-492-6
DOI: <https://dx.medra.org/10.48273/LOT0707>
NUR: 616

Copyright © 2025: Yachan Liang. All rights reserved.

**When speech becomes emotional.
Cross-cultural vocal emotion recognition
in Dutch and Korean**

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 16 december 2025
klokke 10.00 uur

door

YACHAN LIANG

Promotor: Prof. dr. Claartje Levelt
Co-promotores Prof. dr. Vincent J. van Heuven
Prof. dr. Roeland W. N. M. van Hout
(Radboud University)

Promotiecommissie: Prof. dr. Yiya Chen
Prof. dr. Nivja de Jong
Prof. dr. Marc Swerts (Tilburg University)
Dr. John L. A. Huisman (Uppsala University)

Contents

| | |
|--|----|
| Acknowledgements..... | xi |
| Chapter One..... | 1 |
| General Introduction..... | 1 |
| 1.1 Introduction..... | 1 |
| 1.1.1 Theories of emotion..... | 3 |
| 1.1.2 Empirical studies on cross-cultural emotion recognition..... | 4 |
| 1.1.3 Research on affective neuroscience..... | 6 |
| 1.2 Research methodology..... | 6 |
| 1.3 The current study..... | 8 |
| 1.4 Two approaches..... | 9 |
| 1.5 Overview of the dissertation..... | 9 |
| Chapter Two..... | 13 |
| Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design..... | 13 |
| Abstract..... | 13 |
| 2.1 Introduction..... | 14 |
| 2.1.1 Cross-cultural vocal emotion recognition..... | 14 |
| 2.1.2 Methodological considerations..... | 16 |
| 2.1.3 The present study..... | 20 |
| 2.2 Method..... | 23 |
| 2.2.1 Auditory materials..... | 23 |
| 2.2.2 Visual materials..... | 25 |
| 2.2.3 Participants..... | 26 |
| 2.2.4 Procedure..... | 27 |
| 2.3 Results..... | 27 |
| 2.3.1 Above-chance cross-cultural emotion recognition (Hypothesis 1)..... | 28 |
| 2.3.2 The in-group effect in emotion recognition (Hypothesis 2)..... | 29 |
| 2.3.3 The effect of Arousal on emotion recognition (Hypothesis 3)..... | 30 |
| 2.3.4 The effect of Valence on emotion recognition (Hypothesis 4)..... | 36 |
| 2.3.5 The effect of Basicness on emotion recognition (Hypothesis 5)..... | 41 |
| 2.4 Discussion..... | 43 |
| Chapter Three..... | 47 |
| Interpreting the intensity of vocal emotions across cultures..... | 47 |
| Abstract..... | 47 |

| | |
|--|-----|
| 3.1 Introduction..... | 48 |
| 3.1.1 The intensity of emotions | 48 |
| 3.1.2 Cross-cultural perception of emotional intensity | 49 |
| 3.1.3 The relationship between intensity and emotional dimensions..... | 50 |
| 3.1.4 The present study..... | 52 |
| 3.2 Method..... | 53 |
| 3.2.1 Participants | 53 |
| 3.2.2 Stimuli..... | 53 |
| 3.2.3 Procedure | 54 |
| 3.2.4 Statistical analyses..... | 55 |
| 3.3 Results..... | 56 |
| 3.3.1 The in-group bias in intensity ratings across all responses (Hypothesis 1)..... | 56 |
| 3.3.2 The in-group bias in intensity ratings across correct responses (Hypothesis 2)..... | 60 |
| 3.3.3 The effect of Arousal on intensity ratings (Hypothesis 3)..... | 62 |
| 3.3.4 The effect of Valence on intensity ratings (Hypothesis 4) | 65 |
| 3.3.5 The effect of Basicness on intensity ratings (Hypothesis 5) | 69 |
| 3.3.6 Analyses of Arousal, Valence, and Basicness compared..... | 72 |
| 3.4 Discussion..... | 73 |
| 3.5 Conclusion | 78 |
| Chapter Four | 81 |
| Classifying emotions cross-linguistically from acoustic parameters..... | 81 |
| Abstract..... | 81 |
| 4.1 Introduction..... | 82 |
| 4.1.1 Acoustic characteristics of vocal expression of emotions | 82 |
| 4.1.2 Classifying vocal emotions | 84 |
| 4.1.3 The cross-language perspective of our study and the hypotheses . | 85 |
| 4.2 Method..... | 86 |
| 4.2.1 Stimuli..... | 86 |
| 4.2.2 Acoustic analysis of the speech stimuli | 88 |
| 4.2.3 Analysis..... | 91 |
| 4.3 Results..... | 91 |
| 4.3.1 Effects of emotion, speaker language, and gender (Hypotheses 1-3) | 92 |
| 4.3.2 Vocal emotion recognition by Support Vector Machine (SVM) (Hypothesis 4)..... | 102 |
| 4.3.3 Comparison of vocal emotion recognition by machines and by human listeners (Hypothesis 5) | 104 |

| | |
|--|-----|
| 4.4 Discussion..... | 108 |
| 4.5 Conclusion | 111 |
| Chapter Five..... | 113 |
| Recognizing vocal emotions in unfamiliar languages | 113 |
| Abstract..... | 113 |
| 5.1 Introduction..... | 114 |
| 5.1.1 Universality hypothesis | 114 |
| 5.1.2 The Cultural Proximity hypothesis | 115 |
| 5.1.3 The Language Distance hypothesis..... | 116 |
| 5.1.4 The role of prosodic proximity in vocal emotion recognition | 117 |
| 5.1.5 Predicting recognition accuracy based on emotional dimensions | 118 |
| 5.1.6 Research questions and hypotheses | 119 |
| 5.2 Method..... | 120 |
| 5.2.1 Materials | 120 |
| 5.2.2 Participants | 121 |
| 5.2.3 Procedure | 122 |
| 5.3 Results..... | 122 |
| 5.3.1 Above-chance recognition accuracy by both groups of listeners (Hypothesis 1)..... | 124 |
| 5.3.2 Recognition accuracy in Dutch recordings by both groups of listeners (Hypothesis 2) | 126 |
| 5.3.3 Recognition accuracy in Korean recordings by both groups of listeners (Hypothesis 3) | 127 |
| 5.3.4 The effect of Arousal on accuracy (Hypothesis 4) | 127 |
| 5.3.5 The effect of Valence on accuracy (Hypothesis 5) | 131 |
| 5.3.6 The effect of Basicness on accuracy (Hypothesis 6)..... | 134 |
| 5.3.7 Comparing the dimensional and discrete emotional effects..... | 137 |
| 5.4 Discussion..... | 139 |
| 5.5 Conclusion | 141 |
| Chapter Six | 143 |
| Conclusion and discussion..... | 143 |
| 6.1 Introduction..... | 143 |
| 6.2 Main findings..... | 144 |
| 6.2.1 Investigating cross-cultural vocal emotion recognition with an affectively balanced design..... | 144 |
| 6.2.2 Interpreting the intensity of vocal emotions across cultures | 145 |
| 6.2.3 Classifying emotions from acoustic parameters | 146 |

| | |
|--|-----|
| 6.2.4 Universal patterns, Cultural Proximity, Linguistic Proximity, and emotional dimensions in cross-cultural vocal emotion recognition | 147 |
| 6.3 Discussion..... | 148 |
| 6.3.1 Adapting the dimensional approach..... | 148 |
| 6.3.2 Cluster analysis: The cross-cultural perspective: separating and confusing emotions | 149 |
| 6.3.3 In-group advantage..... | 150 |
| 6.3.4 Prosodic structure and vocal emotion recognition | 151 |
| 6.3.5 Acoustic parameters identifying vocal emotions..... | 151 |
| 6.4 General conclusions | 154 |
| 6.4.1 The Cultural Proximity hypothesis | 154 |
| 6.4.2 The Prosodic Proximity hypothesis | 155 |
| 6.4.3 Acoustic parameters of vocal emotions | 155 |
| 6.4.4 Dimensionality of vocal emotions | 155 |
| 6.5 Limitations..... | 156 |
| 6.5.1 Phonological legitimacy..... | 156 |
| 6.5.2 The eight emotions | 157 |
| 6.5.3 The neutral category | 158 |
| 6.5.4 Acoustic correlates of emotional intensity..... | 159 |
| 6.5.5 The impact of stimulus order on recognition accuracy | 160 |
| 6.5.6 The ecological validity of stimuli | 160 |
| 6.6 Future research..... | 161 |
| 6.6.1 Second language acquisition..... | 161 |
| 6.6.2 Acoustic manipulations of stimuli | 161 |
| 6.6.3 Neuroimaging | 162 |
| References..... | 163 |
| Summary | 185 |
| Samenvatting..... | 191 |
| Appendices | 197 |
| Appendix A..... | 197 |
| Appendix B..... | 197 |
| Appendix C | 198 |
| Appendix D..... | 205 |
| Appendix E | 205 |
| Appendix F | 206 |
| Appendix G..... | 206 |
| Appendix H..... | 233 |

| | |
|----------------------------|-----|
| Appendix I | 241 |
| Appendix J | 242 |
| Appendix K | 243 |
| Appendix L | 244 |
| Appendix M | 245 |
| Appendix N | 245 |
| Appendix O | 246 |
| Curriculum Vitae | 251 |
| List of publications | 253 |

Acknowledgements

First, I would like to express my deepest gratitude to my two supervisors, Vincent and Roeland, for their exceptional guidance, dedication, and unwavering support throughout my PhD journey. They are knowledgeable, responsible, and reliable mentors, and it has been a great honour and privilege to be their student. Their quick and thoughtful feedback, together with their patience and encouragement, have inspired me to grow both academically and personally. Their continuous encouragement and kindness became a source of light in the dark days.

Second, I would also like to extend my sincere thanks to Professor Claartje Levelt for acting as my promotor. My sincere thanks also go to the members of my PhD committee for their time, effort, and insightful comments that contributed to the completion of this dissertation.

Finally, I am deeply grateful to everyone who has accompanied and supported me along this journey, whose understanding and encouragement have made this work possible.

Nijmegen, November 2025

Chapter One

General Introduction¹

1.1 Introduction²

Emotions are subjective personal feelings (Ekman, 1992a), which can be expressed either verbally (e.g., words, sentences) or non-verbally (e.g., facial expressions, gestures, prosody). Understanding others' emotions is crucial for effective daily communication and social interactions (Jensen, 2014; Jensen & Pedersen, 2016). Since the publication of Charles Darwin's seminal work *The Expression of the Emotions in Man and Animals* (1872; reprint in 1998), the topic has gained widespread attention in fields like linguistics, biology, psychology, neuroscience, etc.

A controversial issue in human emotion recognition is whether it is universal or culture- and/or language-specific. As a pioneer in affective science, Charles Darwin argued that the production and perception of emotions are biologically determined and universal, inherited through the human genome (1872; reprint in 1998). In contradistinction to this, Harre's (1986) social constructivist theory asserts that emotions are shaped exclusively by culture and language. More recently, Elfenbein and Ambady (2002b) proposed their dialect theory, such that while emotion recognition is universal in principle, it becomes less accurate across cultures due to "nonverbal accents", or cultural differences in expression. To date, there is an increasing consensus that cross-cultural emotion recognition results from an interplay between universal, cultural, and linguistic factors (Elfenbein, 2013; Elfenbein, Mandal et al., 2002; Mesquita & Frijda, 1992).

¹ Chapters 2 to 5 have been written as independent articles, including introduction, methodology, and conclusion sections. Therefore, overlaps between these parts are unavoidable.

² With gratitude to my co-authors, parts of this chapter were based on Liang et al. (2025).

How can emotion be defined? Although the concept of emotion is complex and multifaceted, it can be defined in a way that offers nuances for understanding. Emotion is a subjective personal feeling arising from events or affairs, causing physical and mental changes (Izard, 2010; Widen & Russell, 2010). According to Scherer (2009), emotion is a dynamic process resulting from individuals' evaluation of important affairs depending on their needs, goals, and values. Based on Scherer's Component Process Model (CPM) of emotion (Scherer, 2001), emotion includes five interrelated components—appraisal processes, autonomic physiology, action tendencies, motor expression, and subjective feeling. The terms *emotion* and *affect* are sometimes used interchangeably to describe feelings, but they are not identical. Emotions are multifaceted feelings that combine different aspects such as psychology, cognition, and behavior. On the other hand, affect is a broader term that includes, but is not limited to, emotions and moods (Shuman & Scherer, 2014).

How does the brain process emotions? Emotions are processed through interactive neural networks in different brain regions (e.g., hippocampus, hypothalamus, and thalamus), known as the Papez circuit, highlighting that emotion processing is not an isolated activity (Papez, 1937). Further to this, Panksepp (1998) added two additional regions—the amygdala and the prefrontal cortex, which are involved in the processing of emotions. Also, other neural networks, like the limbic system and prefrontal regions, are involved in the dynamic processing of emotions (Celeghin et al., 2017). Currently, research on affective neuroscience has employed sophisticated neuroimaging technologies, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and positron emission tomography (PET) (Cimino, 2002; Dehghani et al., 2023; Gu et al., 2019; Lim et al., 2024; Palomero-Gallagher & Amunts, 2022). For example, using fMRI, Gu et al. (2019) discovered that basic emotions are active in different yet overlapping regions.

How important are emotions? Emotions are intricately related to perception, memory, and decision-making (Palomero-Gallagher & Amunts, 2022; Turnbull & Salas, 2021). Turnbull and Salas (2021) demonstrate that emotions affect cognitive processing, such that positive emotions enhance working memory and facilitate problem-solving skills, whereas negative emotions interfere with working efficiency. Meanwhile, emotions can be regulated by connectivity-based neurofeedback (Dehghani et al., 2023). Findings from affective neuroscience have practical implications for mental health, psychology, and human-computer interaction (HCI) (Hudlicka, 2008;

Jungilligens et al., 2022; Okon-Singer et al., 2015; Renna et al., 2017; Rolls, 1990).

This dissertation is about the way emotions are expressed in speech and whether these emotions are being recognized. The languages involved are Dutch and Korean, which are typologically different in terms of culture and language. The overall question is whether these languages express emotions in a universal way or not, and whether listeners recognize the intended expression of emotions. We will come back to this question after having discussed the main parts that play a role in answering this overall research question.

When emotions in speech can be recognized by listeners, what cues do they use? Acoustic cues, such as pitch, amplitude, spectral distribution, duration, and laryngeal setting, are pivotal (Banse & Scherer, 1996). By acoustically extracting the above parameters, this dissertation aims to identify both universal and language-specific cues in emotional speech recognition, providing a comprehensive understanding of the production and perception of vocal emotions across Dutch and Korean.

1.1.1 Theories of emotion

Emotions can be examined from either a discrete approach (Ekman, 1992b; Izard, 1977), a dimensional approach (Russell, 1980), or an integration of both (Laukka, 2004).

1.1.1.1 Discrete (basic) emotion theory

The discrete emotion theory proposes that different emotions are characterized by specific physiological and behavioral features (Ekman, 1992a, b; Izard, 1992). Thus, this theory identifies a small set of emotions, referred to as basic emotions (Ekman, 1992b). Basic emotions are biologically conditioned and exhibit specific acoustic patterns that can be recognized above chance levels across cultures (Laukka et al., 2013; Laukka & Elfenbein, 2021).

1.1.1.2 Dimensional emotion theory

In contrast, the dimensional theory classifies emotions based on fundamental dimensions such as arousal, valence, intensity, and potency, providing a complex framework for understanding emotions (Russell, 1980; Russell & Barrett, 1999). According to the circumplex model, emotions are primarily defined by the first two dimensions mentioned—arousal and valence (Russell,

1980). Arousal (high-arousal vs. low-arousal) is the physiological change experienced by the speaker (Russell & Barrett, 1999), while valence determines the emotion's positive (pleasant) or negative (unpleasant) nature. For instance, anger is typically a high-arousal and negative emotion, whereas tenderness is regarded as a low-arousal and positive emotion. To understand how arousal and valence interact, it is essential to analyze these two dimensions together. However, two dimensions are insufficient to fully describe emotional nuances (Larsen & Diener, 1992). To describe the subtle differences between emotions from the same family, it is necessary to include other dimensions, such as intensity (Brehm, 1999) and potency (Russell & Mehrabian, 1977). Intensity distinguishes variations between emotions from the same type of emotion (e.g., hot anger vs. cold anger), underscoring the strength and magnitude of an emotion (Bänziger & Scherer, 2005; Brehm, 1999; Larsen & Diener, 1987; Sonnemans & Frijda, 1994; Wright et al., 1983). Potency refers to the cognitive appraisal of a person's power or influence over a situation (Lazarus & Smith, 1988). Integrating these dimensions offers a more comprehensive understanding of the complex nature of emotional states.

We assigned the distinction between basic and non-basic emotions to the dimensional approaches, and called the dimension "basicness". We did so, as basicness classifies emotions in two general subsets, like arousal and valence.

1.1.1.3 Comparing the discrete and dimensional approaches

The discrete and dimensional approaches are not mutually exclusive. Instead, they describe emotions from different perspectives, enhancing our understanding by emphasizing different facets of emotional experience. On the one hand, the discrete emotion approach categorizes emotions as distinct and individual states such as anger, fear, and happiness. It posits that basic emotions are biologically inherited and have unique acoustic patterns, which serve specific adaptive functions and can be reliably differentiated from one another. On the other hand, the dimensional approach regards emotions as varying dimensions, such as arousal and valence (Russell, 1980). Additional dimensions have been proposed, such as intensity (Brehm, 1999) and potency (Russell & Mehrabian, 1977), to handle more detailed nuances of emotions. The integration of both approaches provides a comprehensive framework for a more holistic understanding of emotions.

1.1.2 Empirical studies on cross-cultural emotion recognition

Over the past few decades, numerous studies have investigated emotion recognition across different cultures (Elfenbein & Ambady, 2002b; Juslin &

Laukka, 2003; Pell, Monetta et al., 2009; Scherer et al., 2001). These studies have mostly adopted experimental designs using either a “one-to-many” approach—presenting stimuli recorded by a single group of speakers to several groups of listeners (Beier & Zautra, 1972; Scherer et al., 2001; Van Bezooijen et al., 1983); or a “many-to-one” approach, presenting stimuli recorded by several groups of speakers to a single group of listeners (Chronaki et al., 2018; Kramer, 1964; Pell, Monetta et al., 2009; Thompson & Balkwill, 2006). Additionally, some studies have used a fully balanced design, presenting stimuli from two or more groups of speakers to the same number of listener groups (Albas et al., 1976; Jiang et al., 2015; Paulmann & Uskul, 2014; Sauter et al., 2010).

Taken together, earlier studies aimed to determine to what extent vocal emotion recognition is universal or culture-/language-specific, and have concluded that emotions are decoded above chance cross-culturally, in line with the universality hypothesis (Elfenbein, 2013; Elfenbein & Ambady, 2002b; Scherer et al., 2001). Furthermore, previous studies reveal an in-group advantage, such that individuals recognize emotions produced in their native language more accurately than those in an unknown language, indicating the existence of language-specific prosodic cues in vocal emotion expressions (Pell, Paulmann et al., 2009). However, prior research in this area has mostly employed unbalanced experimental designs and predominantly focused on basic emotions (Ekman, 1992b). Consequently, empirical research on cross-cultural emotion recognition by listeners and speakers from typologically different languages and cultures remains scarce.

Moreover, most studies on cross-cultural emotion recognition have relied on acted rather than spontaneous speech due to the challenges in controlling the verbal content in spontaneous speech (Jiang et al., 2015; Paulmann & Uskul, 2014; Pell, Monetta et al., 2009; Thompson & Balkwill, 2006; Van Bezooijen et al., 1983), with a few exceptions employing spontaneous speech (Chung, 1999). Additionally, most studies have used pseudo-utterances to eliminate semantic cues that might affect emotion recognition.

Another relevant field is research on emotion classification in speech on the basis of specific acoustic parameters, and classifying vocal emotions via a constellation of acoustic parameters. In affective computing, frequently used machine learning models include Support Vector Machine (SVM), LDA (Linear Discriminant Analysis), Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Convolutional Neural Network (CNN) (Ezhilarasi & Minu, 2012; Lee & Narayanan, 2005; Luengo et al., 2005; Pallewela et al., 2024; see Ververidis & Kotropoulos, 2006 for a review).

Classification rates vary depending on a number of factors, such as the selection of acoustic features and the size of the corpora. To address the limitations of each model, Ezhilarasi and Minu (2012) proposed hybrid systems that integrate both SVM models and deep learning neural networks (DNNs) to improve classification accuracy. Furthermore, Laukka et al. (2011) suggested that classification rates can be improved by combining speech with facial expressions and physiological signals. However, due to the limited number of corpora, current findings of automatic speech recognition may not be generalizable to different languages, cultures, and vocal emotions.

This dissertation employs a balanced two-by-two design, with speakers and listeners from different languages and cultures—Dutch and Korean, aiming to better examine the in-group advantage in a cross-cultural setting. Balanced design is especially important in evaluating the in-group advantage. However, due to the difficulty of recording acted speech, there are limited corpora of vocal emotion expressions. For example, the VENEC corpus is a large database of vocal emotion expressions (Laukka et al., 2010). It includes 19 different emotions and a total of 6,500 stimuli, recorded by 100 professional voice actors from five countries.

1.1.3 Research on affective neuroscience

Emotions are intricately related to perception, memory, and decision-making (Palomero-Gallagher & Amunts, 2022; Turnbull & Salas, 2021). Turnbull and Salas (2021) demonstrate that emotions affect cognitive processing, such that positive emotions enhance working memory and facilitate problem-solving skills, whereas negative emotions interfere with working efficiency. Meanwhile, emotions can be influenced by connectivity-based neurofeedback (Dehghani et al., 2023). Findings from affective neuroscience have practical implications for mental health, psychology, and human-computer interaction (HCI) (Hudlicka, 2008; Jungilligens et al., 2022; Okon-Singer et al., 2015; Renna et al., 2017; Rolls, 1990).

1.2 Research methodology

I used the stimuli from the Demo (Dutch emotion)/Koremo (Korean emotion) corpus (Broersma et al., 2025).³ This corpus was specifically developed for cross-linguistic comparison and is more balanced than the existing corpora in several respects. First, the two sub-corpora contain a comparatively large

³ The scenarios and corpus are publicly available via Radboud University at <https://doi.org/10.34973/5kg3-9852>

number of emotions (eight emotions) which were balanced in arousal (high-arousal vs. low-arousal) and valence (positive vs. negative), and with an equal number of basic and non-basic emotions (see Table 1.1). Second, the eight emotions were expressed by a large number of actors from two typologically different languages (eight Dutch and Korean actors), with the same number of females and males in each language group, accounting for 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 \times 8 actors \times 2 tokens \times 2 languages). Third, the pseudo-sentence /nuto hɔm sɛpikɑŋ/ is phonologically compatible in both Dutch and Korean.⁴ Using a pseudo-sentence can eliminate verbal semantic processing. Further, since vowels are considered to carry more affective meanings than consonants (Majid, 2012), listeners are more affected by vowel duration than consonant duration. In this study, the pseudo-sentence /nuto hɔm sɛpikɑŋ/ was created with a roughly equal number of vowels (/u/, /o/, /ɔ/, /ɛ/, /i/, /ɑ/) and consonants (/n/, /t/, /h/, /m/, /s/, /p/, /k/, /ŋ/). Fourth, the same elicitation technique (the Stanislavski technique) was used by both the Korean and Dutch stage directors. The elicitation methods and recording procedures were the same in both languages. For more details of the elicitation and recording procedure, refer to Chapter 2.

Table 1.1. The eight emotions included in this project in a valence-by-arousal grid (reproduced from Goudbeek & Broersma, 2010b, p. 2212); basic emotions are indicated by *.

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

⁴ According to Goudbeek and Broersma (2010b), the pseudo-sentence /nuto hɔm sɛpikɑŋ/ is phonologically legal in both Dutch and Korean. However, the low-mid vowel [ɔ] does not exist in Korean (Shin, 2015) and was pronounced as high-mid vowel [o].

1.3 The current study

Emotions play a pivotal role in human communication, especially in a cross-cultural setting (Jensen, 2014; Trampe et al., 2015). Although emotions can be recognized above chance across cultures, individuals recognize emotions more accurately when they are expressed by members from the same or similar cultural/linguistic group than by members from a typologically different group. This is referred to as the in-group advantage. However, although research in this field has produced fruitful insights, there remain significant gaps that prevent us from fully understanding how vocal emotions are affected by cultural and linguistic factors. First, most previous studies have either used a “one-to-many” approach, presenting stimuli recorded by a single group of speakers to several groups of listeners (Beier & Zautra, 1972; Scherer et al., 2001; Van Bezooijen et al., 1983); or a “many-to-one” approach, presenting stimuli recorded by several groups of speakers to a single group of listeners (Chronaki et al., 2018; Kramer, 1964; Pell, Monetta et al., 2009; Thompson & Balkwill, 2006). Consequently, it results in an unbalanced design that renders it difficult to test the in-group advantage. Second, most prior research has focused on basic emotions (Cordaro et al., 2016; Laukka et al., 2016). Therefore, current knowledge on emotions cannot be generalized to non-basic emotions. Moreover, the limited number of emotions leads to unbalanced designs in terms of arousal and valence. Third, most studies investigate emotions from a discrete approach (Ekman & Friesen, 1969; Ekman et al., 1987; Laukka et al., 2013), although a limited number of studies explore emotions from a dimensional approach (Barrett, 1998; Laukka et al., 2005; Mozziconacci, 2002; Russell, 1980). Thus, subtle differences between emotions remain unclear, particularly when emotions share similar features. To bridge the gap, I employ the Demo/Koremo corpus (Broersma et al., 2025) using the “two-by-two” design with listeners and speakers from typologically different cultures and languages—Dutch and Korean, and includes a relatively large number of emotions balanced for arousal and valence.

Specifically, Dutch is a stress-accent language with binary trochees, which has a rather restricted pitch range (Gussenhoven, 1993). In Dutch, word stress is employed to differentiate between identical segment strings, for example, *KAnon* /'kanɔn/ “list of saints” versus *kaNON* /ka'nɔn/ “large gun”. Dutch utterances employ two prosodic units above the word level: Intonational Phrase (IP) and Phonological Phrase (PP) (Gussenhoven, 2005). In contrast, Korean does not have minimal stress pairs. Although there are controversies regarding the rhythm classification of Korean, most studies tend to classify it as a syllable-timed language (Arvaniti, 2012). Korean utterances are divided

into two prosodic units above the word level, namely Intonational Phrase (IP) and—different than Dutch—the Accentual Phrase (AP) (Jun, 2005).

The overarching goal of this study is to investigate cross-cultural vocal emotion recognition from both the discrete and the dimensional approaches, focusing on the influence of culture-/language-specific factors, acoustic cues, and emotional dimensions (including also emotional intensity), and basicness on recognition accuracy. To address this broad issue, I conducted four studies targeting several sub-questions. Collectively, the results of these four studies contribute to a better understanding of the perception of vocal emotions in a cross-cultural setting.

1.4 Two approaches

The chapters in this dissertation adopt two approaches—a discrete and a dimensional one. The discrete approach targets the categorical perception of separate emotions, which may hinge on subtle differences between discrete emotions (Chapters 3 and 4). The dimensional approach, as implemented in the present dissertation, makes a three-way overall dimensional characterization of emotions based on arousal (high-arousal vs. low-arousal), valence (positive vs. negative), and basicness (basic vs. non-basic) (Chapters 2 and 5). The discrete approach presents a more precise and perhaps more subtle understanding of the distinctions between emotions, while the dimensional approach aims to distinguish overall underlying properties. For instance, both anger and irritation are negative emotions, whereas they differ in terms of arousal. Anger is a high-arousal emotion with intense energy, while irritation is a low-arousal emotion with mild energy (Spielberger et al., 1995). But do the dimensions cover all discrete emotions adequately? Integrating these two approaches might provide a comprehensive framework to study the complexity of emotions. Therefore, in Chapter 5, we combine these two approaches by analyzing confusion matrices based on the eight emotions, illustrating emotions that are easily misclassified.

1.5 Overview of the dissertation

The rest of this dissertation consists of five chapters, each addressing various aspects of vocal emotion recognition. The final chapter (Chapter 6) summarizes the research chapters 2 to 5 with a discussion and an integration of the results, highlights the dissertation's significant contributions, and, of

course, addresses questions about the limitations of our empirical studies and provides suggestions for future research.

Chapter 2 “Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design” investigates recognition of vocal emotions by listeners from two different cultures and languages, i.e., Dutch and Korean, and examines the so-called in-group advantage in vocal emotion recognition. The in-group advantage hypothesis predicts that listeners recognize vocal emotions produced in their native language more accurately than when expressed in an unknown language. Regardless of the applicability of the in-group advantage, we predict that listeners recognize vocal emotions produced in their native language and in the unknown language above chance, which would show that the expression and perception of vocal emotion has at least a universal component. Finally, the chapter examines the influence of arousal, valence, and basicness on vocal emotion recognition, within and across cultures. As explained above in § 1.2, these are three dimensions that are part of the dimensional approach to emotion production and perception. In our study, we dichotomize the eight target emotions into subsets of four, i.e., high- vs. low-arousal, positive vs. negative valence, and basic vs. non-basic. We will examine whether some subsets (quadruplets) are easier to recognize cross-culturally than others. For instance, basic emotions may be easier to recognize, both within and across cultural divides, than non-basic emotions. We also predict that confusions in the recognition of vocal emotions will be more frequent within than across dimensional quadruplets. In this chapter, the focus is limited to only the accuracy of the emotion identification, and the similarity structure of the emotions is not investigated. This similarity structure will be examined in a later chapter, where the comparison between machine and human identification of emotions and confusion matrices will be presented.

Chapter 3 “Interpreting the intensity of vocal emotions across cultures” analyzes the intensity ratings by Dutch and Korean listeners, as collected in Study 1. The starting point for this chapter is that Intensity should be added (and studied in more detail) as a separate dimension of (vocal) emotions to capture emotional states that cannot be adequately described by the traditional dimensions of Arousal, Valence, and Potency. Emotions are always expressed with different levels of intensity (Mesquita & Frijda, 1992). Intensity refers to the strength of emotions perceived by receivers, and people tend to respond to emotions with higher intensity than those with lower intensity. Moreover, 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 known as the in-group bias

(Kommattam et al., 2019). This chapter examines whether there exists an in-group bias in intensity ratings across accurate and inaccurate trials. Finally, as in the preceding chapter, we examine the effect of arousal, valence, and basicness on intensity ratings. I will do this first for all responses, and then repeat the analysis for the subset of correctly identified emotions (which should have higher intensity ratings).

Chapter 4 “Classifying emotions from acoustic parameters” examines the role of a large number of acoustic cues in recognition accuracy, focusing on the influence of emotion, speaker language, and gender on recognition accuracy. Since we are interested in the vocal (rather than verbal) expression of emotion, this chapter will target the effects of prosodic parameters only. These are properties of human speech that cannot be tied down to specific individuals' speech sounds (phonemes) but are characteristic of larger speech units, such as syllables, phrases, clauses, sentences, and even paragraphs. Specifically, we will examine the role of vocal pitch, acoustic intensity (loudness), articulatory setting (vocal timbre, as conveyed by formants and spectral tilt), and harmonicity (noisiness of the voice). I examine whether recognition accuracy can be reliably predicted by a constellation of acoustic parameters, and compare the recognition accuracy between machine classifiers and human listeners.

This chapter is not only about the effects of acoustic parameters on recognition accuracy but also, and more crucially, on the (cross-cultural) confusion of emotions. If an emotion is not correctly identified, then what is it mistaken for? This opens a window on cross-cultural misunderstanding in the signaling of emotions. So the real research question is not about the accuracy of the emotion perception per se, but on the effects of acoustic parameters on the identification of emotions (whether correct or confused).

Chapter 5 “Recognizing vocal emotions in unfamiliar languages” focuses on cross-cultural vocal emotion recognition by American English and French listeners who had no knowledge of Dutch or Korean. This chapter investigates the relative contributions of the Universality hypothesis, Cultural Proximity, Linguistic Proximity, and emotional dimensions to emotion recognition. According to the Universality hypothesis, some emotions are universally recognized by people across cultures and languages. People from a similar cultural background can recognize emotions more accurately than those from a different one (Elfenbein & Ambady, 2003a). In the vocal domain, listeners find it much easier to identify emotions expressed in a language typologically similar to their native language than to a different one. Furthermore, emotional dimensions, such as arousal, valence, and basicness, affect emotion

recognition. However, it remains unknown how these factors affect recognition accuracy of emotions. Therefore, I aim to study to what extent universal, cultural, linguistic, and emotional dimensions affect the perception of vocal emotions. To achieve this goal, I selected American English listeners and French listeners, since English is a stress-timed language, which is prosodically/rhythmically close to Dutch; French is a syllable-timed language, which is prosodically/rhythmically similar to Korean. By comparing the recognition accuracy between these two groups, we can find out how the above factors affect the perception of vocal emotions. First, I tested whether both listener groups recognized vocal emotions above chance. Second, I tested recognition accuracy in Dutch recordings by both groups of listeners. Third, I tested whether French listeners outperform American English listeners in Korean recordings, since French and Korean are syllable-time languages, which share similar prosodic/rhythmic patterns. Finally, I examined the role of arousal, valence, and basicness in vocal emotion recognition.

To answer the above questions, I conducted three perception experiments and one acoustic analysis of stimulus materials. In the first experiment, Dutch and Korean listeners were asked to listen to each stimulus and identify the emotion it conveyed by choosing one of the eight emotions listed on screen. In the second experiment (using the same moment of data collection), the listeners were asked to estimate the intensity of the emotion as experienced/expressed by the speaker. The third experiment tested the perception of vocal emotions by American English and French listeners. Similar to the first study, these two groups of listeners were asked to listen to the same corpus and select the emotion they thought the stimulus expressed and estimate the intensity of the emotion as expressed by the speaker. Fourth, I acoustically analyzed each of the 256 portrayals based on 17 acoustic parameters, and further examined the correlations between acoustic parameters and recognition accuracy. Moreover, I compared recognition accuracy predicted by machine and human listeners.

Chapters 2 to 5 address the following four main research questions:

Chapter 2: Do Dutch and Korean listeners recognize vocal emotions above chance in Dutch and Korean, and is there an in-group advantage in vocal emotion recognition?

Chapter 3: Is there an in-group bias in intensity ratings of Dutch and Korean vocal emotions by Dutch and Korean listeners?

Chapter 4: How do acoustic parameters of vocal emotions vary across emotions, speaker language, and gender in Dutch and Korean?

Chapter 5: Is cross-cultural/language vocal emotion recognition in unfamiliar languages affected by Universality, Cultural Proximity, Prosodic Proximity, and emotional dimensions?

Chapter Two

Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design⁵

Abstract

This study investigates cross-cultural vocal emotion recognition in a corpus with an affectively and linguistically balanced design. It has two main goals. First, it aims to explore the recognition of emotions in two typologically different languages, Dutch and Korean, within and across cultures. Second, it aims to contribute to the methodological development of the study of cross-cultural vocal emotion recognition by presenting a new corpus for Dutch and Korean emotional speech (the Demo/Koremo corpus), containing portrayals of eight emotions differing in arousal, valence, and basicness (joy, pride, tenderness, relief, anger, fear, sadness, irritation) produced by Dutch and Korean actors (communicated in a single phrase which was viable in both languages). Dutch and Korean participants listened to recordings of all emotions produced by the Dutch and Korean actors and indicated which emotion they thought it expressed. Both groups of listeners recognized emotions significantly above chance in both languages, but more accurately in their native language, in line with the dialect theory of emotion (Elfenbein & Ambady, 2002b; Elfenbein et al., 2007). In addition, we found that low-arousal emotions, negative emotions, and basic emotions were recognized more accurately than their counterparts, both within and across cultures. While some of these results replicate earlier findings, others—the effect of arousal, and the within-cultural effects of valence and basicness—had not been previously investigated. This study provides new insights into cross-cultural vocal emotion recognition and contributes to the methodological toolkit of intercultural emotion recognition research.

Keywords: Dutch, Korean, cross-cultural emotion recognition, speech, in-group advantage

⁵ This chapter is an edited version of Liang et al. (2025).

2.1 Introduction

The ability to understand other people's emotions plays an important role in our daily communication and social interactions (Jensen, 2014). The study of human emotions has a long history: Charles Darwin already proposed that the production and perception of emotions are innate and universal, and that they developed through evolution (1872, republished in 1998).

Since then, emotions have been the topic of many studies, and a much-debated issue is whether emotion recognition is universal or culture- and language-specific. In a seminal study, Ekman et al. (1969) showed that there were striking similarities in the way that individuals from unrelated, vastly different cultures expressed emotions with their facial expressions and recognized these emotions in others. This work cemented the idea that some emotions (originally: anger, fear, happiness, sadness, disgust, and surprise), which they termed "basic emotions", were universal. Numerous studies on facial expressions have replicated this finding, confirming that many emotions can be accurately recognized across cultures (for a meta-analysis, see Elfenbein & Ambady, 2002b). There is also, however, strong evidence that culture and language also play a role in the way humans learn to express and understand emotions, in accordance with Harre's (1986) social constructivist theory of emotions (see also Barrett & Russell, 2014). In an attempt to account for the findings that emotion recognition is to some extent culture-specific, Elfenbein and Ambady (2002b) proposed the dialect theory of emotion that recognition of emotions is to some extent universal, but more accurate within than across cultures. According to this theory, culture-dependent and/or language-dependent factors serve as a "dialect" in cross-cultural emotion recognition. To date, there is a consensus that visual cross-cultural emotion recognition is influenced by both universal and cultural-linguistic factors (Elfenbein, 2013; Elfenbein & Ambady, 2002b; see also Keltner et al., 2019 for a review). The overarching goal of this paper is to test the main tenets of this theory in the domain of vocal emotion recognition in two typologically different languages using a new corpus of vocal emotion stimuli.

2.1.1 Cross-cultural vocal emotion recognition

Whereas research on the expression and interpretation of human emotions started with facial expressions, emotions can be expressed in many other ways, including the semantic content of spoken utterances, paralinguistic characteristics of these utterances like prosody, and bodily signals such as gestures and postures (Mehrabian, 2017; Scherer, 2003; 2019). The vocal expression of emotion has, more recently, become a lively topic of research (Juslin &

Laukka, 2003; Paulmann & Uskul, 2014; Pell, Monetta et al., 2009). Studies of vocal emotion expressions have focused on two main types of utterances: they have either used non-linguistic vocalizations like laughs, growls, and sighs, or linguistic vocalizations like non-words, words, or phrases. A recent meta-analysis of 37 studies of cross-cultural vocal emotion recognition (Laukka & Elfenbein, 2021) showed that emotions expressed both in non-linguistic (Cordaro et al., 2016; Laukka et al., 2013; Sauter et al., 2010; Sauter & Scott, 2007) and linguistic (Juslin & Laukka, 2003; Laukka et al., 2016; Paulmann & Uskul, 2014; Pell, Monetta et al., 2009) vocalizations can be recognized cross-culturally at above-chance level.

At the same time, listeners more accurately recognize emotions expressed by members from the same cultural/linguistic group than by members from another group; they exhibit an *in-group advantage*, similar to the one shown in visual emotion recognition (Laukka & Elfenbein, 2021). So, vocal emotion recognition is—like facial emotion recognition—a product of both universal principles and language-specific factors (Mesquita & Frijda, 1992). Importantly, most of these findings are based on a categorical conceptualization of emotions. However, emotions can also be understood as varying between two dimensions, arousal and valence (Laukka et al., 2005; Russell, 2003; Scherer, 2009). Arousal (or excitement) refers to the intensity with which an emotion is experienced (although the exact nature and definition of arousal are under debate; see Russell, 2003). A person's level of arousal has been shown to exert an influence on their decision-making and judgment, including judgments of the emotions of others, visual processing of pictures, and time perception (Clark et al., 1984; Lane et al., 1999; Mourão-Miranda et al., 2003; Smith et al., 2011). For example, increases in a perceiver's level of positive or negative arousal have been shown to increase the likelihood that they interpret phrases and facial expressions as being high in arousal too, but only for positive emotions (Clark et al., 1984). Arousal also affects the vocal characteristics of speech, as high-arousal emotions are often produced with higher intensity, higher pitch, longer durations, and wider pitch ranges than low-arousal emotions (Breitenstein et al., 2001); arousal, in fact, influences speech more than valence (or the dimension of potency/control; Goudbeek & Scherer, 2010), and listeners can recognize if vocal emotions are high or low in arousal (Laukka et al., 2005). However, little is known about the ease with which listeners recognize low-arousal emotions compared to high-arousal emotions both within and across cultures.

Like arousal, valence, with the poles positive vs. negative, or pleasant vs. unpleasant, plays an important role in emotion recognition (Russell, 1994). Studies have shown that a number of positive and negative emotions can be

identified in vocal signals (Cowen et al., 2019; Laukka & Elfenbein, 2021), and that recognition accuracy is higher for negative than positive emotions (Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2011). The first study to observe such a trend is Sauter et al. (2010), who investigated recognition of emotional vocalization in European English and Himba listeners. The results revealed that while all the negative emotions that they used in their study could be identified both within and cross-culturally, the cross-cultural recognition of the positive emotions was more variable. In their meta-analysis, Laukka and Elfenbein (2021) confirmed that the cross-cultural recognition of negative emotions is more accurate than that of positive emotions. One possible explanation for this effect is that negative emotions are directly associated with danger and survival, while positive emotions are linked to social bonds and, therefore, are more likely to be shared by members from the same culture (Shiota et al., 2004). However, the impact of valence on emotion recognition within cultures is unclear.

Finally, according to basic emotion theory, there is a small set of emotions that all humans share regardless of their cultural background. These emotions are responses to fixed triggers: they cause a fixed physical and behavioral response (Ekman, 1972, 1992a, b; Ekman et al., 1969; but see Gendron et al., 2018, for a different view on this matter).⁶ Numerous studies have demonstrated that facial expressions of basic emotions can be identified by individuals from different countries (Ekman, 1972; Elfenbein & Ambady, 2002b). Similarly, for non-verbal vocalizations, Sauter et al. (2010) found that basic emotions (anger, disgust, fear, joy, sadness, surprise) were reliably decoded by European English and Himba listeners cross-culturally, whereas vocalizations of non-basic emotions were less accurately recognized cross-culturally. An open question is, however, whether basic emotions are also recognized better than non-basic emotions within cultures.

2.1.2 Methodological considerations

Previous studies on cross-linguistic vocal emotion recognition have used a wide array of methodologies (see Laukka & Elfenbein, 2021, for a review). Methodological choices are likely to impact the outcomes of any study, and in particular in studies aiming to investigate interactions involving groups, such as in-group advantages (Matsumoto, 2002). In this paper, we address the following methodological considerations.

⁶ Gendron et al. (2018) challenged the long-standing *Universality* hypothesis, emphasizing the influence of cultural and contextual factors on the perception of emotions.

2.1.2.1 *Balance in the emotion characteristics*

To be able to disentangle the contribution of individual categorical emotions as well as the dimensions valence and arousal, the emotions included in cross-cultural emotion recognition studies should be carefully chosen to represent the emotion characteristics of interest (such as arousal, valence, and basicness) in a balanced way. Many previous studies have exclusively used basic emotions (Bailey et al., 1998; Bryant & Barrett, 2008; Chronaki et al., 2018; Chung, 1999; Huang et al., 2008; Mandal, 2008; Pell, Monetta et al., 2009; Scherer et al., 2001; Thompson & Balkwill, 2006, notable exceptions being Bänziger et al., 2012; Cowen & Keltner, 2017), while other studies have used several basic emotions and only a few non-basic emotions (e.g., Cordaro et al., 2016; Kramer, 1964; Laukka et al., 2016; Shochi et al., 2009).⁷ As for arousal, most prior research has included more high-arousal than low-arousal emotions (Laukka & Elfenbein, 2021), likely related to the fact that the original set of six basic emotions (Ekman, 2016; Ekman & Cordaro, 2011; Ekman et al., 1969) contains only one low-arousal emotion, i.e., sadness. With respect to valence, previous studies have typically included more negative than positive emotions (see Laukka & Elfenbein, 2021): this may again be due to the fact that the original set of six basic emotions (Ekman, 2016; Ekman et al., 1969; Ekman & Cordaro, 2011) contains only one positive emotion, i.e., happiness. As many studies have, exclusively or predominantly, used basic emotions, this has resulted not only in an overrepresentation of high-arousal and negative emotions, but also in a common confound between basicness, arousal, and valence. Such confounds can be addressed by balancing those variables through the choice of emotions included in a study.

2.1.2.2 *Balance in languages used*

The number and typology of the speaker languages and listener languages included in each study will affect the type of questions that can be addressed; e.g., while for some research questions speakers from one language and listeners from multiple languages might be desirable, other research questions require using speakers and listeners from the same two language backgrounds. Some studies, using what we will call a “one-to-many” approach (e.g., Beier & Zautra, 1972; Scherer et al., 2001; Van Bezooijen, 1984), have presented

⁷ The GEMEP (Geneva Multimodal Emotion Portrayals) Corpus contains 18 emotions, including basic and non-basic emotions. Cowen and Keltner (2017) studied 27 self-reported emotional categories elicited by videos, including a relatively large number of basic and non-basic emotions.

stimuli recorded by a single group of speakers to several groups of listeners.⁸ For instance, Scherer et al. (2001) presented stimuli expressing five emotions produced in German to listeners from nine different countries. Other studies have used a “many-to-one” approach, presenting stimuli recorded by several groups of speakers to a single group of listeners (e.g., Chronaki et al., 2018; Kramer, 1964; Pell, Monetta et al., 2009; Thompson & Balkwill, 2006). For example, Thompson and Balkwill (2006) presented English listeners with four basic emotions produced in English, German, Tagalog, Japanese, and Chinese. Finally, others have used a fully crossed design, henceforth referred to as “two-to-two” and “many-to-many” approaches, using speakers and listeners from two or more groups, such that each group of listeners is presented with stimuli from their own language as well as the other language(s) (e.g., Albas et al., 1976; Jiang et al., 2015; Paulmann & Uskul, 2014; Sauter et al., 2010). When interactions between speaker and listener languages are the main interest of a study, fully crossed designs (e.g., “many-to-many” designs) provide more information than other designs.⁹ For example, Paulmann and Uskul (2014) crucially needed a “two-by-two” design, with English and Chinese speakers and listeners, to be able to confirm that there was an in-group advantage in vocal emotion recognition for monolinguals as well as bilinguals in these groups. Laukka et al. (2016) needed a square “many-to-many” design, involving native English speakers and listeners from five different countries (America, Australia, India, Kenya, Singapore) to test the dialect theory of emotion.

In addition to the number of speaker and listener languages included in each study, the typological distance between the languages or variants involved should also be chosen to serve the purpose of the study, as illustrated by Paulmann and Uskul’s (2014) use of two typologically unrelated languages, and Laukka et al.’s (2016) use of different varieties of the same language.

⁸ Note that we follow the terminology used by Goudbeek & Broersma (2010b), whereas Laukka & Elfenbein (2021) refer to the “one-to-many” approach as the “many-on-one” approach, and to the “many-to-one” approach as the “one-on-many” approach.

⁹ A fully crossed design ensures that every speaker language is evaluated by listeners from every listener language. Notably, only square many-to-many designs ($n \times n$ matrix), where the sets of speaker and listener languages are the same, provide complete information regarding these interactions. However, rectangular designs ($n \times m$ matrix), where the number of listener languages is unequal to the number of speaker languages, are also referred to as “many-to-many” but may either miss potential interactions (if $m < n$) or introduce redundancy (if $m > n$).

2.1.2.3 *Similarity of stimuli across languages*

If stimuli are produced in more than one language, the stimuli should be phonologically as similar as possible in those languages, as also proposed by Matsumoto (2002). Traditionally, when cross-cultural emotion studies used linguistic materials produced in two or more languages, the materials differed across those languages, which is unavoidable if the materials involve existing words or phrases from these languages. This, however, introduces two problems. First, if the stimuli are phonologically incompatible with the native language of one or more listener groups (e.g., because they contain speech sounds or combinations of speech sounds that do not occur in that language), this might affect the processing of emotional information. In other words, it creates a confound between cross-cultural effects and linguistic incompatibility. Second, it is conceivable that some sounds carry more affective meaning than others (e.g., vowels versus consonants; Majid, 2012), such that using different materials across languages entails the risk of further confounds.

Such confounds can be avoided, or at least reduced, by using pseudo-words and pseudo-phrases. Nonsense stimuli have the advantage that semantic cues to emotions are avoided, and that the linguistic form can be chosen to be phonologically compatible not only with the speakers' languages but also with the listeners' languages (containing phonemes that occur in all languages involved, and with a phonological structure that is phonologically legal in all those languages).¹⁰

2.1.2.4 *Acted versus spontaneous speech*

Speech materials consist of either acted or spontaneous speech. The most important advantage of acted speech is the opportunity to control relevant aspects of the stimuli, as listed below, while the most important advantage of spontaneous speech is its greater ecological validity. First, in acted speech, the verbal content of the utterances can be controlled, whereas in spontaneous speech it cannot, thus potentially providing information about the emotional state of the speaker. Second, in acted speech, high-quality recordings without

¹⁰ There will always be phonetic differences between the realizations of the same phoneme in different languages. Such differences may show up in a narrow phonetic (but not in a broad phonemic) transcription of the utterances. Slightly deviant realizations of a phoneme will be perceived as non-typical (but identifiable) instantiations of their category—as explained by Best's (1995) Perceptual Assimilation Model (PAM).

background noise can be produced in the laboratory, unlike in spontaneous speech. Third, acted speech can express (or at least aim at the expression of) one emotion per utterance, whereas there might be more than one dominant emotion per utterance in spontaneous speech. While some studies on vocal emotion recognition have used spontaneous speech (Chung, 1999; Jürgens et al., 2013), due to the difficulty of using spontaneous utterances for experimental purposes, most studies have used acted speech instead, typically using pseudo-utterances to avoid semantic cues (Jiang et al., 2015; Paulmann & Uskul, 2014; Pell, Monetta et al., 2009; Thompson & Balkwill, 2006; Van Bezooijen, 1984; Zhu, 2013).

2.1.2.5 Statistical methods capturing all relevant factors

Statistical methods should enable investigating multiple variables of interest in the same analysis, while at the same time accounting for by-participant and by-item variability. Previous studies on cross-cultural emotion recognition have mainly relied on analysis of variance or related techniques (Van Bezooijen, 1984; Scherer et al., 2001), but mixed effects modeling provides a more powerful statistical tool for data analysis involving estimation of and generalization over both fixed and random effects (Barr et al., 2013; Bates et al., 2015). Recent emotion recognition studies, e.g., Jiang et al. (2015), have already started employing these methods.

2.1.3 The present study

This paper has two main goals. First, it aims to contribute to the methodological development of the study of cross-cultural vocal emotion recognition by employing the Demo/Koremo corpus for Dutch and Korean emotional speech (Broersma et al., 2025), adopting a “two-to-two” approach. Second, it aims to explore the recognition of emotions in Dutch and Korean (the latter being a language that is relatively underrepresented in affective science) with affectively and linguistically balanced materials within and across cultures.¹¹

Our first theoretical research aim concerns the recognition of emotions within and across cultures. Based on previous findings from dialect theory (Juslin & Laukka, 2003; Pell, Monetta et al., 2009; Scherer et al., 2001), we hypothesize that listeners will be able to recognize vocal emotions not only within but also across cultures above chance level (Hypothesis 1), but that there will be an in-

¹¹ The scenarios and corpus are publicly available via Radboud University at <https://doi.org/10.34973/5kg3-9852>

group advantage (Elfenbein, 2013; Elfenbein & Ambady, 2002b), such that listeners will be better at recognizing emotions from their own language than from the other language (Hypothesis 2).

Our second theoretical research aim concerns the role of the emotional dimensions: *arousal*, *valence*, and *basicness*. While we have no prior expectations about the influence of arousal on emotion recognition, we test the impromptu hypothesis that high-arousal and low-arousal emotions will be recognized differently, both within and across cultures (Hypothesis 3). Further, we predict that negative emotions will be recognized more accurately than positive emotions (Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2011), both within and across cultures (Hypothesis 4), and, finally, we predict that basic emotions will be recognized more accurately than non-basic emotions (Ekman, 1992b, 1999; Elfenbein & Ambady, 2002b), both within and across cultures (Hypothesis 5).

To address these questions, the methodological considerations outlined above lead to the following design choices. First, as we explore the impact of arousal, valence, and basicness on cross-cultural emotion recognition, it is crucial to have emotions balanced, as far as possible, on all these properties. In the current study, there are eight emotions (see Table 2.1), which are balanced in arousal and valence, with two emotions for each of the combinations: high arousal + positive (joy, pride), low arousal + positive (tenderness, relief), high arousal + negative (anger, fear), and low arousal + negative (sadness, irritation). While there is considerable debate over what constitutes a basic emotion (e.g., some scholars argue that basic emotions should include tenderness, love, and empathy (Kalawski, 2010) or pride (Tracy & Robins, 2007)). However, we adopt Ekman's classification of basic emotions (Ekman, 1992b, 1999; Ekman et al., 1969), which limits the set to anger, fear, happiness, sadness, disgust, and surprise. Due to the composition of the set of basic emotions, they cannot be fully crossed with arousal and valence. Instead, we use equal numbers of basic emotions (joy, anger, fear, sadness) and non-basic emotions (pride, tenderness, relief, irritation) in our corpus.

Table 2.1. The eight emotions used in the current study in a valence-by-arousal grid (reproduced from Goudbeek & Broersma, 2010b, p. 2212); basic emotions are marked with “*”.

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

Second, the study includes speakers and listeners from two languages: Dutch and Korean. Dutch and Korean are two typologically very different languages. Dutch is a stress-timed language. Word stress may be used to differentiate the meaning of segmentally identical word forms (Gussenhoven, 1993), like *KAnon* /'kanɔn/ vs. *kaNON* /ka'nɔn/, meaning “list of saints”, and “large gun”, respectively. Pitch contributes to the marking of one type of prosodic unit in Dutch, below the level of the sentence/utterance, namely the Intonational Phrase (IP) (Gussenhoven, 2005). Korean, however, does not have word stress but uses phrasal stress, so that one of the last syllables of a phrase is marked by a pitch change. Although there are controversies regarding the classification of Korean rhythm, most studies tend to regard it as a syllable-timed language (e.g., Arvaniti, 2012). As in Dutch, pitch contributes to the marking of the IP in Korean; unlike Dutch, Korean also marks prosodic domains within the IP by a pitch movement, namely the Accentual Phrase (AP) (Jun, 2005).¹²

Third, to ensure the similarity of the stimuli across the languages, we use a single pseudo-sentence /nuto hɔm sɛpikaŋ/, which is phonologically similar in these two languages.¹³ Thus, phonologically speaking, the stimuli in this study are compatible with and identical to Dutch and Korean.

¹² According to Jun (2005), Korean also uses the intermediate phrase (ip). Since the ip ends with an AP, both ip and IP are marked by a boundary tone.

¹³ The Korean speakers used slightly different vowel sounds [a] and [ɔ] as substitutes for Dutch [ɑ] and [ɔ], respectively.

Fourth, this study uses acted speech to obtain well-controlled stimuli. We followed the methods developed by Scherer and colleagues (Banse & Scherer, 1996; Bänziger & Scherer, 2007) to ensure that the acted speech was as natural as possible (see Materials, below). To ensure comparability across languages, the same procedures were used by both Korean and Dutch stage directors and actors throughout the recording process.

Finally, to statistically account for the effects of all variables of interest, including by-participant and by-item variability, we used logistic mixed-effects models in our analyses.

2.2 Method

2.2.1 Auditory materials

We used the emotion portrayals from the Demo/Koremo (Dutch emotion/Korean emotion) corpus (Broersma et al., 2025). The corpus contains portrayals of eight different emotions, balanced in valence (positive vs. negative) and arousal (high-arousal vs. low-arousal), and with equal numbers of basic vs. non-basic emotions (Table 2.1). It includes recordings from eight Dutch and eight Korean actors, four females and four males in each group, to account for gender-related differences in the prosodic expression of emotions (Klatt & Klatt, 1990), with two tokens per emotion per actor. The corpus thus contains a total of 256 portrayals (8 emotions \times 8 actors \times 2 tokens \times 2 languages). All portrayals used the single pseudo-sentence /nuto hɔm sɛpikaŋ/. With the exception of the language of communication used between experimenter and participant, the elicitation and recording procedures were the same in Dutch and Korean.

2.2.1.1 Emotion elicitation and recording procedure

Recordings were made with a large membrane microphone at a sampling frequency of 44.1 kHz with 16-bit resolution, in a sound-attenuated room in the Netherlands or in Korea. In addition to the actors, two stage directors (both female) were involved, one Dutch and one Korean, to coach the actors during the recordings. Both stage directors were professionals, and all actors had either graduated from or were still enrolled as students at a colleague-level professional drama school in their own country. Each actor was recorded individually, in their native language and home country, in the presence of the stage director with the same native language. Actors and directors were paid for their service.

We adopted the “method acting” technique developed by Stanislavski (1988), which aims to achieve maximal naturalness of the acted emotions. Following this technique, the stage directors coached the actors to act out emotions by reliving a personal episode in which the actors had experienced the target emotion. All the actors and directors were highly experienced with this technique. In addition, following Banse and Scherer (1996), three scenarios per emotion were provided to illustrate the emotions prior to reenactment.

Different emotions were recorded separately, with a break in between. Actors and directors worked on reliving and recording each emotion for an average of 15 minutes (with a large variation across actors and emotions). The actors were asked to improvise, using any speech or movement they wanted, while reliving the target emotion, and to start uttering the pseudo-sentence into the microphone (and to cease moving) when they felt ready for it.

The director determined which utterances represented the emotion well, and stopped when the actor had recorded a sequence of at least five good portrayals. From those selected sequences, the final four portrayals per emotion per actor were used for the judgement study. If any of those four had any imperfections in sound quality (e.g., due to the actor moving) or recording quality (e.g., due to clipping), that portrayal was replaced with one of the remaining earlier portrayals that the director had approved of.

2.2.1.2 Judgement study

To determine the quality and naturalness of each emotion portrayal, we conducted a judgement study (see also Goudbeek & Broersma, 2010a, b) with native Dutch and Korean listeners who evaluated the portrayals in their respective native languages.

Participants were 24 native speakers of Dutch (11 males, 13 females) and 24 native speakers of Korean (12 males, 12 females). All were students (from Radboud University Nijmegen, the Netherlands, and Korea University, Seoul, respectively), who received course credits or a small payment. None reported any hearing or speech problems.

A total of 512 utterances ($8 \text{ actors} \times 8 \text{ emotions} \times 4 \text{ tokens} \times 2 \text{ languages}$) were included in the study. Each participant was only presented with the 256 stimuli in their native language, in a semi-random order. A computer screen showed nine response options, namely the eight emotions and “Neutral”, written in the participant’s native language, in nine equally-sized squares.

Response options had the same position throughout the experiment.¹⁴ The computer screen simultaneously showed a four-point scale from 1 (labeled “very unnatural”) to 4 (labeled “very natural” in the participants’ native language).

On each trial, participants heard an auditory stimulus and first identified it by clicking with the mouse on one of the nine response options (i.e., the eight emotions or “Neutral”), and then indicated the naturalness of the emotion expression by clicking on the four-point scale. There was no time limit for the responses. The experiment was run with the Praat MFC module (Boersma, 2001).

2.2.1.3 Corpus selection

For each portrayal, an “unbiased hit rate” was computed (Wagner, 1993) as a measure of how well the same-language native listeners recognized the intended emotion in the portrayal, while correcting for the participants’ biases to certain response options. The two most accurately recognized portrayals per actor per emotion (i.e., with the highest unbiased hit rates) were selected for the final Demo/Koremo corpus. When two portrayals per actor per emotion were equally well recognized, the one with the highest naturalness rating was selected. For an analysis of all unbiased hit rates and a further description of the unbiased hit rates of the portrayals included in the corpus, see Goudbeek and Broersma (2010b).

2.2.2 Visual materials

The main experiment used two adapted versions of the Geneva Emotion Wheel (Sacharin et al., 2012; Scherer, 2005; Scherer et al., 2010), representing the eight emotions of interest in this study—a Dutch version and a Korean version (Figure 2.1). The emotion wheels showed the names of the eight emotions (written in Dutch and Korean, respectively) in a circle, with the four quadrants representing all combinations of valence and arousal; clockwise, starting at the top right: positive/high (joy, pride), positive/low (relief, tenderness), negative/low (sadness, irritation), and negative/high (anger, fear). Each emotion was represented by four circles, with the small circles towards the center standing for low emotional intensity, and the big circles at the perimeter standing for high emotional intensity. A single circle in the middle of the wheel represented the response option “Neutral”.

¹⁴ From left to right, in the top row: “Relief”, “Tenderness”, “Pride”, “Joy”; on the middle row: “Neutral”; in the bottom row: “Sadness”, “Irritation”, “Anger”, “Fear”.

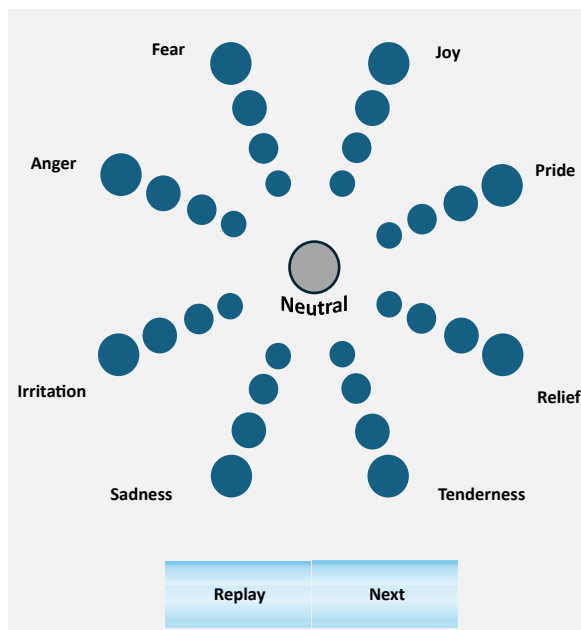


Figure 2.1. The emotion wheel in English (reproduced from Liang et al., 2023). Translation in Dutch and Korean, Joy: “Blijdschap”, “행복”; Pride: “Trots”, “자랑스러움”; Relief: “Opluchting”, “안도감”; Tenderness: “Vertederling”, “애정”; Sadness: “Verdriet”, “슬픔”; Irritation: “Irritatie”, “짜증”; Anger: “Woede”, “분노”; Fear: “Angst”, “공포”; Neutral: “Neutral”, “중립”.

2.2.3 Participants

There were two groups of participants: 31 native listeners of Dutch (27 females, 4 males, age: $M = 20.87$, $SD = 2.17$), all were students at Radboud University Nijmegen in the Netherlands, and 24 native listeners of Korean (12 females, 12 males, age: $M = 23.46$, $SD = 2.59$), all of whom were students at the University of Seoul, Korea. Participants took part in this experiment for a small payment or course credits. None of them had any knowledge of the language or culture of the other group, and none reported any speech or hearing problems. Furthermore, none of the participants had participated in the judgement study that was used for the selection of the portrayals (described above).

2.2.4 Procedure

Participants were tested individually in a sound-attenuated booth at Radboud University and at the University of Seoul. They were seated in front of a computer screen showing the emotion wheel in the participant's native language. Recordings were played at a comfortable loudness level over high-quality closed-back headphones. The experiment was implemented in Java and conducted on a standard laboratory computer.

Written instructions were provided in the participants' native language, asking them to listen to each stimulus, and to identify the emotion it conveyed to them by choosing from the eight emotions on the screen, as well as the intensity with which they thought the speaker had experienced the emotion, or, alternatively to choose Neutral (without intensity specification), and to indicate their answer by clicking on one of the circles on the screen. In the current paper, only the categorical responses, i.e., the chosen emotions, are analyzed; analyses of the perceived intensity of the emotion expressions will be presented in the next chapter. The instructions explained that participants could choose two emotions on a single trial if they felt that the stimulus conveyed more than one emotion (note that only the first emotion chosen is analyzed in the present paper), that they could listen to each stimulus more than once if they wanted to, and that they could correct a given response; they were, however, also asked to follow their first impression.

Presentation of the stimuli was blocked by language, with both blocks containing all 128 stimuli for that language, and always started with the block with the Korean recordings. Within each block, stimuli were presented in a randomized order. Participants were told before each block which language they were about to listen to. Each block started with eight practice trials, containing unique stimuli (i.e., not used in the main experiment). There was no time limit for the responses. The experiment took approximately 35-45 minutes.

2.3 Results

The data were analyzed in R (R Core Team, 2018). We ran one-sample *t*-tests to address Hypothesis 1 and the first part of Hypothesis 5, and a sequence of logistic mixed-effects models with the *lme4* package (Bates et al., 2015) to address all other hypotheses. The models used a combination of five predictors (fixed factors) as outlined in each analysis below: Speaker Language (Dutch vs. Korean recordings), Listener Language (Dutch vs. Korean listeners),

Arousal (high-arousal vs. low-arousal emotions), Valence (positive vs. negative emotions), and Basicness (basic vs. non-basic emotions). The outcome variable in all analyses was accuracy of emotion recognition (correct vs. incorrect). All logistic models used regression-style contrast coding for the five predictors ($-.5$ and $.5$ contrast codes for the variable levels listed first and second above).

The models included the maximal random structure justified by the design and leading to convergence (random intercepts for participants and items in all models, as well as random slopes for participants and items leading to convergence as detailed in each model below). In case of non-convergence, models were simplified by iteratively removing the random slopes accounting for the smallest amount of variance (Barr et al., 2013) until convergence was reached.¹⁵

2.3.1 Above-chance cross-cultural emotion recognition (Hypothesis 1)

The first research question concerned the accuracy of vocal emotion recognition within and across cultures. The first leg of this question is whether listeners can recognize emotions produced in an unknown language with above-chance accuracy. We expected above-chance performance (with a chance level in a 9-alternative forced-choice task, i.e., 1 out of 9, being .11) in both listener groups and for both recordings. This hypothesis was tested with four one-sample *t*-tests, which compared the average recognition accuracy of Dutch listeners in Dutch recordings and in Korean recordings, as well as the recognition accuracy of Korean listeners in Dutch recordings and in Korean recordings, to the chance level (see Figure 2.2).

¹⁵ Specifically, all models included random intercepts for participants and items. The maximal random structure for all models also included random by-participant slopes for Speaker Language and for the remaining variables of interest (Arousal, Valence, and Basicness in different models) as these variables were manipulated within participants but between items, and random by-item slopes for Listener Language as this variable was manipulated between participants and within items. When models with the maximal random structure did not converge, we removed random slopes one at time, starting with the random slope that accounted for the least variance. Thus, we report models with the maximal random structure allowing convergence. We further verified whether each random slope improved model fit significantly or not (with a series of model comparisons against models that did not include these slopes), as indicated for transparency for each model in the text below. However, we report models with all slopes for completeness (i.e., models with random slopes that did and did not improve model fit significantly but that allowed models to converge).

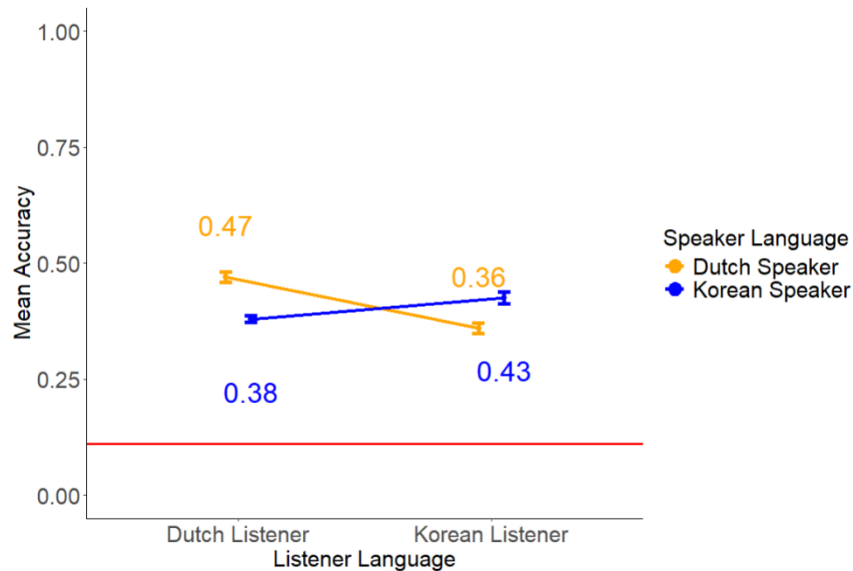


Figure 2.2. Accuracy (proportion of correct responses) for Dutch and Korean recordings by Dutch and Korean listeners. The red line indicates chance performance (.11). Error bars are ± 1 SE in all figures.

Performance was above chance in all conditions, always $t > 22$, and $p < .001$ (see Appendix A). Thus, consistent with earlier studies (Laukka et al., 2016; Laukka & Elfенbein, 2021), our data revealed that both groups of listeners were capable of recognizing vocal emotion expressions above chance, not only in their own language but also in an unknown language.

2.3.2 The in-group effect in emotion recognition (Hypothesis 2)

The second leg of the question concerns the in-group effect in emotion recognition. We hypothesized that listeners would recognize emotions from their own language more accurately than emotions from another language. We tested this hypothesis by assessing the joint effects of Speaker Language and Listener Language on emotion recognition (Model 1, Table 2.2). The model included Speaker Language and Listener Language as fixed effects, as well as random by-participant slopes for Speaker Language and random by-item slopes for Listener Language. Random by-item slopes for Listener Language improved model fit significantly, but random by-participant slopes for Speaker Language did not. The model showed a significant main effect of Listener Language, as Dutch listeners had generally higher accuracy than

Korean listeners (recognition accuracy was .06 higher in Dutch listeners than Korean listeners) and, crucially, a significant interaction between Speaker Language and Listener Language (removing this interaction resulted in a poorer model fit, $\chi^2(1) = 33.04, p < .001$). There was an in-group recognition benefit of .09 for Dutch listeners responding to Dutch over Korean recordings (mean accuracy: .47 vs. .38; see Figure 2.2), and an in-group recognition benefit of .07 for Korean listeners responding to Korean over Dutch recordings (mean accuracy: .43 vs. .36, see Figure 2.2). Thus, both groups of listeners displayed an in-group advantage: they recognized emotions produced by same-language speakers correctly more often than emotions produced by different-language speakers, consistent with the dialect theory of emotion (Elfenbein, 2013; Elfenbein & Ambady, 2002b).

Table 2.2. Summary of results of the logistic mixed-effects model analyses for Hypothesis 2. In all tables, coefficients (β) are transformed back to odds ($exp(\beta)$) for ease of interpretation. (Interpretations of the highest-level significant interactions in terms of differences in the odds of correct responses across conditions are reported below each table. Interpretations in terms of differences in proportions across conditions are reported in the main text.

| Model 1 (Hypothesis 2) | Estimates | | | | |
|-------------------------------|-----------|--------------|------|-------|--------|
| | β | $Exp(\beta)$ | SE | z | p |
| Intercept | -0.56 | 0.57 | 0.11 | -5.28 | < .001 |
| Speaker Language (SL) | -0.08 | 0.92 | 0.20 | -0.42 | .674 |
| Listener Language (LL) | -0.22 | 0.80 | 0.10 | -2.20 | < .050 |
| SL \times LL | 0.92 | 2.51 | 0.15 | 6.08 | < .001 |

Note. The Speaker Language \times Listener Language interaction showed a reliable in-group effect ($\beta = .92$). For Dutch listeners, the odds of a correct response were 1.45 times (= .37 log odds) higher when listening to Dutch than to Korean recordings. For Korean listeners, the odds of a correct response were 1.32 times (= .28 log odds) higher when listening to Korean than Dutch recordings.

2.3.3 The effect of Arousal on emotion recognition (Hypothesis 3)

The second research aim concerns the role of arousal, valence, and basicness in cross-cultural emotion recognition. First, Hypothesis 3 proposes that arousal influences recognition accuracy, both within and across cultures. This hypothesis was addressed with three models.

We first examined the impact of Arousal on emotion recognition in the entire dataset (positive and negative emotions), testing for a three-way interaction between Speaker Language, Listener Language, and Arousal. The best-fitting model included Speaker Language, Listener Language, and Arousal as fixed effects, and interacting random by-participant slopes for Speaker Language and Arousal, as well as random by-item slopes for Listener Language (see Model 2a in Table 2.3). Random by-item slopes for Listener Language and random by-participant slopes for Arousal improved model fit significantly, but random by-participant slopes for Speaker Language did not. This model showed the expected two-way interaction between Listener Language and Speaker Language (i.e., the in-group effect), and importantly, a main effect of Arousal on emotion recognition: recognition accuracy was .26 higher for low-arousal than high-arousal emotions. Further, interactions with Arousal were weak: there was a marginally significant two-way interaction between Arousal and Speaker Language and a marginal three-way interaction (removing the three-way interaction also resulted in a marginally poorer model fit, $\chi^2(1) = 2.87, p = .09$). This was due to the fact that the in-group effect was weaker for high-arousal emotions than low-arousal emotions. As shown in Figure 2.3a, Dutch listeners correctly recognized both high-arousal and low-arousal emotions more often in Dutch than in Korean recordings (an in-group recognition benefit of .12 for high-arousal emotions and .07 for low-arousal emotions). In contrast, Korean listeners correctly recognized low-arousal emotions more often in Korean than in Dutch recordings (an in-group recognition benefit of .14), but had similar accuracy for both speaker groups for high-arousal emotions.

Further, we tested the impact of Arousal on emotion recognition in two sub-analyses for positive emotions (joy, pride, tenderness, relief) and negative emotions (anger, fear, sadness, irritation).

For positive emotions, the best-fitting model included Speaker Language, Listener Language, and Arousal as fixed effects, and random by-participant slopes for Speaker Language and Arousal, as well as random by-item slopes for Listener Language (see Model 2b in Table 2.3). Random by-item slopes for Listener Language and random by-participant slopes for Arousal improved model fit significantly, but random by-participant slopes for Speaker Language did not. There was a main effect of Arousal (recognition accuracy was .21 higher for low-arousal than high-arousal emotions) and the expected two-way interaction between Listener Language and Speaker Language, but no three-way interaction with Arousal, suggesting that the in-group effect in positive emotions was not modulated by Arousal (Figure 2.3b). As expected,

removing the three-way interaction did not result in a poorer model fit, $\chi^2(1) = .47, p = .49$.

For negative emotions, the best-fitting model included Speaker Language, Listener Language, and Arousal as fixed effects, as well as interacting random by-participant slopes for Speaker Language and Arousal, and random by-item slopes for Listener Language (see Model 2c in Table 2.3). In this model, all random slopes improved the model fit significantly. Consistent with the results from Models 2a and 2b, this model showed a main effect of Arousal (recognition accuracy was .31 higher for low-arousal than for high-arousal emotions), and the expected two-way interaction between Listener Language and Speaker Language. There was also a three-way interaction with Arousal (Figure 2.3c; removing the three-way interaction resulted in a poorer model fit, $\chi^2(1) = 7.26, p < .01$). Dutch listeners correctly recognized both high-arousal and low-arousal emotions more often in Dutch than in Korean recordings (an in-group recognition benefit of .06 for high-arousal emotions and .07 for low-arousal emotions). Korean listeners correctly recognized low-arousal emotions more often in Korean than in Dutch recordings (an in-group recognition benefit of .16), but an analogous in-group benefit was not observed for high-arousal emotions (instead, there was an out-group recognition benefit of .03).

Our findings showed, importantly, that both groups of listeners recognized low-arousal emotions accurately more often than high-arousal emotions. Also, the in-group effect was confirmed in these three analyses. The in-group effect was marginally stronger for low-arousal than high-arousal emotions in the entire dataset. Arousal did not modulate the in-group effect in positive emotions, but the in-group effect in negative emotions was attenuated by Arousal in Korean listeners.

All models showed the expected two-way interaction between Speaker Language and Listener Language, and two models showed a three-way interaction. In Model 2a (including positive and negative emotions), there was a marginally significant interaction between Speaker Language, Listener Language, and Arousal. In Dutch listeners, the odds of a correct response were 1.69 times higher when listening to Dutch than Korean recordings (i.e., .52 log odds higher when listening to Dutch than Korean recordings) for high-arousal emotions, and 1.32 times higher when listening to Dutch than Korean recordings (i.e., .28 log odds higher when listening to Dutch than Korean recordings) for low-arousal emotions. In Korean listeners, the odds of a correct response were 1.05 times higher when listening to Dutch than Korean recordings (i.e., .05 log odds higher when listening to Dutch than Korean recordings)

for high-arousal emotions, and 1.76 times higher when listening to Korean than Dutch recordings (i.e., .57 log odds higher when listening to Korean than Dutch recordings) for low-arousal emotions.

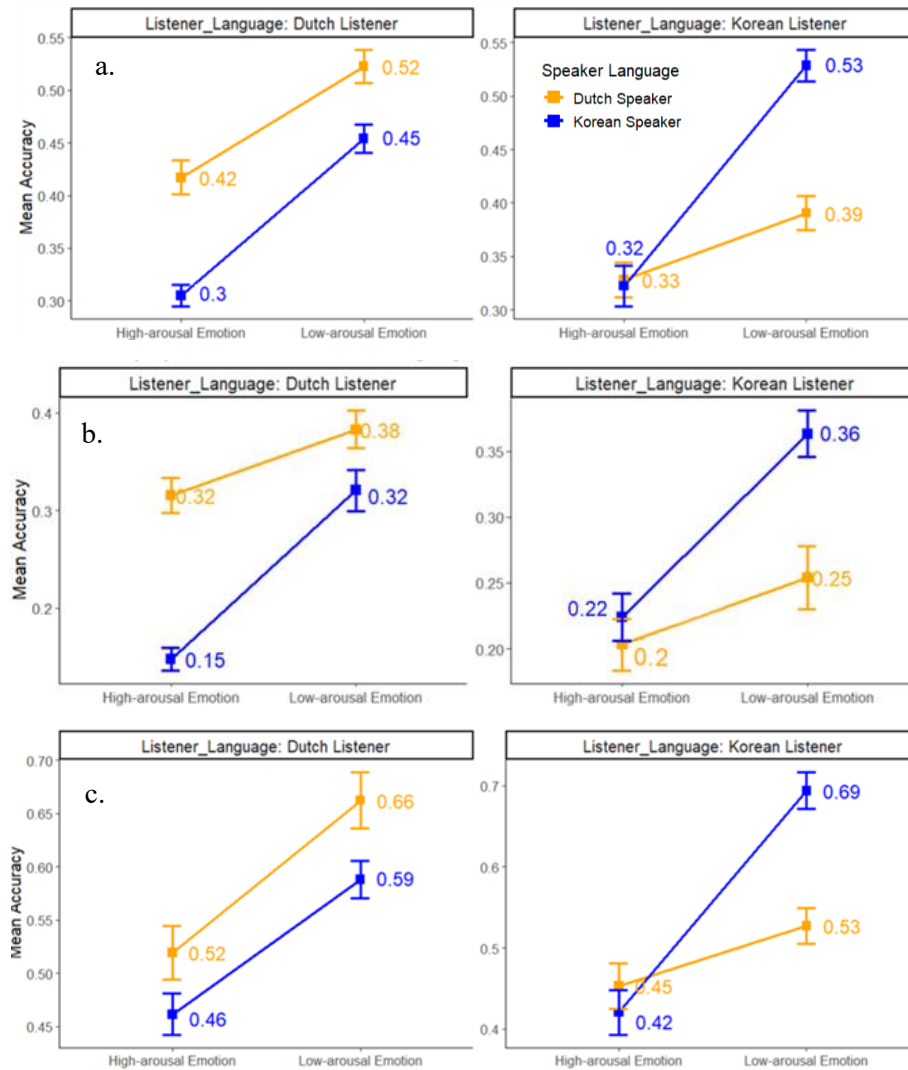


Figure 2.3. Recognition accuracy (Proportion correct) for high-arousal and low-arousal emotions in Dutch and Korean recordings by Dutch and Korean listeners (a) in the entire dataset (positive and negative emotions), (b) for positive emotions, and (c) for negative emotions.

The same three-way interaction with Arousal was again reliably found in Model 3c (negative emotions) but not in Model 3b (positive emotions only). In Model 3c, in Dutch listeners, the odds of a correct response were 1.27 times higher when listening to Dutch than Korean recordings (i.e., .24 log odds higher when listening to Dutch than Korean recordings) for high-arousal emotions. The odds of a correct response were 1.35 times higher when listening to Dutch than Korean recordings (i.e., .30 log odds higher when listening to Dutch than Korean recordings) for low-arousal emotions. In Korean listeners, the odds of a correct response were 1.14 times higher when listening to Dutch than Korean recordings (i.e., .12 log odds higher when listening to Dutch than Korean recordings) for high-arousal emotions. The odds of a correct response were 1.97 times higher when listening to Korean than Dutch recordings (i.e., .68 log odds higher when listening to Korean than Dutch recordings) for low-arousal emotions.

Table 2.3. Summary of results of the logistic mixed-effects model analyses for Hypothesis 3.

| | Estimates | | | | |
|--|-----------|--------------|------|--------|--------|
| | β | $Exp(\beta)$ | SE | z | p |
| Model 2a (Hypothesis 3) | | | | | |
| Intercept | -0.56 | 0.57 | 0.10 | -5.48 | < .001 |
| Speaker Language (SL) | -0.09 | 0.91 | 0.20 | -0.46 | .643 |
| Listener Language (LL) | -0.22 | 0.80 | 0.10 | -2.25 | < .050 |
| Arousal (A) | 0.82 | 2.27 | 0.21 | 4.01 | < .001 |
| SL \times LL | 0.92 | 2.51 | 0.15 | 6.09 | < .001 |
| SL \times A | 0.70 | 2.01 | 0.40 | 1.77 | .076 |
| LL \times A | 0.07 | 1.07 | 0.19 | 0.37 | .711 |
| SL \times LL \times A | 0.55 | 1.73 | 0.32 | 1.71 | .087 |
| Model 2b: Positive emotion dataset (Hypothesis 3) | | | | | |
| Intercept | -1.39 | 0.25 | 0.13 | -10.70 | < .001 |
| Speaker Language (SL) | -0.24 | 0.79 | 0.24 | -1.00 | .318 |
| Listener Language (LL) | -0.22 | 0.80 | 0.15 | -1.40 | .161 |
| Arousal (A) | 0.76 | 2.14 | 0.25 | 3.09 | < .010 |
| SL \times LL | 1.22 | 3.39 | 0.23 | 5.30 | < .001 |
| SL \times A | 0.97 | 2.64 | 0.47 | 2.05 | < .050 |
| LL \times A | -0.22 | 0.80 | 0.25 | -0.87 | .387 |
| SL \times LL \times A | -0.31 | 0.73 | 0.44 | -0.70 | .487 |
| Model 2c: Negative emotion dataset (Hypothesis 3) | | | | | |
| Intercept | 0.25 | 1.28 | 0.13 | 1.98 | < .050 |
| Speaker Language (SL) | 0.04 | 1.04 | 0.24 | 0.17 | .867 |
| Listener Language (LL) | -0.16 | 0.85 | 0.13 | -1.20 | .232 |
| Arousal (A) | 0.92 | 2.51 | 0.26 | 3.56 | < .001 |
| SL \times LL | 0.77 | 2.16 | 0.21 | 3.64 | < .001 |
| SL \times A | 0.45 | 1.57 | 0.49 | 0.92 | .359 |
| LL \times A | 0.30 | 1.35 | 0.28 | 1.10 | .273 |
| SL \times LL \times A | 1.23 | 3.42 | 0.44 | 2.76 | < .010 |

2.3.4 The effect of Valence on emotion recognition (Hypothesis 4)

Hypothesis 4 proposes that listeners recognize negative emotions more accurately than positive emotions. This question was addressed with three different analyses.

First, we examined the effect of Valence on emotion recognition in the entire dataset (high-arousal and low-arousal emotions), testing for a three-way interaction between Speaker Language, Listener Language, and Valence. The best-fitting model included Speaker Language, Listener Language, and Valence as fixed effects, and interacting random by-participant slopes for Speaker Language and Valence, as well as random by-item slopes for Listener Language (see Model 3a in Table 2.4, Figure 2.4a). Random by-item slopes for Listener Language and random by-participant slopes for Valence improved model fit significantly, but random by-participant slopes for Speaker Language did not. The model showed a significant main effect of Valence: recognition accuracy was .27 higher for negative than positive emotions, which was consistent with our predictions. The model also yielded the expected two-way interaction between Speaker Language and Listener Language, as in previous analyses (Models 1 and 2). However, there was no three-way interaction with Valence, indicating that the in-group effect was not modulated by Valence (removing the three-way interaction did not result in a poorer model fit, $\chi^2(1) = 1.94, p = .16$).

To further explore the effect of Valence on emotion recognition in high-arousal and low-arousal emotions, two further sub-analyses were run after splitting the dataset into two subsets: the high-arousal emotions (joy, pride, anger, fear) and the low-arousal emotions (tenderness, relief, sadness, irritation).

Model 3b tested the impact of Valence in the high-arousal emotions, including Speaker Language, Valence, and Listener Language as fixed effects, and interacting random by-participant slopes for Speaker Language and Valence, as well as random by-item slopes for Listener Language. In this model, all random slopes improved the model fit significantly. The model showed the expected interaction between Speaker Language and Listener Language. There was also a significant main effect of Valence, as recognition accuracy was .24 higher for negative than positive emotions, and a three-way interaction with Valence (removing this interaction resulted in a poorer model fit, $\chi^2(1) = 6.86, p = .01$): an in-group effect was found in Dutch listeners for negative and positive emotions, but not in Korean listeners (see Model 3b in Table 2.4, Figure 2.4b). Specifically, Dutch listeners recognized both negative and positive emotions correctly more often in Dutch than in Korean recordings

(an in-group recognition benefit of .06 for negative emotions and .17 for positive emotions). Korean listeners recognized positive emotions slightly better in Korean than in Dutch recordings (an in-group recognition benefit of .02), but had higher accuracy for negative emotions in Dutch than in Korean recordings (an out-group recognition benefit of .03).

In Model 3c, we focused on the modulation of the in-group effect by Valence in the low-arousal emotions. The model included Speaker Language, Valence, and Listener Language as fixed effects, and interacting by-participant slopes for Speaker Language and Valence, as well as random by-item slopes for Listener Language. In this model, random by-item slopes for Listener Language and random by-participant slopes for Valence improved model fit significantly, and random by-participant slopes for Speaker Language improved model fit marginally. The model showed a significant main effect of Valence, as recognition accuracy was .29 higher for negative than positive emotions (see Model 3c in Table 2.4, Figure 2.4c). As predicted, the interaction between Speaker Language and Listener Language reached significance. However, this model yielded no three-way interaction, indicating that the in-group effect was not modulated by Valence (removing the three-way interaction did not result in a poorer model fit, $\chi^2(1) = .38, p = .54$).

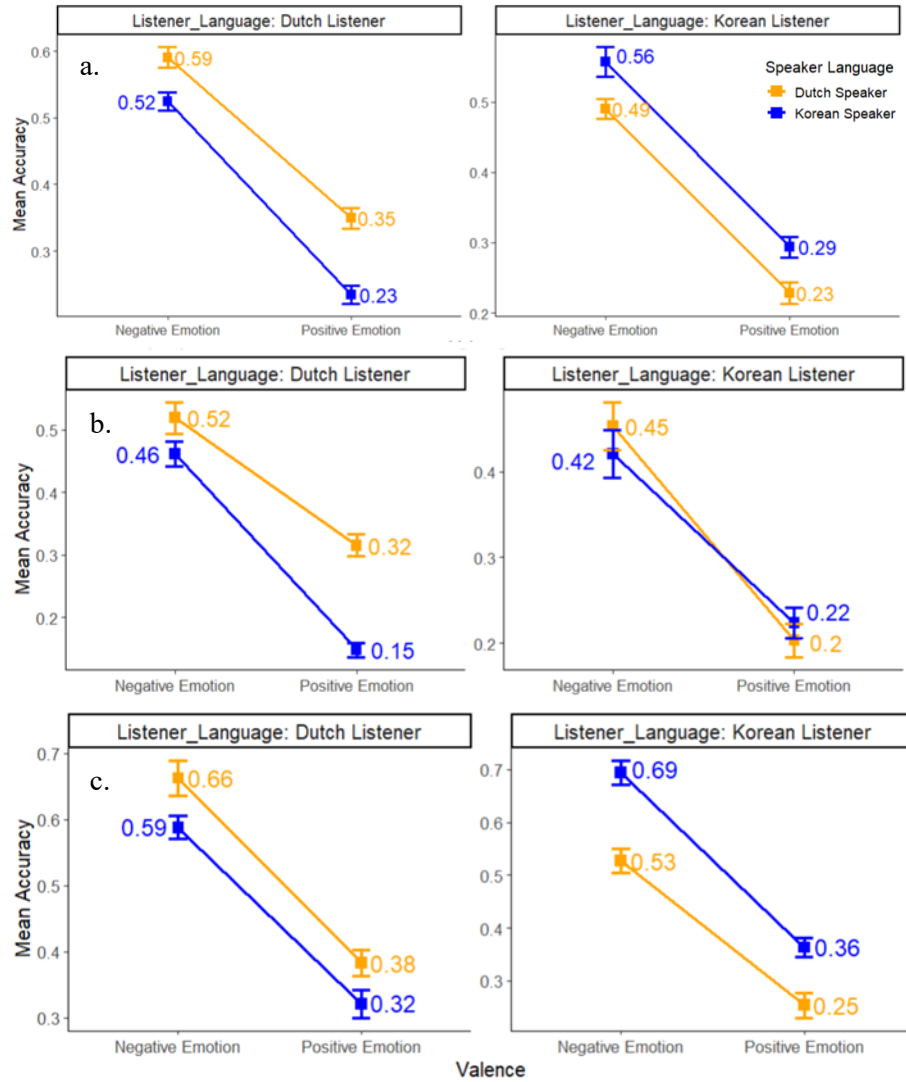


Figure 2.4. Recognition accuracy (Proportion correct) for positive and negative emotions in Dutch and Korean recordings by Dutch and Korean listeners in the entire dataset (high-arousal and low-arousal emotions), (b) for high-arousal emotions, and (c) for low-arousal emotions.

All three models showed the expected two-way interaction between Speaker Language and Listener Language, but only Model 3b showed a significant three-way interaction with Valence. For Dutch listeners, the odds of a correct response were 2.67 times (= .98 log odds) higher when listening to Dutch than Korean recordings of positive emotions. The odds of a correct response were 1.27 times (= .24 log odds) higher when listening to Dutch than to Korean recordings of negative emotions. For Korean listeners, the odds of a correct response were 1.13 times (= .12 log odds) higher when listening to Korean than to Dutch recordings of positive emotions. The odds of a correct response were 1.13 times (= .12 log odds) higher when listening to Dutch than to Korean recordings of negative emotions.

In sum, we built three models in three datasets (the entire dataset, the high-arousal emotion dataset, and the low-arousal emotion dataset), testing for recognition accuracy for positive and negative emotions. Importantly, as predicted, these models showed a significant main effect of Valence, indicating that recognition accuracy was higher for negative than positive emotions, both across and within cultures. While this finding is in line with previous work that recognition accuracy is higher for negative than positive emotions across cultures (Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2011), no previous studies have shown this to be the case within cultures, to the best of our knowledge.

Further, there was an in-group advantage in these three analyses, confirming again that listeners identified emotions produced in their native language correctly more often than emotions produced in an unknown language. However, Valence modulated the in-group effect in high-arousal emotions but not in low-arousal emotions.

Table 2.4. Summary of results of the logistic mixed-effects model analyses for Hypothesis 4. *P*-values of significant effects and interactions are in boldface.

| | Estimates | | | | |
|--|-----------|--------------|-----------|----------|----------------|
| | β | $Exp(\beta)$ | <i>SE</i> | <i>z</i> | <i>p</i> |
| Model 3a (Hypothesis 4) | | | | | |
| Intercept | -0.57 | 0.57 | 0.09 | -6.06 | < . 001 |
| Speaker Language (SL) | -0.08 | 0.92 | 0.18 | -0.47 | .640 |
| Listener Language (LL) | -0.21 | 0.81 | 0.10 | -2.13 | < . 050 |
| Valence (V) | -1.62 | 0.20 | 0.19 | -8.65 | < . 001 |
| SL \times LL | 0.95 | 2.59 | 0.15 | 6.20 | < . 001 |
| SL \times V | -0.25 | 0.78 | 0.35 | -0.71 | .477 |
| LL \times V | -0.11 | 0.90 | 0.20 | -0.54 | .591 |
| SL \times LL \times V | 0.45 | 1.57 | 0.32 | 1.41 | .160 |
| Model 3b: High-arousal emotion dataset (Hypothesis 4) | | | | | |
| Intercept | -0.98 | 0.38 | 0.12 | -8.03 | < . 001 |
| Speaker Language (SL) | -0.45 | 0.64 | 0.23 | -1.96 | < . 050 |
| Listener Language (LL) | -0.21 | 0.81 | 0.14 | -1.47 | .142 |
| Valence (V) | -1.55 | 0.21 | 0.24 | -6.56 | < . 001 |
| SL \times LL | 0.76 | 2.14 | 0.22 | 3.42 | < . 001 |
| SL \times V | -0.52 | 0.59 | 0.46 | -1.14 | .256 |
| LL \times V | 0.19 | 1.21 | 0.26 | 0.73 | .466 |
| SL \times LL \times V | 1.22 | 3.39 | 0.45 | 2.68 | < . 010 |
| Model 3c: Low-arousal emotion dataset (Hypothesis 4) | | | | | |
| Intercept | -0.15 | 0.86 | 0.13 | -1.14 | .254 |
| Speaker Language (SL) | 0.28 | 1.32 | 0.25 | 1.09 | .276 |
| Listener Language (LL) | -0.20 | 0.82 | 0.14 | -1.45 | .147 |
| Valence (V) | -1.74 | 0.18 | 0.27 | -6.45 | < . 001 |
| SL \times LL | 1.25 | 3.50 | 0.22 | 5.74 | < . 001 |
| SL \times V | 0.02 | 1.02 | 0.51 | 0.05 | .962 |
| LL \times V | -0.38 | 0.68 | 0.29 | -1.32 | .188 |
| SL \times LL \times V | -0.27 | 0.76 | 0.43 | -0.63 | .531 |

2.3.5 The effect of Basicness on emotion recognition (Hypothesis 5)

First, we tested whether listeners could recognize basic and non-basic emotions above chance, both within and across cultures. We compared recognition accuracy of each listener group for recordings from each speaker group, separately in basic and non-basic emotions, to the chance level (.11) with eight one-sample *t*-tests (see Appendix B). Performance was above chance in all conditions (always $t > 8.45$, and $p < .006$), indicating that listeners were capable of identifying basic as well as non-basic emotions above chance within and across cultures.

Further, we tested whether basic emotions are recognized more accurately than non-basic emotions, not only across cultures (Ekman, 1992a, 1999; Elfenbein & Ambady, 2002b; Hypothesis 5), but also within cultures. This hypothesis was addressed in Model 4 (see Table 2.5). The best-fitting model included Speaker Language, Listener Language, and Basicness as fixed effects, as well as interacting random by-participant slopes for Speaker Language and Basicness and random by-item slopes for Listener Language. Random by-item slopes for Listener Language and random by-participant slopes for Basicness improved model fit significantly, but random by-participant slopes for Speaker Language did not. The model showed the expected two-way interaction between Speaker Language and Listener Language (as in previous models). Importantly, as predicted, there was a significant main effect of Basicness: recognition accuracy was .69 higher in basic than non-basic emotions. Further, there was a significant three-way interaction between Speaker Language, Listener Language, and Basicness (removing this interaction resulted in a poorer model fit, $\chi^2(1) = 11.66$, $p < .001$). As shown in Figure 2.5, Dutch listeners recognized both basic and non-basic emotions correctly more often in Dutch than in Korean recordings (an in-group recognition benefit of .09 for both basic and non-basic emotions). Korean listeners recognized non-basic emotions correctly more often in Korean than in Dutch recordings (an in-group recognition benefit of .16), but had similar accuracy for both speaker groups for basic emotions. Thus, an in-group effect was found in Dutch listeners for both basic and non-basic emotions, but only for non-basic emotions in Korean listeners.

Table 2.5. Summary of results of the logistic mixed-effects model analyses for Hypothesis 5.

| Model 4 (Hypothesis 5) | Estimates | | | | |
|-------------------------------|------------------|--------------|------|-------|--------|
| | β | $Exp(\beta)$ | SE | z | p |
| Intercept | -0.55 | | 0.10 | -5.48 | < .001 |
| Speaker Language (SL) | -0.08 | 0.92 | 0.19 | -0.44 | .664 |
| Listener Language (LL) | -0.22 | 0.80 | 0.10 | -2.26 | < .050 |
| Basicness (B) | -0.96 | 0.38 | 0.10 | -4.80 | < .001 |
| SL \times LL | 0.92 | 2.51 | 0.15 | 6.20 | < .001 |
| SL \times B | 0.44 | 1.55 | 0.38 | 1.15 | .251 |
| LL \times B | 0.01 | 1.01 | 0.19 | 0.07 | .947 |
| SL \times LL \times B | 1.04 | 2.83 | 0.30 | 3.50 | < .001 |

Model 4 showed the expected two-way interaction between Speaker Language and Listener Language, and a significant three-way interaction with Basic/Non-Basic Emotion. In Dutch listeners, the odds of a correct response were 1.44 times higher when listening to Dutch than Korean recordings (i.e., .36 log odds higher when listening to Dutch than Korean recordings) for basic emotions. The odds of a correct response of Non-basic Emotions were 1.50 times higher when listening to Dutch than Korean recordings (i.e., .41 log odds higher when listening to Dutch than Korean recordings) for non-basic emotions. In Korean listeners, the odds of a correct response were 1.13 times higher when listening to Dutch than Korean recordings (i.e., .12 log odds higher when listening to Dutch than Korean recordings) for basic emotions. The odds of a correct response were 2.14 times higher when listening to Korean than Dutch recordings (i.e., .76 log odds higher when listening to Korean than Dutch recordings) for non-basic emotions.

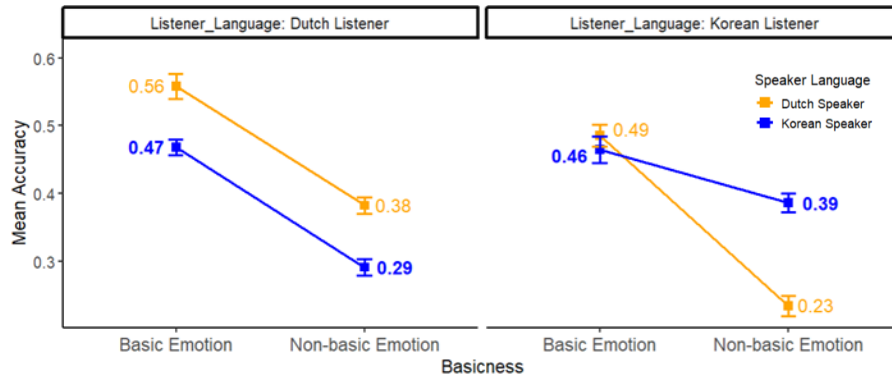


Figure 2.5. Recognition accuracy (Proportion correct) for basic and non-basic emotions in Dutch and Korean recordings by Dutch and Korean listeners.

The results showed that both groups of listeners recognized basic emotions more accurately than non-basic emotions across cultures, which is consistent with basic emotion theory (Ekman, 1992a, b, 1999). Importantly, as predicted, the results also showed, for the first time as far as we are aware, that listeners recognized basic emotions more accurately than non-basic emotions within cultures. The in-group effect was found in Dutch listeners for both basic and non-basic emotions, but only for non-basic emotions in Korean listeners. Korean listeners, on the other hand, identified basic emotions similarly in both Dutch and Korean recordings. In sum, our data showed that listeners recognized basic and non-basic emotions above chance within and across cultures.

2.4 Discussion

This study investigated cross-cultural emotion recognition with a carefully balanced design. We replicated and extended earlier findings and provided a number of novel insights into vocal emotion recognition. Our first aim (expressed in Hypotheses 1 and 2) was to test the predictions of dialect theory. First, as predicted in Hypothesis 1, both groups of listeners (Dutch and Korean) recognized emotions significantly above chance, not only in their native language, but also in an unknown language.¹⁶ Our study has thus replicated

¹⁶ In Chapter 2, we focus exclusively on the research questions and hypotheses related to the accuracy of emotion identification instead of the similarity structure of the emotion. The similarity structure will be examined in Chapter 4, where the performance by human listeners and machine classifiers will be compared, and the confusion matrices for the emotion identification will be presented.

the well-established finding that listeners can recognize vocally expressed emotions cross-culturally above chance, which is taken as evidence for universal principles in cross-cultural emotion recognition (Laukka & Elfenbein, 2021; Scherer et al., 2001). Second, as predicted in Hypothesis 2, we found an in-group advantage in both groups of listeners, such that listeners recognized emotions more accurately in their native language than in the unknown language. This in-group advantage is in line with previous studies that have consistently shown in-group advantages for emotions expressed by speakers of one's own peer group (Pell, Monetta et al., 2009), due to cultural norms and language-specific prosodic cues influencing intercultural emotion recognition (Elfenbein & Ambady, 2002b; Pell, Monetta et al., 2009; Scherer et al., 2001). Taking the results for Hypotheses 1 and 2 together, the present study provides support for dialect theory (Juslin & Laukka, 2003; Pell, Monetta et al., 2009; Scherer et al., 2001), which proposes the existence of universal principles in emotion recognition, while at the same time leaving room for culture-dependent and/or language-dependent factors (Elfenbein, 2013; Elfenbein & Ambady, 2002a).

Our second aim (expressed in Hypotheses 3-5) was to investigate the effects of valence, arousal, and basicness on the accuracy of cross-cultural and within-cultural emotion recognition. With a design that was aimed at optimally balancing the emotions on these three properties, we obtained new insights into their role in vocal emotion recognition.

First, we found that low-arousal emotions were recognized more accurately than high-arousal emotions within and across cultures. While it has been shown that the level of arousal of a speaker affects various characteristics of their speech production (e.g., pitch and duration) (Breitenstein et al., 2001; Goudbeek & Scherer, 2010), this is the first study that, to the best of our knowledge, has directly compared the recognition of low-arousal and high-arousal emotions. While we did not have prior expectations about the direction of the effect (Hypothesis 3), our finding that low-arousal emotions were recognized better than high-arousal emotions is in line with earlier reports that listeners can distinguish between emotions that are high or low in arousal (Laukka et al., 2005). These findings add additional nuance to the role of arousal in the communication of emotion.

Second, we found that negative emotions were recognized more accurately than positive emotions within and across cultures, as predicted in Hypothesis 4. As far as we are aware, this study is the first to compare recognition of positive and negative emotions *within* cultures. Our results *across* cultures are in accordance with the pattern first observed by Sauter et al. (2010), and confirmed by Scherer et al. (2011), as well as by the meta-analysis performed

by Laukka and Elfenbein (2021), who all showed recognition accuracy to be higher for negative than positive emotions across cultures in non-linguistic vocalizations. Further, our findings provide corroborating evidence that vocal cues can be used to distinguish between positive and negative emotions, which has been demonstrated by earlier studies (Cowen et al., 2019; Laukka & Elfenbein, 2021). Our results support the notion that recognizing valence is imperative for accurate emotion recognition (Russell, 1994).

Third, we found that basic emotions were recognized more accurately than non-basic emotions within and across cultures, as predicted in Hypothesis 5. As far as we are aware, this study has been the first to compare the recognition of basic and non-basic emotions within cultures. Our cross-cultural findings are consistent with earlier findings that basic emotions can be decoded more accurately than non-basic emotions across cultures in non-linguistic vocalizations (Sauter et al., 2010) as well as in facial expressions (Ekman, 1972; Elfenbein & Ambady, 2002b). The results are in line with the predictions of basic emotion theory, which posits that a small number of emotions are shared across cultures (Ekman, 1972, 1992a, b; Ekman et al., 1969). However, the finding that basic emotions were recognized more accurately than non-basic emotions *within* cultures, and that listeners recognized not only our four basic emotions but also our four non-basic emotions above chance across (as well as within) cultures, provides a challenge for the strong version of basic emotion theory (Gendron et al., 2018). We further observe a close relationship between valence and basicness. Among the four basic emotions in our experiment, only a single one was positive (joy), while the other three were negative (anger, fear, sadness). This is a direct result of the definition of basic emotions; among the six basic emotions that Ekman et al. (1969) originally proposed (anger, fear, happiness, sadness, disgust, and surprise), most emotions are negative; the only exceptions are happiness (positive) and surprise (which can be either negative or positive). The findings showed that negative emotions were recognized more accurately than positive emotions, and that basic emotions were recognized more accurately than non-basic emotions, both within and across cultures. The high recognition accuracy of negative and basic emotions reflects that valence and basicness are closely related. It is therefore no coincidence that positive emotions are seen to be closely connected to the formation and maintenance of social bonds (Shiota et al., 2004) and that non-basic emotions are sometimes referred to as the “social emotions” (Shiota et al., 2017), which are shared among members with similar cultural backgrounds.

To conclude, in the current study, we have replicated previous findings of above-chance cross-cultural vocal emotion recognition and of the in-group advantage in cross-cultural vocal emotion recognition, with an affectively and linguistically balanced design. The issue of whether emotion recognition is universal or culture-/language-specific has been a long-standing debate. The present results support the current consensus that the expression and recognition of emotions are affected by both universal and cultural/linguistic factors. Second, the affectively and linguistically balanced design has enabled us to shed new light on the respective influence of arousal, valence, and basicness on intercultural emotion recognition. Finally, we have presented and demonstrated the Demo/Koremo corpus for Dutch and Korean emotional speech (Broersma et al., 2025) with the aim of contributing to the methodological development of the study of cross-cultural vocal emotion recognition. Thus, with the current study, we hope to have contributed to a better understanding of cross-cultural emotion recognition and to the methodological toolkit of intercultural emotion recognition research.

Chapter Three

Interpreting the intensity of vocal emotions across cultures¹⁷

Abstract

This study investigates cross-cultural intensity ratings using the Demo (Dutch) and Koremo (Korean) corpora, with listeners and speakers from typologically different languages. Our results corroborate earlier findings and shed new light on intensity ratings of vocal emotions. First, contrary to previous findings (Ekman et al., 1987; Kommattam et al., 2019), we did not find an in-group bias in intensity ratings, such that neither listener groups gave higher ratings to emotions produced in their native language than in the unknown language. Second, intensity ratings were higher for high-arousal than for low-arousal, higher for negative than for positive, and higher for basic than for non-basic emotions. Notably, intensity ratings are more strongly correlated with arousal and basicness than valence, supporting earlier findings that high-arousal emotions are characterized by increased intensity (Laukka et al., 2005). Despite the significant effects of arousal, valence, and basicness on intensity ratings, they do not yield a successful dichotomy of emotions in intensity, since some particular emotions violate the general patterns of intensity ratings based on these three dimensions. Additionally, intensity ratings were higher for correct than incorrect responses. Together, these findings contribute to a better understanding of the role of intensity in vocal emotion across cultures.

Keywords: Dutch, Korean, cross-cultural, vocal emotions, intensity, in-group bias, arousal, valence, basicness

¹⁷ Liang, Y., van Hout, R., van Heuven, V. (submitted). Interpreting the intensity of vocal emotions across cultures.

3.1 Introduction

Emotions are ubiquitous in daily life and play a crucial role in interpersonal communication (Hall et al., 2009; Trampe et al., 2015). The fundamental role of emotions in communication has promoted exploration into the complexity of emotions (Elfenbein & Ambady, 2003b; Larsen et al., 1987; Liu et al., 2012). Most studies have mainly investigated emotions from a discrete approach (Biehl et al., 1997; Ekman & Friesen, 1969; Ekman et al., 1987; Elfenbein, 2013; Elfenbein et al., 2002; Elfenbein & Ambady, 2002; Laukka et al., 2013; see Laukka & Elfenbein, 2021 for a review), although some have explored emotions from a dimensional approach (Barrett, 1998; Mozziconacci, 2002; Russell, 1980). Emotions not only differ in quality (intended category) but also in quantity (intensity/strength), ranging from subdued to intense (Flett et al., 1986; Larsen & Diener, 1987; Laukka et al., 2005). For instance, anger can vary from mild irritation to intense rage (Spielberger et al., 1995). Misinterpretations of emotional intensity may lead to misunderstandings and even conflicts (Guerrero & La Valley, 2006). Therefore, intensity is classified as an important dimension of emotions (Larsen & Diener, 1985). Despite these findings, knowledge on intensity remains limited, especially in the vocal domain (Frijda et al., 1992; Reisenzein, 1994). The present study, therefore, examines the perception of intensity from a dimensional approach, aiming to provide a more comprehensive understanding of the role of intensity in vocal emotions, especially in a cross-cultural setting.

3.1.1 The intensity of emotions

Emotions are never produced neutrally. Instead, they are always expressed with some intensity (Mesquita & Frijda, 1992; Sonnemans & Frijda, 1994). Intensity refers to the strength or magnitude of individuals' responses evoked by emotions (Bänziger & Scherer, 2005; Diener et al., 1985; Larsen & Diener, 1987; Sonnemans & Frijda, 1994). Intensity measures how strongly an emotion is perceived by the receiver. People react to emotions with stronger intensity sooner than to those with weaker intensity (Kommattam et al., 2019). Intensity of emotion affects individuals' physiological and behavioral responses, such as decision-making (Frijda et al., 1992; Sonnemans & Frijda, 1994). For example, highly intense emotions can trigger increased heart rate and blood pressure on the part of the receiver.

3.1.2 Cross-cultural perception of emotional intensity

Research on cross-cultural perception of emotional intensity has predominantly concentrated on the visual domain (e.g., facial expressions), (Holz et al., 2021; Juslin & Laukka, 2001; Morningstar et al., 2021; Zhang & Pell, 2022). Findings reveal that both universal and culture-specific factors affect intensity ratings (Ekman et al., 1987; Kommattam et al., 2019; Matsumoto & Ekman, 1989). These studies can be broadly divided into two categories, i.e., (1) objective recognition accuracy based on stimulus intensity, and (2) subjective judgment of intensity.

3.1.2.1 Objective recognition accuracy based on intensity

Recognition accuracy of emotions increases with stimulus intensity, but this pattern varies across emotions (Hess et al., 1997; Juslin & Laukka, 2001; Shioiri et al., 1999). In the visual domain, Hess et al. (1997) found that while recognition accuracy of most basic facial expressions improved linearly with the physical intensity level, happiness was accurately identified at a lower intensity level. Moreover, Shioiri et al. (1999) reported that Japanese participants gave higher intensity ratings to facial expressions than Americans, but had lower recognition accuracy.

In the vocal domain, Juslin and Laukka (2001) discovered that listeners can easily decode emotions with stronger intensity than those with weaker intensity. However, this study has merely focused on two levels of intensity (strong and weak intensity). Expanding on this, Morningstar et al. (2021) manipulated the intensity of vocal emotions in 10% increments, ranging from 0% (neutral) to 100% (full-intensity). Recognition accuracy of some emotions (i.e., anger) increased linearly with intensity, whereas the accuracy of happiness stayed stable across low levels of intensity but increased at high levels of intensity.

However, Holz et al. (2021) argued that intensity plays a paradoxical role in emotion recognition. While emotions with moderate and strong intensity are accurately identified, those with peak intensity become ambiguous and difficult to recognize. One possible reason is that emotions with peak intensity may involve more than one emotion, making it difficult to classify them as a single emotion.

3.1.2.2 Subjective intensity judgments based on stimulus intensity

Cultural and linguistic factors affect how individuals rate emotional intensity in both visual and vocal domains, displaying different evidence for an in-group bias in a cross-cultural setting. Ekman et al. (1987) found that Western participants gave higher intensity ratings to basic emotions than non-Western participants for Caucasian facial expressions. However, Matsumoto and Ekman (1989) noticed inconsistent in-group bias across emotions, such that American participants consistently gave higher intensity ratings than Japanese participants to each emotion except disgust, whereas Japanese participants always rated disgust as the most intense emotion. More recently, in a meta-analysis, Kommattam et al. (2019) found that participants rated emotions produced by in-group members with higher intensity than those expressed by out-group members. Notably, some facial expressions that are difficult to recognize, i.e., contempt, embarrassment, and pride, were rated as less intense when evaluated by out-group members. These differences were attributed to cultural influence, supporting the idea that cultural norms affect emotion recognition. Therefore, Kommattam et al. (2019) proposed the notion of *in-group bias*, indicating that individuals rate emotions produced within their culture as more intense than those produced in an unfamiliar culture. Zhang and Pell (2022) adopted a fully cross-cultural design and found an overall in-group bias for both Canadian and Mandarin listeners. Particularly, both listener groups gave higher intensity ratings to anger and fear than to happiness and sadness. However, no in-group intensity bias was found for sadness (Canadian listeners) and for both anger and fear (Mandarin listeners).

Taken together, previous studies on intensity ratings of facial expressions present different views on the in-group bias of intensity ratings. Ekman et al. (1987) and Kommattam et al. (2019) reported higher intensity ratings of facial emotions for in-group than for out-group members. Notably, both studies provided participants with the corresponding emotion labels for each facial expression, informing participants of the specific emotion they were rating. However, Matsumoto and Ekman (1989) found no evidence for the in-group bias in intensity ratings in two separate experiments—one with and one without showing the labels of the intended emotions.

3.1.3 The relationship between intensity and emotional dimensions

In addition to investigating intensity separately, studies have examined intensity with other emotional dimensions, as emotions are intricate psychological experiences with multiple dimensions (Barrett & Russell, 2014; Russell, 1980; Russell, 2009; Russell & Barrett, 1999). One of the most well-

known models is the circumplex model, which classifies emotions into two fundamental dimensions: arousal and valence (Russell, 1980). Arousal, also referred to as activation, is the perceivers' physiological responses caused by emotions, ranging from low-arousal to high-arousal (Russell & Barrett, 1999). On the other hand, valence is the perceivers' personal experience affected by emotions, which can be either positive (pleasant) or negative (unpleasant). Furthermore, other theories propose that intensity should be added as one of the emotional dimensions (Larsen & Diener, 1987; Smith & Ellsworth, 1985). Intensity is the strength or magnitude of the emotional experience (Brehm, 1999), which is arguably related to arousal (Laukka et al., 2005; Mesquita & Frijda, 1992; Reisenzein, 1994). Although intensity and arousal are closely related, they are not interchangeable. While intensity concentrates on the overall strength of emotion, arousal emphasizes the physiological activation of perceivers caused by an emotion (Sonnemans & Frijda, 1995; Zsidó, 2023). For instance, intense emotions can be either high-arousal or low-arousal. Likewise, positive and negative emotions can be either intense or mild.

Basicness is another important attribute of emotions, although there is an argument on whether basicness should be one of the dimensions of emotions. As mentioned above, emotion theory distinguishes six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), which are universally recognized. Basic emotions are universal in the sense that they are expressed and understood in all cultures worldwide (Ekman, 1992a, b).

Extending this work, Laukka et al. (2005) investigated vocal emotions from a dimensional approach, and found that intensity is positively related to arousal (activation) in terms of ratings by listeners. However, their study employed a "many-to-one" design rather than a fully cross-cultural design, since they presented vocal emotions produced by two groups of speakers (British English and Swedish speakers) to only Swedish listeners.

Collectively, then, arousal, valence, intensity, and basicness categorize emotions for different aspects, providing a multidimensional framework to better understand the complexity of emotions. However, compared to these dimensions, intensity has been less studied, and our understanding of the interplay between intensity and arousal, valence, and basicness remains limited. To address this problem, it is necessary to explore the perception of emotional intensity and the extent to which it correlates with arousal, valence, and basicness.

3.1.4 The present study

This study investigates intensity ratings in Dutch and Korean listeners, whose culture and language are typologically different. It has two primary goals. The first goal aims to examine the in-group bias in intensity ratings. According to the notion of intensity bias, individuals tend to give higher intensity ratings to emotions expressed by members of their own group than to those expressed by out-group members (Ekman et al., 1987; Kommattam et al., 2019). Therefore, we hypothesize that both listener groups will exhibit an in-group bias in intensity ratings. First, we test the in-group bias across all responses, including correct and incorrect responses. We hypothesize that both listener groups will give higher intensity ratings to emotions produced in their native language than in the unknown language (Hypothesis 1). Second, we test the in-group bias across correct responses. Previous research on intensity ratings presented participants with the corresponding emotion labels, which may affect the ratings of intensity. However, in our materials, listeners were asked to rate the intensity of emotions without knowing the specific emotion (Goudbeek & Broersma, 2010). To avoid the effect of emotion labels on intensity ratings, we further examine intensity ratings on correct responses only. Based on the literature reviewed above, we hypothesize that there will exist an in-group bias across correct responses (Hypothesis 2).

The second goal aims to examine the effect of arousal, valence, and basicness on intensity ratings across all responses and correct responses. First, we examine the effect of Arousal on intensity ratings. According to the literature reviewed above, arousal is positively related to intensity (Laukka et al., 2005). Therefore, we hypothesize that intensity ratings will be higher for high-arousal than low-arousal emotions (Hypothesis 3). Second, we examine the effect of Valence on intensity ratings. Negative emotions are more related to increased levels of intensity than positive emotions (Schröder et al., 2001), and emotions with higher intensity are typically recognized as negative rather than positive (Scherer, 2003). Therefore, we hypothesize that intensity ratings will be higher for negative than positive emotions (Hypothesis 4). Third, we examine the effect of Basicness on intensity ratings. According to the emotion theory, there is a small number of emotions shared by all humans, which cause fixed behavioral responses (Ekman, 1972, 1992; Ekman et al., 1969). Basic emotions include more negative (anger, fear, sadness, and disgust) than positive emotions (happiness/joy), while surprise can be either positive or negative (Ekman, 1992b). In this study, we included four basic emotions, one positive (i.e., joy), and the other three (anger, fear, and sadness) negative. Since negative emotions are usually rated as more intense than positive emotions (see above, see also Kuppens et al., 2008), we hypothesize that

intensity ratings will be higher for basic than non-basic emotions (Hypothesis 5).

3.2 Method

To address the above research questions, we conducted an intensity rating experiment. During the experiment, all participants were asked to rate the intensity of each emotion they perceived on a four-point scale from 1 (low intensity) to 4 (high intensity).

3.2.1 Participants

Two groups of participants took part in this experiment. Thirty-one native listeners of Dutch (27 females, 4 males, age: $M = 20.87$, $SD = 2.17$) were students at Radboud University, the Netherlands; and 24 native listeners of Korean (12 females, 12 males, age: $M = 23.46$, $SD = 2.59$) were students at the University of Seoul, Korea. Prior to the experiment, each participant filled in a questionnaire to confirm they had no prior knowledge of the other group's culture or language. None of them reported any hearing problems. All participants got a small payment or course credits as a reward for their participation.

3.2.2 Stimuli

The auditory materials were vocal expressions from the Demo and Koremo (Dutch emotions and Korean emotions) corpus (Broersma et al., 2025).¹⁸ Each language corpus is based on 8 emotions produced by 8 actors (64 stimulus types). Each stimulus type was recorded twice (giving two tokens), resulting in a total of 128 portrayals per language (Table 3.1). For more information regarding the corpora and the recording procedure, see Liang et al. (2025, Chapter 2). All emotions were portrayed on the pseudo-sentence /nuto hɔm sɛpikaŋ/, which is phonologically compatible in Dutch and Korean.¹⁹ Using pseudo-sentences avoids semantic cues that may affect intensity ratings (Bhatara et al., 2016).

¹⁸ The scenarios and corpus are publicly available via Radboud University at <https://doi.org/10.34973/5kg3-9852>

¹⁹ The Korean speakers used slightly different vowel sounds /a/ and /o/ as substitutes for Dutch /ɑ/ and /ɔ/, respectively.

Table 3.1. The eight emotions in an arousal-valence grid (after Goudbeek & Broersma, 2010b, p. 2212), with basic emotions marked with “*”.

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

3.2.3 Procedure

Participants completed the experiment individually in a sound-attenuated booth at Radboud University in the Netherlands or at the University of Seoul in South Korea. The emotion wheel, with eight emotions and four circles in different sizes indicating different levels of intensity (Figure 3.1), was displayed on the computer screen with buttons labeled in the participants’ native language (Dutch or Korean). Participants listened to the recordings via high-quality headphones. The whole experiment was administered by a JavaScript on a standard laboratory computer, and took approximately 35-45 minutes.

All participants were given instructions in their native language. Participants were asked to listen to each stimulus and to select the target emotion from the eight options shown on the screen. The next step was to rate the emotional intensity they thought the stimulus conveyed. As a last option, they could choose Neutral (no category, no emotional intensity). Participants could listen to each stimulus more than once if they wished. In the previous study, we investigated only the first categorical response given. In the present study, we focused on the intensity ratings.

The whole experiment included two parts, blocked by language, with 128 stimuli for each language. The experiment started with the first part of the Korean stimuli. Before the experiment, there was a practice session, and participants were informed about the language they would hear.

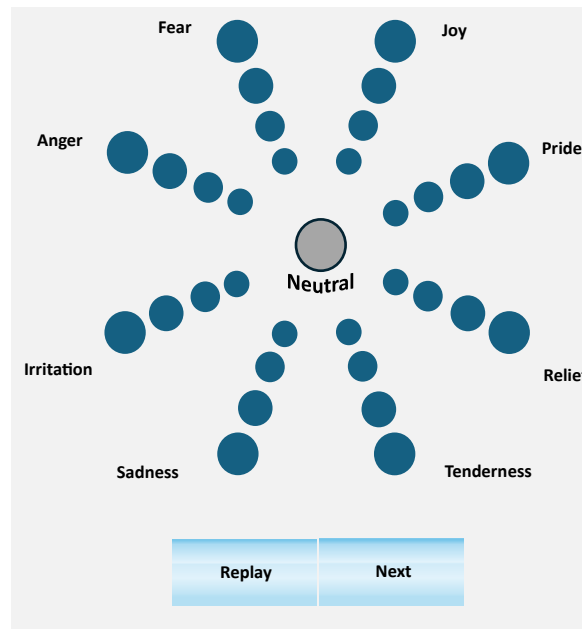


Figure 3.1. The emotion wheel in English (reproduced from Liang et al., 2023). Translation in Dutch and Korean, Joy: “Blijdschap”, “행복”; Pride: “Trots”, “자랑스러움”; Relief: “Opluchting”, “안도감”; Tenderness: “Vertederling”, “애정”; Sadness: “Verdriet”, “슬픔”; Irritation: “Irritatie”, “짜증”; Anger: “Woede”, “분노”; Fear: “Angst”, “공포”; Neutral: “Neutral”, “중립”.

3.2.4 Statistical analyses

The data analyses were performed in R (R Core Team, 2022). To address all hypotheses, we conducted a series of linear mixed-effects models with the *lme4* package (Bates et al., 2015), and pairwise comparisons for Hypotheses 1 and 2.

For Hypotheses 1 and 2, we performed separate analyses for the Dutch and the Korean listener groups, as their in-group bias, if present, works the other way around with respect to the Dutch and Korean speakers, and should be an independent group effect. Each group of listeners should show their group-specific bias. The analyses included two fixed predictors: Speaker Language (Dutch vs. Korean recordings) and Emotion (Joy, Pride, Anger, Fear, Tenderness, Relief, Sadness, and Irritation). The outcome variable was Intensity, which is a continuous variable. Additionally, we included two

random intercepts (Listener and Speaker) in all analyses. Random slopes with Emotion involved systematically returned non-convergent models. Therefore, we added only random by-listener slopes for Speaker Language. As for the fixed effects, the interaction between Speaker Language and Emotion was significant in all analyses, and removing it always resulted in higher AICs. Therefore, we kept the interaction effect. To investigate the in-group bias per Emotion, we completed the analyses by pairwise comparisons using the package (EMMEANS). The statistical summaries are given in Appendix C. Models 1 and 2 deal with all responses, models 3 and 4 with the correct responses.

For Hypotheses 3 to 5, the analyses included three fixed predictors: Speaker Language, Listener Language, plus one of the binary distinctions: Arousal (high-arousal vs. low-arousal emotions)/Valence (negative vs. positive emotions)/Basicness (basic vs. non-basic emotions). Again, we repeated the analysis, first for all responses, followed by an analysis on the correct responses only. Consequently, we have 3 (the three binary distinctions) \times 2 (all responses, correct responses) analyses. Their statistical summaries can be found in Models 5 to 10 in Appendix C. The random parts of these models were treated in the same way. We included all three random intercepts and the slopes for their interaction with three fixed variables. In Model 6, one of the slopes was removed as there was no random variation. We preferred to keep all interactions between the three fixed variables in the analysis to make sure that the outcomes are directly comparable. Reducing the models did not return deviant estimates of the effects involved. Our final step in these analyses was to add Emotion as a random effect to visualize how individual emotions deviated from the group they were assigned to. Again, this step did not change the estimates of the fixed effects of the overall model.

All linear mixed-effects models used regression-style contrast coding for the predictors with two levels (-0.5 and 0.5 contrast coding for the first and second levels shown above).

3.3 Results

3.3.1 The in-group bias in intensity ratings across all responses (Hypothesis 1)

The first research question examined whether there exists an in-group bias in intensity ratings across all responses. We hypothesized that both listener groups would give higher intensity ratings to emotions produced in their

native language than in the unknown language across all responses (Hypothesis 1). This hypothesis was tested in Model 1 and Model 2, which tested intensity ratings of the eight emotions aggregated over correct and incorrect responses but separately for Dutch and Korean listeners. Therefore, we split the whole dataset into two subsets according to listeners' language. Each model included two fixed predictors: Speaker Language and Emotion, and random intercepts for listeners and speakers, as well as random by-listener slopes for Speaker Language, which improved model fit significantly, resulting in the lowest AIC score. The random effects structure showed that intensity ratings varied across listener and speaker groups, highlighting individual differences in intensity ratings. The mean scores and their confidence intervals are given in Figure 3.2A-B. In both panels, we see that the two connecting lines tend to go down, but neither of the two lines (red and blue) is always higher than the other one. They cross each other.

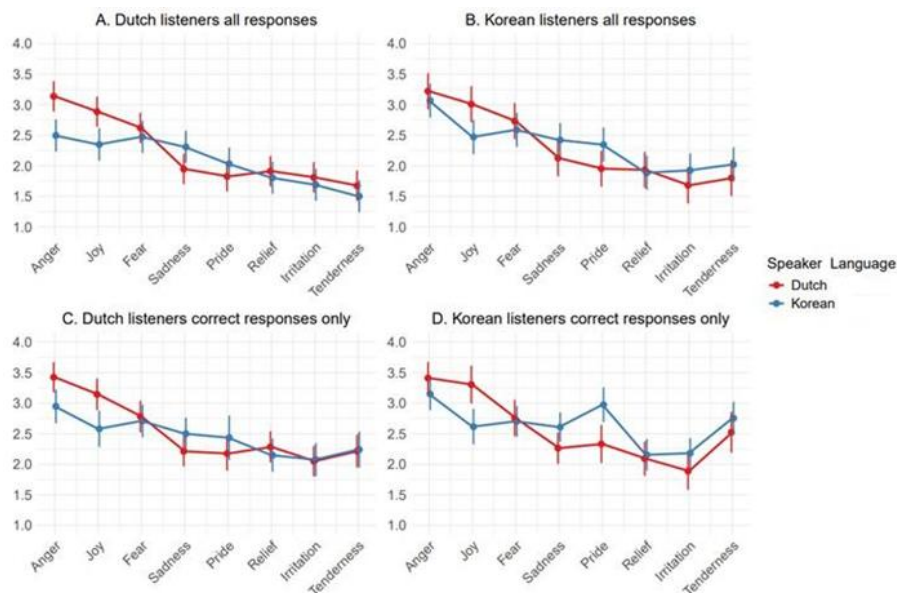


Figure 3.2. Intensity ratings by Dutch and Korean listeners (a) across all responses (correct and incorrect responses), (b) across the correct responses. Error bars represent ± 2 SE in all figures.

Intensity ratings of emotions in the whole dataset by Dutch listeners

The results of Model 1 (see Appendix C) revealed a significant main effect of Speaker Language on Intensity ($\Delta = .64$) in favor of the Dutch speakers. Furthermore, there was a significant main effect of Emotion on intensity, such that intensity ratings were higher for Anger than for any other emotion by Dutch listeners across all responses. More importantly, there were six significant two-way interactions between Speaker Language and Emotion (Pride/Fear/Tenderness/Relief/Sadness/Irritation), indicating that intensity ratings differed across speaker languages and emotions. To further examine the specific differences of intensity ratings for each emotion across Dutch and Korean recordings, we performed detailed Estimated Marginal Means (EMMEANS) analyses for pairwise comparisons across the eight emotions between Dutch and Korean recordings. To avoid Type I error caused by multiple comparisons, we used the *Tukey* adjustment. The results demonstrated that Dutch listeners gave significantly higher intensity ratings to Anger and Joy in Dutch than in Korean recordings, whereas they gave higher intensity ratings to Sadness in Korean than in Dutch recordings (see Table 3.2).

Table 3.2. Summary of EMMEANS analyses for Dutch listeners across all responses (standard error in parentheses, in all tables; significant *p*-values in boldface).

| Emotion | Speakers | | <i>df</i> | <i>t</i> | <i>p</i> |
|------------|-------------|-------------|-----------|----------|--------------|
| | Dutch | Korean | | | |
| Anger | 3.14 (0.12) | 2.50 (0.13) | 20.10 | 3.91 | 0.001 |
| Joy | 2.89 (0.12) | 2.35 (0.13) | 20.10 | 3.27 | 0.004 |
| Pride | 1.83 (0.12) | 2.03 (0.13) | 20.10 | -1.25 | 0.224 |
| Fear | 2.62 (0.12) | 2.47 (0.13) | 20.10 | 0.89 | 0.387 |
| Tenderness | 1.67 (0.12) | 1.50 (0.13) | 20.10 | 1.05 | 0.309 |
| Relief | 1.91 (0.12) | 1.80 (0.13) | 20.10 | 0.65 | 0.522 |
| Sadness | 1.95 (0.12) | 2.31 (0.13) | 20.10 | -2.20 | 0.040 |
| Irritation | 1.81 (0.12) | 1.69 (0.13) | 20.10 | 0.74 | 0.469 |

Intensity ratings of emotions in the whole dataset by Korean listeners

The results of Model 2 (see Appendix C) demonstrated no significant main effect of Speaker Language on intensity, as Korean listeners gave slightly different intensity ratings on average to emotions regardless of the type of recordings ($\Delta = .15$), but the effect is not significant. Furthermore, there was

a significant main effect of Emotion on intensity, indicating that intensity ratings varied across emotions. Additionally, there were five significant two-way interactions between Speaker Language and Emotion (Joy/Pride/Tenderness/Sadness/Irritation), showing that Korean listeners gave different intensity ratings to emotions across Dutch and Korean recordings. EMMEANS analyses (Table 3.3) showed that Korean listeners gave significantly higher intensity ratings to Pride in Korean than in Dutch recordings, while they gave higher intensity ratings to Joy in Dutch than in Korean recordings.

Table 3.3. Summary of EMMEANS results for Korean listeners across all responses.

| Emotion | Speakers | | <i>df</i> | <i>t</i> | <i>p</i> |
|------------|-------------|-------------|-----------|----------|--------------|
| | Dutch | Korean | | | |
| Anger | 3.22 (0.15) | 3.07 (0.14) | 21.90 | 0.90 | 0.380 |
| Joy | 3.01 (0.15) | 2.47 (0.14) | 21.90 | 3.13 | 0.005 |
| Pride | 1.95 (0.15) | 2.35 (0.14) | 21.90 | -2.29 | 0.032 |
| Fear | 2.73 (0.15) | 2.59 (0.14) | 21.90 | 0.84 | 0.413 |
| Tenderness | 1.80 (0.15) | 2.02 (0.14) | 21.90 | -1.29 | 0.210 |
| Relief | 1.93 (0.15) | 1.89 (0.14) | 21.90 | 0.27 | 0.787 |
| Sadness | 2.12 (0.15) | 2.42 (0.14) | 21.90 | -1.73 | 0.097 |
| Irritation | 1.68 (0.15) | 1.92 (0.14) | 21.90 | -1.43 | 0.168 |

Together, Dutch and Korean listeners displayed similar patterns in intensity ratings for the eight emotions across Dutch and Korean recordings, such that they gave higher intensity ratings to Anger, Joy, Fear, and Relief in Dutch than in Korean recordings, whereas they gave higher intensity ratings to Pride and Sadness in Korean than in Dutch recordings. However, both listener groups gave slightly higher intensity ratings to Tenderness and Irritation in their native language than in the unknown language, although this interaction did not reach statistical significance.

All in all, Hypothesis 1 has to be rejected. Neither listener group gave consistently higher intensity ratings to emotions produced in their native language than in the unknown language. Instead, they gave higher intensity ratings to particular emotions, even when these emotions were produced in the unknown language.

3.3.2 The in-group bias in intensity ratings across correct responses (Hypothesis 2)

As shown in Figure 3.2, both listener groups gave higher intensity ratings to correct responses than to all responses, indicating that intensity ratings are higher when emotions were recognized. More importantly, panels C and D in Figure 3.2 show again two crossing lines for the correct responses. The second hypothesis focuses on the presence of an in-group bias in the correct responses. Again, we performed two analyses for the two listener groups. We applied the same analyses as for all responses.

Intensity ratings of emotions in the subset of correct responses by Dutch listeners

The results of Model 3 (see Appendix C) demonstrated a significant main effect of Speaker Language on intensity, as Dutch listeners gave higher intensity ratings to emotions produced in Dutch than in Korean recordings ($\Delta = .48$). Also, there was a significant main effect of Emotion (Joy/Pride/Fear/Tenderness/Relief/Sadness/Irritation) on intensity, such that intensity ratings were higher for Anger than for any other emotions. Importantly, there were significant two-way interactions between Speaker Language and Emotion (Pride/Fear/Tenderness/Relief/Sadness/Irritation), suggesting that intensity ratings varied across emotions and speaker languages. Subsequent EM-MEANS analyses (Table 3.4) revealed that Dutch listeners gave higher intensity ratings to Anger, Joy, Fear, and Relief in Dutch than in Korean recordings, while they gave higher intensity ratings to Pride, Tenderness, Sadness, and Irritation in Korean than in Dutch recordings. However, the difference between Korean and Dutch recordings reached statistical significance only in the case of Anger and Joy.

Table 3.4. Summary of EMMEANS analyses for Dutch listeners in the correct responses.

| Emotion | Speakers | | <i>df</i> | <i>t</i> | <i>p</i> |
|------------|-------------|-------------|-----------|----------|--------------|
| | Dutch | Korean | | | |
| Anger | 3.42 (0.12) | 2.95 (0.14) | 22.90 | 2.84 | 0.009 |
| Joy | 3.14 (0.13) | 2.58 (0.15) | 30.50 | 3.14 | 0.004 |
| Pride | 2.18 (0.14) | 2.43 (0.18) | 61.70 | -1.21 | 0.232 |
| Fear | 2.78 (0.13) | 2.71 (0.13) | 23.10 | 0.43 | 0.670 |
| Tenderness | 2.22 (0.13) | 2.24 (0.15) | 32.40 | -0.14 | 0.894 |
| Relief | 2.28 (0.13) | 2.15 (0.14) | 23.70 | 0.80 | 0.431 |
| Sadness | 2.21 (0.12) | 2.50 (0.13) | 19.40 | -1.80 | 0.088 |
| Irritation | 2.05 (0.13) | 2.08 (0.14) | 22.70 | -0.15 | 0.883 |

Intensity ratings of emotions in the subset of correct responses by Korean listeners

The results of Model 4 (see Appendix C) revealed no significant main effect of Speaker Language on Intensity ($\Delta = .26$). Also, there was a significant main effect of Emotion on intensity, such that intensity ratings were higher for Anger than for any other emotion. Additionally, there were six significant two-way interactions between Speaker Language and Emotion (Joy/Pride/Tenderness/Relief/Sadness/Irritation). We performed Estimated Marginal Means (EMMEANS) analyses to further examine the differences in intensity ratings for each emotion across Dutch and Korean recordings (Table 3.5). The results demonstrated that Korean listeners gave higher intensity ratings to Anger, Joy, and Fear in Dutch than in Korean recordings, whereas they gave higher intensity ratings to Pride, Tenderness, Relief, Sadness, and Irritation in Korean than in Dutch recordings. However, only the ratings for Joy, Pride, and Sadness reached statistical significance.

Table 3.5. Summary of EMMEANS analyses for Korean listeners in the correct responses.

| Emotion | Speakers | | <i>df</i> | <i>t</i> | <i>p</i> |
|------------|-------------|-------------|-----------|----------|--------------|
| | Dutch | Korean | | | |
| Anger | 3.41 (0.13) | 3.15 (0.13) | 37.30 | 1.64 | 0.109 |
| Joy | 3.30 (0.15) | 2.61 (0.15) | 72.00 | 3.66 | 0.001 |
| Pride | 2.33 (0.15) | 2.97 (0.14) | 68.50 | -3.45 | 0.001 |
| Fear | 2.76 (0.15) | 2.70 (0.13) | 46.60 | 0.31 | 0.757 |
| Tenderness | 2.52 (0.17) | 2.75 (0.13) | 76.30 | -1.20 | 0.234 |
| Relief | 2.09 (0.14) | 2.15 (0.13) | 42.20 | -0.37 | 0.710 |
| Sadness | 2.26 (0.13) | 2.61 (0.12) | 26.90 | -2.34 | 0.027 |
| Irritation | 1.89 (0.15) | 2.18 (0.12) | 47.40 | -1.73 | 0.090 |

All in all, Hypothesis 2 is rejected. Neither listener group gave consistently higher intensity ratings to emotions produced in their native language than in the unknown language, even when they recognized the emotion. However, both listener groups gave higher intensity ratings to certain emotions, although these emotions were produced in the unknown language.

3.3.3 The effect of Arousal on intensity ratings (Hypothesis 3)

The third hypothesis concerns the effect of Arousal on intensity ratings, both across all responses and correct responses. The hypothesis was that both listener groups would give higher intensity ratings to high-arousal than to low-arousal emotions across all responses and especially across correct responses (Hypothesis 3). The statistical results of the mixed-effects analysis are given in Appendix C, Model 5 (all responses) and Model 6 (correct responses). The predicted outcomes are visualized in Figure 3.3.

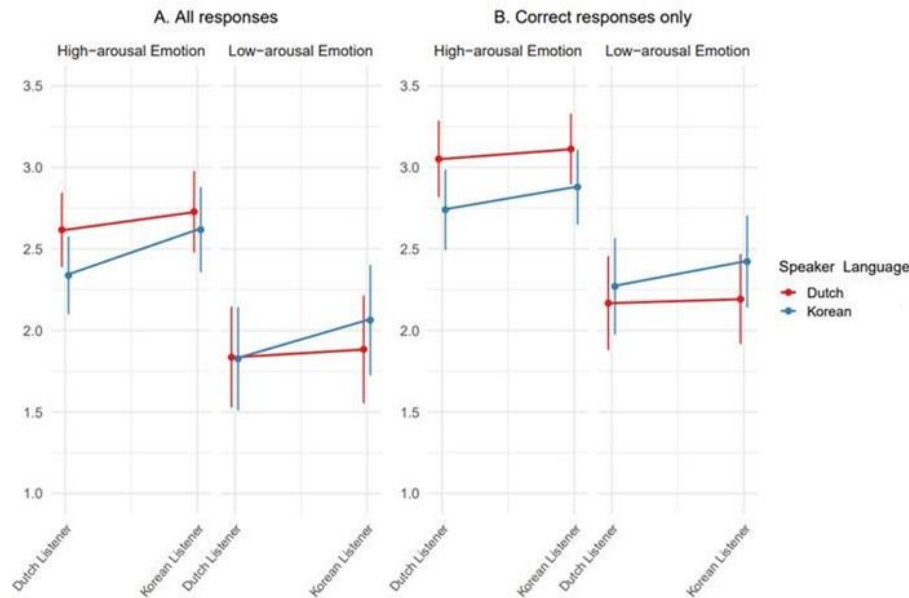


Figure 3.3. Means score of the intensity ratings for high-arousal and low-arousal emotions by Dutch and Korean listeners (A) across all responses (marginal R^2 : 0.098, conditional R^2 : 0.235), (B) across correct responses (marginal R^2 : 0.141, conditional R^2 : 0.234), and their confidence intervals (2SE).

Panels A and B exhibit a quite similar pattern, but with higher scores overall in panel B. More importantly, the outcomes show a clear split between high-arousal and low-arousal emotions. The main effect of arousal is .67 in the left panel (all responses) and .68 in the right panel (correct responses). Two interactions are significant as well. In panel A, it is the interaction between Speaker Language and Listener Language. In panel B, it is the interaction between Speaker Language and Arousal.

Panel A shows that Dutch listeners gave .29 higher ratings to emotions produced in Dutch than to those produced in Korean, whereas Korean listeners gave .07 higher ratings to emotions produced in Korean than to those produced in Dutch. However, there was no three-way interaction between Speaker Language, Listener Language, and Arousal, indicating that the two-way interaction between Speaker Language and Listener Language was not modulated by Arousal.

In the subset of correct responses only, there was a significant two-way interaction between Speaker Language and Arousal, such that the intensity rating was .67 higher for high-arousal emotions produced in Dutch than in

Korean, whereas both listener groups gave .36 higher ratings to low-arousal emotions produced in Korean than in Dutch.

Overall, the effect of arousal is obviously significant and by far the strongest one in the analysis. On the other hand, the R^2 values given in Figure 3.3 predict a substantial part (.234 to .235) of the variance present in the fixed effects (the conditional R^2). This effect is far from perfect, but including the random effects renders the effect size (unconditional R^2) substantially smaller: .100 in panel A, .140 in panel B. Since the effect is medium, we should have a look at the variation between the emotions. The success of the split can be investigated by including emotions within their categories, making it possible to evaluate their contribution. These intercepts are visualized in Figure 3.4, in combination with their confidence interval (high-arousal emotions on the left, and low-arousal emotions on the right). Within each group of four emotions, the mean is zero by definition. The difference between high- and low-arousal emotions was captured already by the main effect of arousal ($\Delta = .67$ and $.68$). If the emotions are a random effect within their category, their confidence intervals would include the zero value. Anger is extremely high and Pride is extremely low within both panels of Figure 3.4. Sadness is too high in the left panel in relation to the other three low-arousal emotions. Pride is about $-.52$, a value that is almost the difference between high- and low-arousal emotions. It is a clear violation of the predicted effect of high-arousal emotions. The same can be said for Sadness in panel A with respect to the low-arousal emotions. These results indicate that although intensity ratings are positively related to the level of arousal as predicted by Hypothesis 3, the binary split in high-arousal versus low-arousal does not yield an across-the-board dichotomy of the eight emotions in terms of intensity.

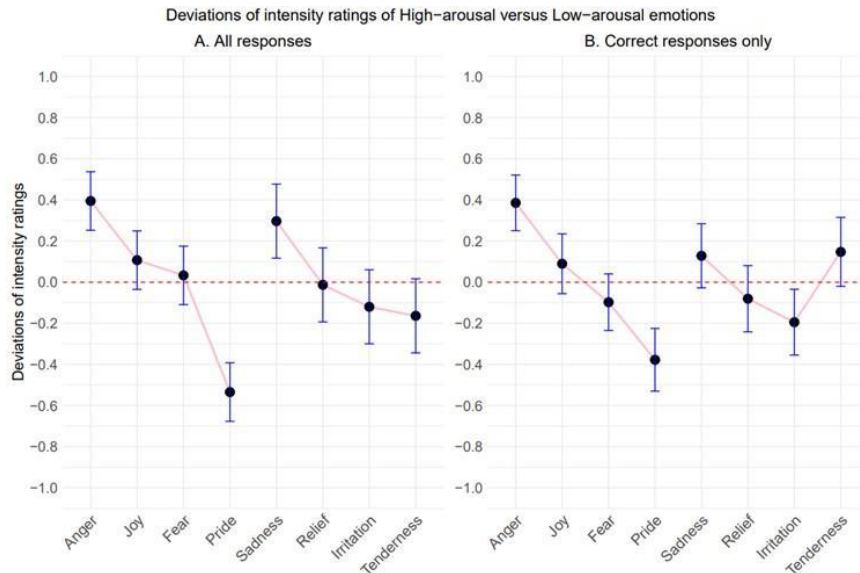


Figure 3.4. Intercepts of the intensity ratings of high-arousal (Anger, Joy, Fear, Pride) and low-arousal emotions (Sadness, Relief, Irritation, Tenderness) (A) across all responses, (B) across the correct responses, with their confidence interval (2SE) The order of emotions from left to right is listed as follows: (1) the four high-arousal emotions are presented left, and the four low-arousal emotions are presented right; (2) within each group (high-arousal or low-arousal), emotions are listed according to the size of their intercepts in the analysis of all responses (from high to low). For numerical values of intercepts in panels A and B, see Appendix D.

3.3.4 The effect of Valence on intensity ratings (Hypothesis 4)

The fourth hypothesis concerns the effect of Valence on intensity ratings, both across all responses and correct responses. The hypothesis was that both listener groups would give higher intensity ratings to negative than to positive emotions across all responses, especially in correct responses (Hypothesis 4). The statistical results of the mixed-effects analysis are given in Appendix C, Model 7 (all responses) and Model 8 (correct responses). The predicted outcomes are visualized in Figure 3.5.

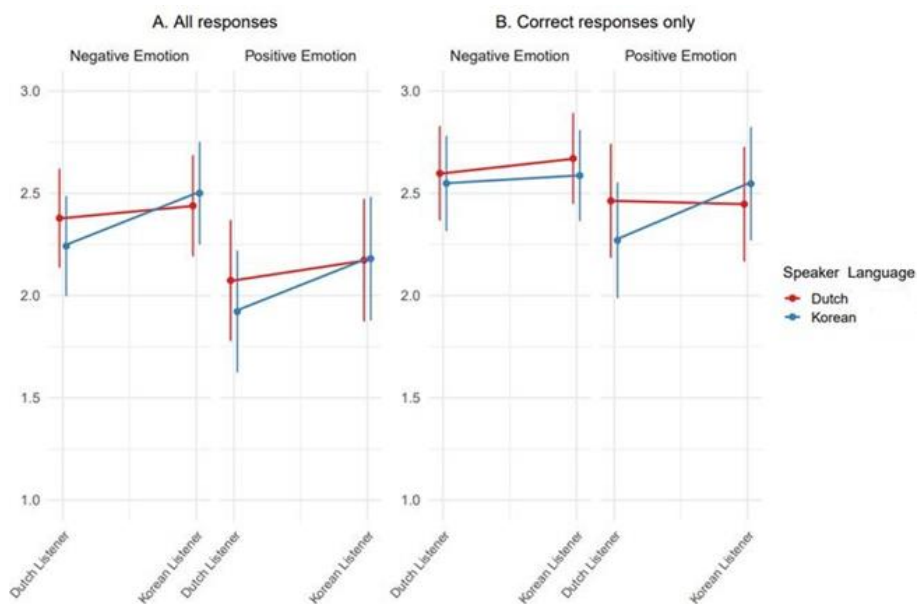


Figure 3.5. Means score of the intensity ratings for negative and positive emotions by Dutch and Korean listeners (A) across all responses (marginal R^2 : 0.028, conditional R^2 : 0.084), (B) across correct responses (marginal R^2 : 0.011, conditional R^2 : 0.087), and their confidence intervals (2SE).

Panels A and B display a similar pattern, with higher intensity ratings in panel B. Notably, the outputs exhibit an obvious split between negative and positive emotions. The main effect of valence is .30 in the left panel (all responses) and .17 in the right panel (correct responses). In panel A, there is a significant two-way interaction between Speaker Language and Listener Language. In panel B, there was a significant three-way interaction between Speaker Language, Listener Language, and Valence.

Panel A demonstrates that Dutch listeners gave .29 higher ratings to emotions produced in Dutch than in Korean, whereas Korean listeners gave only slightly higher ratings ($\Delta = .07$) to emotions produced in Korean than in Dutch. However, the three-way interaction between Speaker Language, Listener Language, and Valence did not reach statistical significance, revealing that the two-way interaction between Speaker Language and Listener Language was not modulated by Valence.

In the subset of correct responses, there was no significant two-way interaction between Speaker Language and Listener Language. However, there was a

significant three-way interaction between Speaker Language, Listener Language, and Valence. Specifically, both listener groups gave slightly higher intensity ratings to negative emotions produced in Dutch than in Korean (Δ for Dutch listeners: .03; Δ for Korean listeners: .08). However, both listener groups gave higher intensity ratings to positive emotions produced in their native language than in the unknown language (Δ for Dutch listeners: .32; Δ for Korean listeners: .08).

Together, the effect of valence is significant but weaker than that of the split by Arousal. Moreover, the R^2 values given in Figure 3.5 predict a small part of the variance: .084 to .087 of the variance shown in the fixed effects (the conditional R^2), but including the random effects results in much lower outcomes: .028 in panel A, .011 in panel B. Since the effect is small, we further examined the differences between the emotions within the categories of negative and positive emotions. It is possible to assess their relative contribution by examining the intercepts of each emotion within their category. The results for these intercepts are visualized in Figure 3.6, along with their confidence intervals (negative emotions on the left, and positive emotions on the right).

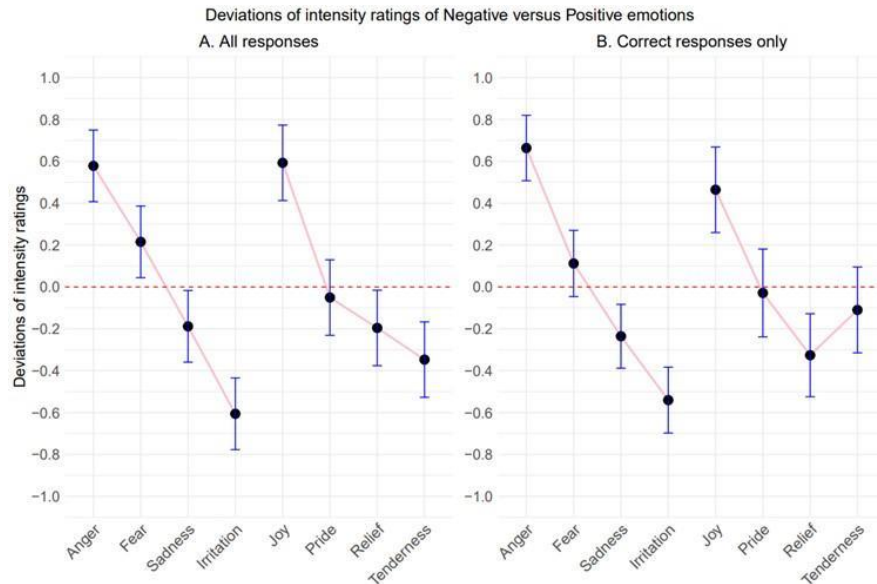


Figure 3.6. Intercepts of the intensity ratings of negative (Anger, Fear, Sadness, Irritation) and positive emotions (Joy, Pride, Relief, Tenderness) (A) across all responses, (B) across the correct responses, with their confidence interval (2SE). The order of emotions from left to right is listed as follows: (1) the four negative emotions are presented left, and the four positive emotions are presented right; (2) within each group (negative or positive), emotions are listed according to the size of their intercepts in the analysis of all responses (from high to low). For numerical values of intercepts in panels A and B, see Appendix D.

By definition, the mean of each group of four emotions is set to zero. The difference between negative and positive emotions was shown by the main effect of valence ($\Delta = .30$ and $.17$). When the emotions are regarded as a random effect within their category, their confidence intervals should contain the zero value. Anger is the second highest in panel A and the highest in panel B of Figure 3.6, while Irritation is the lowest in both panels of Figure 3.6. Joy is too high in the left panel compared to the other three positive emotions. Irritation is around $-.60$, a value that is almost twice the difference between negative and positive emotions. It is an obvious violation of the predicted effect of negative emotions. The same can be Joy in panel A in relation to the positive emotions. These results reveal that although intensity ratings are positively correlated with valence as predicted by Hypothesis 4, the binary split in negative versus positive does not result in a clear dichotomy of the eight emotions in intensity ratings.

3.3.5 The effect of Basicness on intensity ratings (Hypothesis 5)

The fifth hypothesis concerns the effect of Basicness on intensity ratings, both across all responses and correct responses. Both listener groups were expected to give higher intensity ratings to basic than to non-basic emotions across all responses, especially in correct responses (Hypothesis 5). This hypothesis was addressed with models 9 and 10. The statistical results of the mixed-effects analysis are given in Appendix C, Model 9 (all responses) and Model 10 (correct responses). The predicted outcomes are visualized in Figure 3.7.

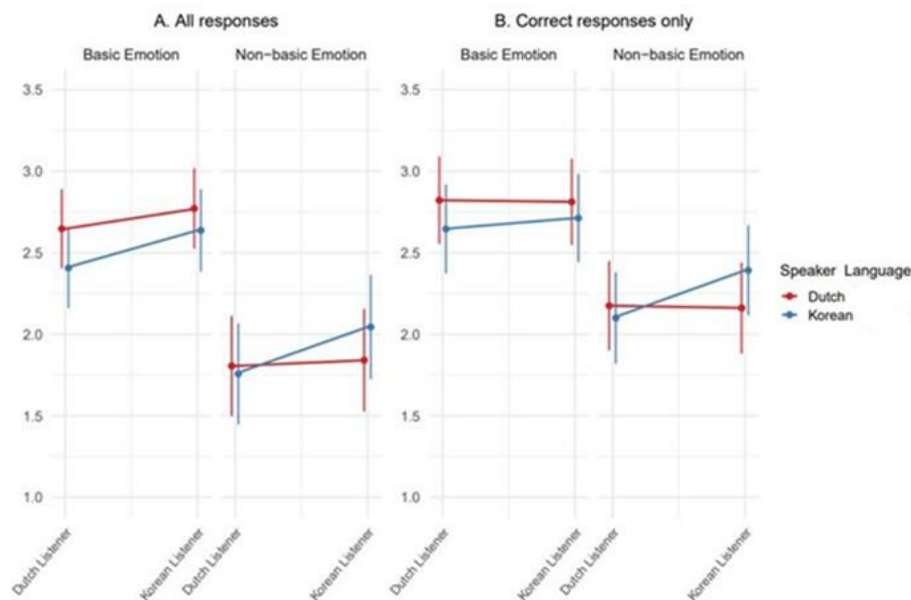


Figure 3.7. Mean intensity ratings for basic and non-basic emotions by Dutch and Korean listeners (A) across all responses (marginal R^2 : 0.121, conditional R^2 : 0.258), (B) across correct responses (marginal R^2 : 0.087, conditional R^2 : 0.163), and their confidence intervals (2SE).

Panels A and B exhibit similar patterns, with slightly higher intensity ratings in panel B. Importantly, they reveal a clear split between basic and non-basic emotions. The main effect of Basicness is .75 in the left panel, and .54 in the right panel. In panel A, there is a significant two-way interaction between Speaker Language and Listener Language. In panel B, there is a significant two-way interaction between Speaker Language and Listener Language.

Dutch listeners (panel A) gave .29 higher intensity ratings to emotions produced in Dutch than in Korean, whereas Korean listeners gave slightly higher ratings to emotions produced in Korean than in Dutch ($\Delta = .07$). Furthermore, there was a significant three-way interaction between Speaker Language, Listener Language, and Basicness, indicating that the two-way interaction between Speaker Language and Listener Language was modulated by Basicness.

Panel B shows a significant two-way interaction between Speaker Language and Listener Language. Here, both listener groups gave higher ratings to emotions produced in their native language than in the unknown language, with .19 and .15 higher in-group than out-group intensity ratings for Dutch and Korean listeners, respectively. More importantly, there was a significant three-way interaction between Speaker Language, Listener Language, and Basicness. Specifically, Dutch listeners gave .19 higher ratings to basic emotions in Dutch than in Korean, while they gave slightly higher ratings ($\Delta = .01$) to non-basic emotions in Dutch than in Korean. Korean listeners gave .12 higher ratings to basic emotions in Dutch than in Korean, whereas they gave .27 higher ratings to non-basic emotions in Korean than in Dutch.

Altogether, the effect of basicness is statistically significant and strong in the analysis. Moreover, the R^2 values presented in Figure 3.7 account for a considerable part of the variance: .258 and .163 of the variance shown in the fixed effects (the conditional R^2), but including the random effects leads to lower results: .121 in panel A, and .087 in panel B. Given that the effect is medium, we further investigated the variations between the emotions within the categories of basic and non-basic emotions. The success of the split can be analyzed by incorporating emotion as a random effect in the analysis, allowing for an evaluation of their relative contribution by examining the intercepts of individual emotions within each category. These intercepts are visualized in Figure 3.8, together with their confidence intervals (basic emotions on the left, and non-basic emotions on the right). The mean within each group of four emotions is defined as zero by definition.

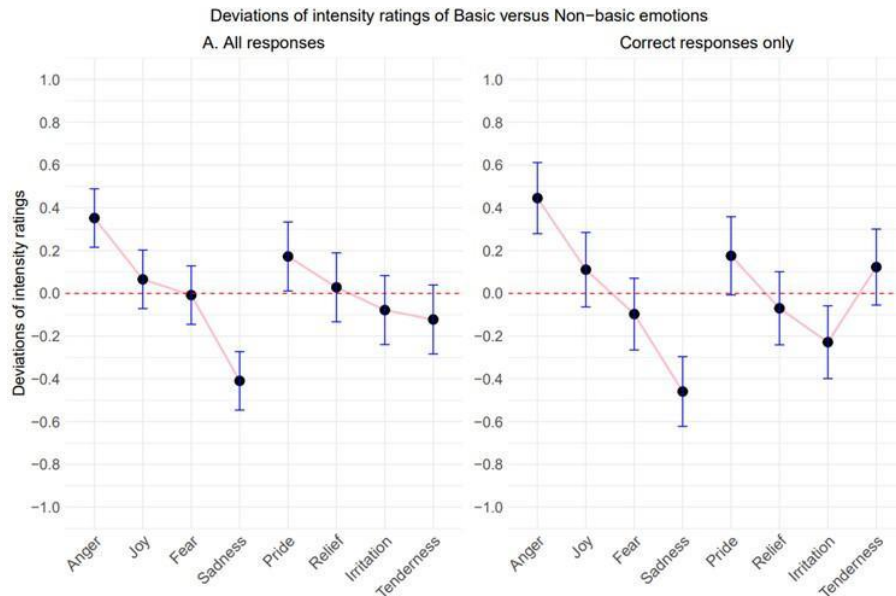


Figure 3.8. Intercepts of the intensity ratings of basic (Anger, Joy, Fear, Sadness) and non-basic emotions (Pride, Relief, Irritation, Tenderness) (A) across all responses, (B) across correct responses, with their confidence interval (2SE). The order of emotions from left to right is listed as follows: (1) the four basic emotions are presented left, and the four non-basic emotions are presented right; (2) within each group (basic or non-basic), emotions are listed according to the size of their intercepts in the analysis of all responses (from high to low). For numerical values of intercepts in panels A and B, see Appendix D.

The difference between basic and non-basic emotions was manifested by the main effect of basicness ($\Delta = .75$ and $.54$). When the emotions are considered as a random effect within their category, their confidence intervals would involve the zero value. Anger is the highest, and Sadness is the lowest within both panels of Figure 3.8. Pride is too high in the left panel compared to other non-basic emotions. Sadness is at about $-.40$, a value that is close to the difference between basic and non-basic emotions. Clearly, it is a violation of the predicted effect of basic emotions. These results suggest that although intensity ratings are positively related to basicness overall, the binary split in basic versus non-basic emotions does not yield a clear dichotomy of the eight emotions in intensity.

3.3.6 Analyses of Arousal, Valence, and Basicness compared

In Table 3.6, we compare the statistical outcomes of the three binary splits presented above in order to evaluate their relative success in predicting the emotional intensity ratings.

Table 3.6. Summary of the analysis of the dimensions of Arousal, Valence, and Basicness across all responses and correct responses only; violations are intercepts of emotions with an outlier value in the wrong direction.

| Dimension | Data | R ² | | Effect size [CI] | Violations |
|-----------|---------|----------------|-------------|------------------|------------|
| | | Marginal | Conditional | | |
| Arousal | All | 0.098 | 0.235 | -0.67 | 2 |
| | Correct | 0.141 | 0.234 | -0.68 | 1 |
| Valence | All | 0.028 | 0.084 | -0.30 | 3 |
| | Correct | 0.011 | 0.087 | -0.17 | 3 |
| Basicness | All | 0.121 | 0.258 | -0.75 | 2 |
| | Correct | 0.087 | 0.163 | -0.54 | 1 |

The results revealed that conditional R² was higher, of course, than marginal R² in each dimension, indicating that the combination of both fixed predictors and random effects explains the variance of intensity ratings better than fixed predictors alone, but the values of .084 and .087 are quite low for Valence. None of the three dimensions perfectly dichotomizes intensity. The higher conditional R² is medium. Furthermore, we found that Arousal and Basicness had clearly higher conditional R² values than Valence, with the highest value for Basicness across all responses (.258), and the highest value for Arousal across correct responses (.234), revealing that intensity ratings were more strongly related to Arousal and Basicness than to Valence. Also, Valence has more deviant emotions than Arousal and Basicness, suggesting again that the impact of Valence on intensity ratings is minimal. This finding corroborates earlier findings that arousal is positively related to intensity (Laukka et al., 2005). All three splits gave an effect as predicted by the hypotheses, but overall, none of the dimensions returned an across-the-board split between the emotions.

3.4 Discussion

This study investigated intensity ratings of vocal emotions with a balanced design involving speakers and listeners from typologically different cultures and languages, i.e., Dutch and Korean. Our results partially corroborate earlier findings and shed new light on intensity in affective sciences. Prior studies on intensity ratings have predominantly focused on facial expressions (Ekman et al., 1987; Kommattam et al., 2019; Matsumoto & Ekman, 1989; Shioiri et al., 1999; Yrizarry et al., 1998). Some studies investigated intensity ratings in the vocal domain, but not cross-culturally (Holz et al., 2021; Juslin & Laukka, 2001; Laukka et al., 2005). Most of these studies target basic emotions only (Holz et al., 2021; Juslin & Laukka, 2001; Laukka et al., 2005), and/or studied an unbalanced number of emotions in terms of arousal and valence. Consequently, the current literature on emotional intensity cannot be generalized to non-basic emotions. In our study, therefore, we included four basic emotions (Joy, Anger, Fear, and Sadness) and four non-basic emotions (Pride, Tenderness, Relief, and Irritation), which are balanced in arousal and valence, making it possible to examine the relative contributions of arousal, valence, and basicness to intensity ratings. This would be the first study that explores intensity ratings in vocal emotions with three dimensions, i.e., arousal, valence, and basicness.

This study had two main goals. First, we tested the in-group bias in intensity ratings across all responses (Hypothesis 1) and then in the subset of correct responses only (Hypothesis 2). The notion of in-group bias entails that individuals rate facial expressions of emotion produced by members of their own (or similar) ethnic group as more intense than expressions produced by members of a different group (Kommattam et al., 2019). The in-group bias was attested for facially expressed emotions, but it is unclear if the principle also applies to vocal expression of emotions. First, our results contradicted the in-group bias hypothesis, since neither listener group gave higher intensity ratings to emotions produced in their native language than in the unknown language (Hypothesis 1). Rather, intensity ratings were higher for some specific emotions, even when these emotions were produced in the unknown language. Instead of in-group bias, intensity ratings are more emotion-dependent, which affects the degree of perceived intensity. For instance, Dutch listeners exhibited an in-group bias for emotions like Anger and Joy, whereas Korean listeners displayed a bias for these two emotions, but not an in-group bias, as they rated Anger and Joy with higher intensity in Dutch than in Korean. Possibly, then, some emotions may be produced with more intensity than others in all (or at least a majority of) languages and/or cultures.

Second, we tested the in-group bias in intensity ratings in the subset of correct responses only. Similar to the findings in Hypothesis 1, we found no general in-group bias in intensity ratings in this subset (Hypothesis 2). Here, too, both listener groups gave higher intensity ratings to Anger and Joy in Dutch than in Korean, whereas they gave higher ratings to Sadness in Korean than in Dutch. These findings provide evidence against a general in-group bias in intensity ratings, even when the observations are restricted to correct responses only. Again, listeners rated some particular emotions higher, even though they were not produced in their native language.

The in-group intensity bias was originally formulated and tested in the cross-cultural perception of facial expression of emotion. We adopted the in-group intensity bias hypothesis from Kommattam et al. (2019). These authors present a survey of the cross-cultural perception of emotional intensity for three groups of white European speakers and listeners, i.e., Dutch, English, and Finnish. The first nine studies investigated the perception of a single emotion, i.e., embarrassment—a non-basic emotion. Three more studies were reported in the survey; these tested nine different emotions but reported the results aggregated over all nine without providing a breakdown by specific emotion. It is impossible, therefore, to verify whether indeed all nine emotions consistently showed the in-group bias or whether one or more emotions deviated from the general effect. Moreover, the twelve studies reported by Kommattam and associates presented facial expressions of emotions produced by white (Dutch, American) versus Arab (Moroccan, Turkish) models, while the perceivers were exclusively white Europeans (or European Americans). The complementary condition, with observers from an Arab background, was not implemented, so that it is possible that Arab observers, had they been included in the experiments, might also have judged the European expressions of emotion more intense than the same emotions expressed by fellow Arabs. To address this problem, we adopted a “two-by-two” design by involving both speakers and listeners whose culture and language are typologically different. This fully cross-cultural design enables us to test the in-group bias in intensity ratings effectively. Finally, we cannot exclude the possibility that the facial and vocal transmission channels differ fundamentally, so that effects found reliably in the facial modality may be absent or even reversed in the vocal mode. Interestingly, the in-group bias was found to be strongest for facially expressed non-basic emotions (contempt, pride, embarrassment). This finding supports the idea that secondary emotions are expressed less clearly than basic emotions, so that it takes an in-group perceiver to pick up the cues. Interpreting emotions was found to be more challenging from vocal than from visual cues (App et al., 2011). Its absence (or emotion specificity) in the vocal domain may be attributed to the complexity and strength of the vocal modality

(e.g., the audible differences of sound features even across languages). Despite this, the in-group intensity bias was less clear in weakly expressed vocal emotions than in strongly expressed emotions in the recent study by Zhang and Pell (2022), and the in-group bias was not consistently found for all four emotions tested. Our results confirm Zhang and Pell (2022) insofar as the in-group intensity bias did not apply to all emotions tested. This begs the question of whether the in-group intensity bias operates through the same mechanism in the facial and vocal perception of emotion.

Further, we observed that both listener groups rated certain vocal emotions more intensely when produced in the unknown language than in their native language. This pattern was inconsistent, depending on the emotion and the speaker's language. It can probably be attributed to the fact that Dutch actors produced Anger more intensely in Dutch than the Korean actors in Korean. This highlights the need to ensure "stimulus equivalence" (Matsumoto, 2002) when testing the in-group bias cross-culturally. Differences caused by language-specific prosodic features may override any potential in-group bias.

The second goal of our study was to examine the role of arousal, valence, and basicness, as binary properties of emotional expressions, in intensity ratings across all responses, and for correct responses only (Hypotheses 3 to 5). With the relatively larger number of eight emotions balanced in these three dimensions, our results provide new insight into their relative contributions to intensity ratings.

First, high-arousal emotions were rated as more intense than low-arousal emotions (Hypothesis 3), which confirms earlier studies (Holz et al., 2021; Laukka et al., 2005). The level of arousal is closely related to the rated intensity of emotions, which were found to share similar acoustic cues, such as fundamental frequency (F0), fundamental frequency variation, speech rate, etc. (Laukka et al., 2005).

Second, negative emotions were rated as more intense than positive emotions (Hypothesis 4). Little is known about the relationship between valence and intensity, although it was found earlier that non-verbal vocalizations with high and peak intensities were more prone to be rated as negative emotions (Holz et al., 2021). Notably, Holz et al. (2021) asked participants to rate the dimensions of emotions by presenting the dimensions (i.e., arousal and valence) in their experiment, while these dimensions were neither mentioned nor rated in our study. Our study examined the role of valence as a binary characteristic (assigned to emotions on a theoretical basis) on intensity ratings. The difference in intensity ratings we found between negative and positive

emotions may be caused by the fact that negative emotions are more related to threats and dangers than positive emotions (Shiota et al., 2004). That negative emotions tend to be produced with higher intensity than positive emotions would be beneficial for survival.

Third, basic emotions were rated as more intense than non-basic emotions (Hypothesis 5). Ours is arguably the first study comparing intensity ratings between basic and non-basic emotions in vocal emotions, although Kommattam et al. (2019) included three non-basic emotions (secondary emotions) in intensity ratings of facial emotions. In this study, we included four basic (anger, fear, joy, sadness) and four non-basic emotions (irritation, pride, relief, tenderness). However, due to the unequal number of positive and negative emotions in the classic set of six basic emotions, we could not balance the four basic emotions (anger, fear, happiness, and sadness) in terms of arousal and valence, resulting in a non-orthogonal design in basicness. The high-intensity ratings for negative and basic emotions further demonstrate that valence and basicness are strongly correlated, playing a fundamental role in daily life.

In sum, this study investigated intensity ratings from a theoretical perspective. Our participants were directly asked to rate the level of intensity rather than rate the dimensions of arousal, valence, and basicness. Therefore, participants' ratings of intensity are based on their understanding of the concept of intensity, independent of the notions of arousal, valence, and basicness. These three dimensions, reduced to theory-based binary characteristics rather than empirically measured gradients, explained intensity ratings to different extents. Arousal and Basicness are more strongly related to intensity than to Valence, corroborating previous findings that arousal and intensity are positively correlated (Laukka et al., 2005). Although four dimensions (arousal, valence, potency, and intensity) have been reported in many studies (Laukka et al., 2005; Smith & Ellsworth, 1985), our results demonstrate that basicness is another important dimension of emotions. Additionally, we observed that although intensity ratings were affected by arousal, valence, and basicness, they varied across the eight emotions. Particularly, Anger and Joy were consistently rated as more intense than other emotions, both across all responses and in the subset of correct responses only, highlighting their pivotal role in human emotional experience.

Also, we speculated that intensity ratings were higher for correct than for incorrect identifications of the emotion type, with the size of these differences varying across the eight emotions (see Figure 3.2). In the literature, there are inconsistent results regarding the impact of intensity on recognition accuracy.

Some studies have demonstrated that emotions with higher intensity are better recognized than those with lower intensity (Bänziger et al., 2012; Hess et al., 1997; Juslin & Laukka, 2001; Livingstone & Russo, 2018; Wingenbach et al., 2016). As emotions become more intense, they are more pronounced, making them easier to identify. On the other hand, emotions with peak intensity interfere with recognition accuracy and with the rating of valence (Holz et al., 2021). Our data revealed a positive correlation between intensity and recognition accuracy, as the intensity rating was higher for correct than for incorrect ones (see Appendix E), which is consistent with the notion that emotions with stronger acoustic cues are more prominent and easier to recognize (Bachorowski & Owren, 2003; Scherer, 1986). However, due to the lack of peak emotions, we were unable to examine the (negative) effect of peak intensity on emotion recognition in the current study.

To assess whether the models had sufficient statistical power to test the in-group hypothesis, we have performed post-hoc analyses using the *simr* package in R to investigate the main effect of speaker language. The predicted power (100 simulations) was 0.95 (95% CI [0.933, 0.962]) to identify an in-group bias in Dutch listeners in the entire dataset, and 0.76 (95% CI [0.729, 0.783]) in the subset of correct responses only. The predicted power was above or close to the threshold of 0.80, indicating sufficient power to identify an in-group bias. For Korean listeners, the estimated power was 0.15 (95% CI [0.127, 0.173]) to test an in-group bias in the entire dataset, and was 0.37 (95% CI [0.339, 0.400]) to identify an in-group bias in the subset (correct responses only), suggesting that we might have found an in-group bias with a much larger sample of participants. This approach is not the right one, however, to evaluate the in-group hypothesis. Figure 3.2 shows that the speaker language lines keep crossing each other. One line is not systematically higher than the other one, and the differences between emotions are stronger than the differences between speaker language. This pattern points out that the interaction between emotions and language outweighs any consistent in-group bias. This can be deduced as well from the statistical information in Models 1 to 4 in Appendix C. The interaction effects between speaker language and emotion are stronger than the main effect of speaker language. The main effects of emotion are stronger than the main effects of speaker language. In addition, the number of participants in previous studies on intensity in acoustic cues justifies the sample size of our study. For example, Juslin and Laukka (2001) tested 15 participants and found that listeners decoded portrayals with strong intensity better than those with weak intensity. Recently, Morningstar et al. (2021) tested 190 listeners to examine the impact of intensity on the recognition of vocal socioemotional expressions. Compared to previous

studies, the total number of participants in our study (Dutch: 31, Korean: 24) falls within the range typical of vocal studies.

As a final note, the language order of presentation of stimuli constitutes a limitation of this study. In our design, the Korean recordings of 128 portrayals were presented before the Dutch recordings, and the stimuli for each participant were randomized. The effect of the sequential order on intensity ratings was minimal, as the intensity ratings of Korean and Dutch recordings did not change systematically over time. As shown in Appendix F, the mean intensity ratings across 256 stimuli, indicated as time series, display a shallow U-shaped trend, with a slight decrease in the first half and a slight increase in the second half. The variation is small given the wide range of intensity ratings (the blue and red points). Importantly, there is no clear discontinuity between the Korean and the Dutch parts.

3.5 Conclusion

Intensity, as a fundamental dimension of emotions, is an intriguing topic in affective science (Juslin & Laukka, 2001; Kommattam et al., 2019; Laukka et al., 2005; Smith & Ellsworth, 1985). However, theoretical and empirical studies on intensity, particularly in the vocal domain, remain sparse (Baum & Nowicki, 1998). To fill this gap, we investigated cross-cultural perception of intensity of vocal emotions by comparing ratings from Dutch and Korean listeners on the full set of stimuli in the Demo (Dutch speakers) and Koremo (Korean speakers) corpora. To our knowledge, this is the first study that investigates intensity ratings on vocal emotions from both discrete and dimensional approaches.

The first aim was to examine the hypothesized in-group bias in intensity ratings (Ekman et al., 1987; Kommattam et al., 2019). However, this hypothesis was not convincingly supported by our results. Instead of an in-group bias, listeners rated particular emotions as more intense even though these emotions were not produced in their native language. Therefore, intensity ratings are more dependent on emotions than on the language or culture shared by both listeners and speakers. Although there were discrepancies in intensity ratings between native and non-native listeners, the rating patterns remained basically the same. Additionally, intensity ratings are generally higher for the subset of correctly identified emotions than for the total undifferentiated response set, but correctness of the response does not substantially interact with other factors.

The second aim was to examine the effect of arousal, valence, and basicness on intensity ratings. Arousal, valence, and intensity are three of the four fundamental dimensions of emotions (Larsen & Diener, 1987; Smith & Ellsworth, 1985). The fourth dimension, potency, was not included in the study. Although basicness is not classified as one of the fundamental dimensions, it plays a pivotal role in emotions (Ekman & Cordaro, 2011). However, the extent to which intensity correlates with other dimensions turned out to be limited. The results reveal that intensity ratings were affected by arousal, valence, and basicness, such that intensity ratings were higher for high-arousal than low-arousal, higher for negative than positive, and higher for basic than non-basic emotions. However, none of these dimensions affords a sharp dichotomy of the eight emotions in intensity, since intensity ratings for certain emotions cannot be reliably predicted from the general patterns.

Our results partially replicate earlier findings and provide new insights into intensity ratings of vocal emotions. Further studies are needed to explore additional emotional characteristics (e.g., potency) in the exploration of determinants of perceived emotional intensity.

Chapter Four

Classifying emotions cross-linguistically from acoustic parameters²⁰

Abstract

This study acoustically measured a total of 256 vocal emotions (8 emotions \times 8 actors \times 2 tokens \times 2 languages) from the Demo/Koremo corpus, produced in two typologically different languages—Dutch and Korean. All emotions were analyzed acoustically according to 17 acoustic parameters divided into five categories: pitch, amplitude, spectral, duration, and perturbation. Linear mixed-effects models were used to analyze the emotion-specific acoustic patterns across language and gender. The results revealed that the eight emotions displayed distinct acoustic patterns, which varied between female and male actors and were highly language-dependent. Support Vector Machine (SVM) models were then employed to examine the extent to which machine classifiers mimic human listeners. The results demonstrated that the machine classifiers can identify vocal emotions successfully, especially in within-group conditions, which, in turn, agrees with human performance.

Keywords: Dutch, Korean, vocal emotions, acoustic parameters, recognition accuracy, Support Vector Machine (SVM)

²⁰ Liang, Y., van Heuven, V., & van Hout, R. (Submitted). Classifying emotions cross-linguistically from acoustic parameters.

4.1 Introduction

The voice, as carrier of vocal emotions (Darwin, 1998), conveys a wealth of information about the speaker beyond the literal meanings of words (Banse & Scherer, 1996; Johnstone & Scherer, 2000; Mozziconacci & Hermes, 1999). In the realm of vocal emotion communication, variations in acoustic parameters are referred to as *emotional* (or *affective*) *prosody* (Frick, 1985), which distinguishes emotions and affects how information is produced and perceived (Banse & Scherer, 1996; Juslin & Laukka, 2001; Mozziconacci, 2002; Murray & Arnott, 1993; Scherer, 1986). The overarching goal of this study is to examine language-specific acoustic features in vocal emotions via comparing two typologically different languages, i.e., Dutch and Korean. In this introduction, we discuss research on the acoustic characteristics of the vocal expressions of emotions, and we discuss classifying these emotions given their acoustic characteristics, before arriving at the cross-language perspective of our study and the hypotheses we want to test.

4.1.1 Acoustic characteristics of vocal expression of emotions

Emotions are conveyed via a constellation of acoustic parameters, known as cue configuration, which are used to describe discrete vocal emotions (Banse & Scherer, 1996; Juslin & Laukka, 2003; Laukka et al., 2016; Pell, Monetta, et al., 2009; Scherer & Oshinsky, 1977). These acoustic parameters include, but are not limited to, fundamental frequency (f_0), intensity, formants, tempo, and perturbation (e.g., unsteadiness, breathiness), resulting in particular acoustic patterns for specific emotions (Scherer & Oshinsky, 1977). For example, sadness is expressed with a comparatively low pitch and a slow speech rate, whereas anger and happiness are conveyed by a higher pitch with a fast speech rate (Banse & Scherer, 1996; Juslin & Laukka, 2003). It is acknowledged that pitch, intensity, tempo, and voice quality greatly affect vocal emotion recognition (Juslin, 2000; Scherer & Oshinsky, 1977).

In the past decades, empirical studies have demonstrated that listeners can identify vocal emotions from acoustic parameters above chance across cultures, with higher recognition accuracy for emotions produced in their native language (the so-called in-group advantage) or languages typologically similar to their native language (Pell, Monetta, et al., 2009; Scherer et al., 2001; Thompson & Balkwill, 2006; Van Bezooijen, 1984). The above-chance recognition accuracy indicates that there exists a set of acoustic cues that can be recognized across languages/cultures, but language-specific features cause the in-group advantage.

To examine the role of linguistic factors in vocal emotion recognition, a number of studies have examined the effect of acoustic parameters on vocal emotions (Juslin & Laukka, 2001, 2003; Mozziconacci & Hermes, 1997; Murray & Arnott, 1993; Scherer, 1986, 1989). The parameters can be broadly divided into four categories: f_0 (or pitch), intensity (loudness), tempo (speech/articulation rate, pauses), and voice quality (timbre) (Banse & Scherer, 1996; Johnstone & Scherer, 2000; Murray & Arnott, 1993; Pell, Paulmann et al., 2009; Van Bezooijen, 1984; see Bachorowski & Owren, 2003 for a review). The influential study by Banse and Scherer (1996) measured 29 acoustic parameters across 14 emotions produced by 12 German voice actors. Their findings revealed the pivotal role of f_0 and intensity in vocal emotion expressions. Other acoustic parameters, formants, and tempo (including pauses), also characterize the production of vocal emotions. Anolli et al. (2008) examined vocal emotions produced by Chinese and Italian speakers. They proposed that emotions are conveyed via variations of acoustic cues in both populations, but with subtle differences in acoustic patterns observed between speakers from these two cultures. Likewise, Juslin and Laukka (2001) observed that listeners' ratings of emotional intensity (strength) are reliably predictable from acoustic parameters. Jointly, there is a close relationship between acoustic parameters and vocal emotions, such that vocal emotions are characterized by specific acoustic patterns (Bachorowski & Owren, 2008; Banse & Scherer, 1996; Darwin, 1998; Scherer, 1986), and f_0 - and intensity-related parameters are stronger correlates of vocal emotions than other parameters (Banse & Scherer, 1996).

Along this line, Pell, Monetta, et al. (2009) investigated the influence of language on vocal emotions by examining how six basic emotions are identified and acoustically characterized in four different languages (Arabic, English, German, and Hindi). They found that all emotions were recognized above chance, with some particular emotions having higher recognition accuracy than others. The acoustic analysis highlighted the significant role of (mean and variability of) f_0 in signaling vocal emotions across languages (Banse & Scherer, 1996; Pell, Paulmann, et al., 2009).

Moreover, the gender of speakers and individual differences also influence vocal emotion recognition (Bachorowski & Owren, 2008; Matsumoto, 2006). Gender is a biological factor affecting the production and perception of emotional speech (Bonebright et al., 1996; Lausen & Schacht, 2018). Typically, due to differences in the anatomy and physiology of the vocal organs, female and male voices differ in terms of fundamental frequency and formant frequency (Collins, 2000; Klatt & Klatt, 1990; Latinus & Taylor, 2012; Titze, 1989). A few studies have investigated the impact of decoders' gender

on the perception of vocal emotions and found that females are more sensitive to emotional prosody than males (Keshtiari & Kuhlmann, 2016; Vigil, 2009). However, research on the role of gender in the production of vocal emotions remains scarce (Scherer et al., 1991).

4.1.2 Classifying vocal emotions

Based on the extracted acoustic parameters, research in human-computer interaction (HCI) has proposed different machine classifiers, such as Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM), for automatic recognition of emotion (Ezhilarasi & Minu, 2012; Lee & Narayanan, 2005; Luengo et al., 2005; Pallewela et al., 2024, see Ververidis & Kotropoulos, 2006 for a review). These studies have collectively demonstrated that it is possible to classify vocal emotions according to a series of acoustic features. Classification rates vary depending on the machine learning model, corpora, and the number of acoustic features (Cowie et al., 2001; Laukka et al., 2011; Ververidis & Kotropoulos, 2006).

Despite the well-established knowledge of the descriptions of specific acoustic patterns associated with different emotions, discrepancies remain (Banse & Scherer, 1996). First, although acoustic parameters have been extensively studied in Indo-European languages (e.g., English, German) (Dromey et al., 2005; Scherer et al., 2001; Van Bezooijen, 1984), and to some extent in non-Indo-European languages, such as Chinese (Li et al., 2023), Japanese (Lubis et al., 2016), and Arabic (Meddeb et al., 2016), research on the acoustic measurement of Korean remains scarce. As a result, current knowledge of acoustic parameters and machine learning models for vocal emotions may not be generalizable to the underrepresented language—Korean. Second, the existing studies have primarily focused on basic emotions produced by a small number of speakers, resulting in an unbalanced representation of arousal and valence (Banse & Scherer, 1996; Laukka et al., 2005; Sauter et al., 2010). Consequently, the acoustic characteristics of non-basic emotions are underexplored. This limitation underscores the need for new studies to include a larger number of emotions, especially non-basic emotions. Therefore, the present study compares acoustic differences of eight vocal emotions, including four non-basic ones, in two typologically different languages—Dutch and Korean (a less-studied language), aiming to find out the similarities and differences of 17 acoustic parameters across gender, speaker language, and emotion. We assess whether human classification rates can be reliably predicted from acoustic parameters by Support Vector Machine (SVM) models.

4.1.3 The cross-language perspective of our study and the hypotheses

The overarching goal of this study on vocal emotions is to examine language-specific acoustic characteristics for two typologically different languages, i.e., Dutch and Korean. Dutch is a stress-timed language, while Korean is syllable-timed. Specifically, Dutch allows complex syllables with up to three onset consonants, up to four coda consonants, a vowel-length contrast, and reduction to schwa in unstressed syllables, placing Dutch at the stress-timed pole of the typological rhythm continuum. Moreover, Dutch has word stress, such that each word has its own fixed stress position, which (Jun & Fougeron, 2000; Van Heuven et al., 2008) is always marked by lengthening and/or a prominence-lending pitch change (Van Heuven, 2018). Korean has no word stress but phrasal stress, and the last syllable in the Accentual Phrasal (AP) must have a high pitch target (H), which can be construed as either a pitch accent or a boundary tone (this cannot be decided in a language with phrasal stress, which is also the claim made for prosodically similar languages such as French, Malay and Indonesian (see e.g., Jun & Fougeron, 2000; Van Heuven et al., 2008; Maskikit-Essed & Gussenhoven, 2016). If a word in Korean is not the last word in the AP, then there is no prosodic mark at all (Jun, 2006).

In our study on acoustic parameters, we address the following five issues and their concomitant hypotheses.

1. We examine the effect of emotion on variations in acoustic parameters. According to the literature reviewed above, the acoustic properties of vocal emotions, such as f_0 , intensity, and speech rate, vary systematically with emotional states (Banse & Scherer, 1996; Breitenstein et al., 2001). Therefore, we hypothesize that the eight emotions we investigate (Anger, Fear, Irritation, Joy, Pride, Relief, Sadness, and Tenderness) will display different acoustic patterns (Hypothesis 1), depending on the language.
2. We examine the effect of speaker language on variations in acoustic parameters. These parameters differ across languages and cultures (Scherer et al., 2001). Therefore, we hypothesize that acoustic parameters will exhibit distinct acoustic patterns in different languages (Hypothesis 2).
3. We examine the effect of gender on the variations of acoustic parameters. Since vocal cues differ between females and males (Klatt & Klatt, 1990), gender will affect acoustic parameters, though patterns of variation will depend on the specific acoustic parameter and emotion (Hypothesis 3).
4. We investigate whether vocal emotions can be accurately classified automatically based on acoustic parameters.

- a. Hypothesis 4a: When we have appropriate acoustic parameters, an automatic classification algorithm (machine learning) will perform above chance level, for both speaker languages (in-group condition).
 - b. Hypothesis 4b: When we apply the solution for one language cross-culturally to the other, the results will be lower but still above chance level (out-group condition).
5. We examine the extent to which machines and human listeners differ in their use of acoustic parameters in recognizing vocal emotions. Specifically, we test whether machines can mimic human listeners if provided with the necessary information. We hypothesize that machines and human listeners use acoustic parameters similarly when classifying vocal emotions (Hypothesis 5).

4.2 Method

4.2.1 Stimuli

All stimuli were from the existing Demo/Koremo (Dutch emotion/Korean emotion) corpus (Broersma et al., 2025).²¹ The corpora include a total number of 256 portrayals (8 emotions × 8 speakers × 2 tokens × 2 languages). The stimuli were balanced in terms of arousal and valence, with an equal number of basic and non-basic emotions (Table 4.1). All stimuli were portrayed using the same pseudo-sentence /nuto hɔm sɛpikaŋ/.²² For more information about the corpora, we refer to Liang et al. (2025; Chapter 2).

²¹ The scenarios and corpus are publicly available via Radboud University at <https://doi.org/10.34973/5kg3-9852>

²² According to Goudbeek and Broersma (2010a, b), the pseudo-sentence /nuto hɔm sɛpikaŋ/ is phonologically legal in both Dutch and Korean. However, the rhyme /aŋ/ is not allowed in Dutch (but /aŋ/ is). Furthermore, there is no low-mid vowel /ɔ/ in Korean (Shin, 2015); the substitute is high-mid /o/. Therefore, Dutch and Korean voice actors used slightly different vowel sounds, which were compatible with their native language.

Table 4.1. The eight emotions used in the current study in a valence-by-arousal grid (reproduced from Goudbeek & Broersma, 2010b, p. 2212), with basic emotions marked with “*”.

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

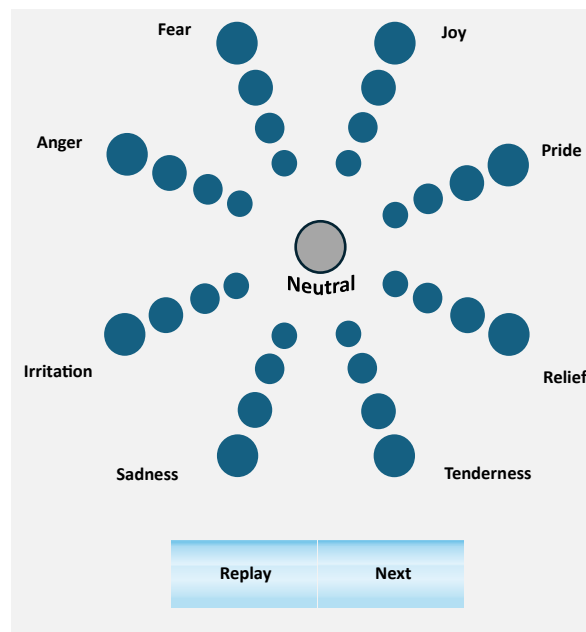


Figure 4.1. The emotion wheel in English (reproduced after Liang et al., 2023). Translation in Dutch and Korean, Joy: “Blijdschap”, “행복”; Pride: “Trots”, “자랑스러움”; Relief: “Opluchting”, “안도감”; Tenderness: “Vertederling”, “애정”; Sadness: “Verdriet”, “슬픔”; Irritation: “Irritatie”, “짜증”; Anger: “Woede”, “분노”; Fear: “Angst”, “공포”; Neutral: “Neutral”, “중립”.

4.2.2 Acoustic analysis of the speech stimuli

A total number of 256 stimuli were acoustically analyzed with different acoustic parameters in Praat (Boersma & Van Heuven, 2001). In addition to the classic acoustic parameters, i.e., f0, intensity, speech/articulation rate, and spectral distribution (Banse & Scherer, 1996; Breitenstein et al., 2001; Laukka et al., 2005; Sauter et al., 2010a), we included three additional acoustic parameters: jitter, shimmer, and harmonics-to-noise ratio, which may be related to the influence of perturbations on the human perception of vocal emotions, resulting in a total of 17 acoustic parameters (Table 4.2).

Table 4.2. Overview of the acoustic parameters measured in this study.

| Acoustic parameter | Abbreviation | Perceptual cue |
|--|--------------|--------------------------|
| <i>Pitch parameters</i> | | |
| Mean fundamental frequency (semitones, st) | F0-M | Mean pitch |
| Standard deviation of fundamental freq. (st) | F0-SD | Pitch variability |
| Minimum fundamental frequency (st) | F0-min | Minimum pitch |
| Maximum fundamental frequency (st) | F0-max | Maximum pitch |
| Fundamental frequency slope (st/s) | Slope | Pitch slope |
| Synchronization | Sync | Location of pitch change |
| <i>Amplitude parameters</i> | | |
| Mean intensity (dB) | Int-M | Mean loudness |
| Standard deviation of intensity (dB) | Int-SD | Loudness variability |
| Spectral slope (dB/oct) | Tilt | Mean loudness |
| <i>Spectral parameters</i> | | |
| Formant 1 (Bark) | F1 | Voice quality |
| Formant 2 (Bark) | F2 | Voice quality |
| Formant 3 (Bark) | F3 | Voice quality |
| <i>Duration parameters</i> | | |
| Articulation rate (syll/s, minus pause) | AR | Tempo, fluency |
| Pause rate (% Pause) | PR | Tempo, fluency |
| <i>Laryngeal parameters</i> | | |
| Jitter (PPQ5) | PPQ | Pitch instability |
| Shimmer (APQ5) | APQ | Loudness instability |
| Harmonicity (dB) | HNR | Breathiness |

In the following section, we will present an overview of the acoustic parameters measured in this study and explain how they are measured.

- **Fundamental frequency:** Fundamental frequency (F0) is the frequency of the vibration (opening and closing) of the vocal folds during speech

(Scherer et al., 1991), which is measured in hertz (Hz). The higher the F0, the higher the perceived pitch. Slow variations in fundamental frequency are perceived as changes in pitch, but rapid changes in fundamental frequency may be perceived as jitter (instability of the voice). F0 and derived parameters such as F0 range and F0 contour are correlated with the production and perception of vocal emotions (Banse & Scherer, 1996; Cosmides, 1983; Paeschke et al., 1999; Sauter et al., 2010; Van Bezooijen, 1984). For instance, neutrality is characterized by relatively low F0 and gradual changes in pitch (Carl et al., 2022; Williams & Stevens, 1972). Compared to neutrality, anger is featured by faster and larger F0 changes, whereas sadness is marked by slow and small F0 changes (Banse & Scherer, 1996; Juslin & Laukka, 2001). Four F0-related parameters, i.e., F0-M, F0-SD, F0-min, and F0-max, were extracted from each token directly by Praat, and expressed in semitones (re. 50 Hz), a logarithmic scale that better aligns with human pitch perception. Pitch tiers were checked visually, and errors were corrected manually. F0-SD reflects the variability of pitch of each token. Since F0-M and F0-SD provide no information on the shape of the pitch change in the speech fragments, we took slope into consideration. Slope was computed by dividing the difference between F0-max and F0-min (in semitones, st) by the time interval between these two. If F0-max is larger than F0-min, the pitch slope is positive, indicating a rising pitch pattern, and vice versa. F0-M is the average pitch of the entire speech signal. F0-min is the lowest point of the pitch (in st) across the entire speech utterance, while F0-max is the highest point of the pitch (in st). Sync is the temporal midpoint of the pitch slope, which is computed as a percentage of the duration of the emotion token, indicating whether the pitch change comes early or late in the token. It reflects the timing of pitch changes across the speech signal.

- **Intensity:** Intensity, measured in decibels (dB), is the energy level in speech production, which is (all else equal) perceived as loudness (Bachorowski & Owren, 2003). The perception of loudness also depends on the general distribution of intensity over the sound spectrum. Typically, as the intensity rises from low to high frequencies, greater vocal effort expended by the speaker is perceived, so that the speech sounds louder (Sluijter & Van Heuven, 1996). Different emotions exhibit different levels of intensity. Acoustic intensity is different from emotional intensity. Emotional intensity is the strength of emotions experienced by individuals (For instance, intensely expressed sadness will have low acoustic intensity). Intensity is difficult to measure since it is affected by the distance between the speaker's mouth and the microphone, as well as the equipment used for recordings (Scherer et al., 1991). To capture the nature of intensity, we measured several intensity-related parameters: Mean

intensity (Int-M), the standard deviation of the intensity (Int-SD), and Tilt. Int-M is expressed in decibels (dB). Int-SD is the variability of the intensity of the voice. Spectral tilt (Tilt) is the overall increase or decrease of intensity from low to high frequencies in the sound spectrum. It captures the general trend in the distribution of energy over the spectrum. Tilt is expressed in either decibels per hertz (a linear scale) or in decibels per octave (a logarithmic scale).

- **Formants:** Formants are used to describe the shape of the resonance characteristics of the vocal tract. Formants are concentrations of acoustic energy caused by resonance in the vocal tract. They are characterized by their center frequency, amplitude, and bandwidth. In emotion studies, the measurements are typically restricted to only the center frequencies of the lowest two or three formants (F1, F2, F3). The center frequencies of these formants affect the sound quality, and thereby the perception of vocal emotions (Scherer, 1986). Formants can be estimated by Praat. To better reflect the auditory effects of different formant frequencies, we applied perceptual scaling by transforming formant frequencies in hertz to perceptually defined Barks (Traunmüller, 1990), as is customary in vowel quality studies.
- **Articulation rate:** Articulation rate (AR) and speech rate (SR) are measures of the pace of speaking. Both AR and SR are expressed in terms of the number of linguistic units (sounds, syllables, or words) per unit time (seconds or minutes). In the computation of AR, the duration of (silent) pauses and (filled) hesitations is not included (Laver, 1994), whereas these elements are included in SR (Buller, 2005; Robb et al., 2009). We argue that SR is a compound measure, which we decompose into its constituents: AR and speech/pause ratio. The latter parameter (PR) is conveniently expressed as the proportion of pauses (silent and filled) in the total duration of the speech materials on which SR is based.
- **Jitter:** Jitter measures cycle-to-cycle fluctuations in the f_0 of the speech signal, assessing subtle instabilities of the voice (Farrús et al., 2011; Zamek & Zamek, 2005). An often-used implementation is the Jitter Period Perturbation Quotient (ppq5). This is the mean absolute difference between the middle period in a symmetrical window of five adjacent periods and the mean of all periods in the window, as the window is moved in one-period steps from the beginning to the end of the utterance (Farrús et al., 2011).
- **Shimmer:** Shimmer measures the peak-to-peak changes in the amplitude of the vocal period, evaluating the irregularity and instability of the amplitude (Farrús et al., 2011). Higher levels of shimmer reflect less stability in the voice. Shimmer is closely related to vocal emotions. For instance, emotions like anger and sadness display higher levels of

shimmer, indicating subtle fluctuations in intensity during vocalizations. However, stable emotions like neutrality and calm exhibit lower levels of shimmer, suggesting slow period-to-period variations in intensity during vocalizations. Shimmer (apq5) refers to the five-point Amplitude Perturbation Quotient, calculated as ppq5 above, substituting amplitude (or intensity) for frequency of the period (Gorris et al., 2020).

- **Harmonics-to-noise ratio:** Harmonics-to-noise ratio (HNR), measured in decibels (dB), evaluates the relationship between harmonic and noise components in the voice (Yumot & Gould, 1982). It correlates with voice quality, specifically breathiness. A higher level of HNR indicates a clearer sound with less noise. A lower level of HNR indicates a rather more whispery sound with more noise.

4.2.3 Analysis

We adopted both linear mixed-effects models and Support Vector Machine algorithms to analyze acoustic parameters of vocal emotions across Dutch and Korean speakers. First, we employed linear mixed-effects models to examine how acoustic parameters varied across Emotion, Speaker Language, and Gender (Hypotheses 1-3). Second, Support Vector Machine classifiers were trained to predict vocal emotions according to acoustic parameters (Hypotheses 4a & 4b). In the Results section, we summarized the main results and interpretations. Detailed statistical tables are provided in Appendix G. Plots of the effects of emotion, language, and gender can be found in Appendix H.

4.3 Results

The data analyses were performed in R (R Core Team, 2023) using the *lme4* package (Bates et al., 2015). To address Hypotheses 1 to 3, first, we built a series of linear mixed-effects models for the whole dataset, including both Dutch and Korean recordings. Each model included Emotion (Joy, Pride, Anger, Fear, Tenderness, Relief, Sadness, and Irritation), Speaker Language, and Gender as predictors, and each of the acoustic parameters as the outcome variable. For the predictor Emotion, we used Fear as the reference, since it had the largest value in most of the acoustic parameters. Second, to eliminate the impact of Speaker Language on acoustic parameters, we split the whole data into two subsets—Dutch and Korean recordings. We performed a number of linear mixed-effects models on the Dutch and Korean subsets, respectively, with Gender and Emotion (Fear is the reference) as fixed variables, and each of the acoustic parameters as the outcome variable. However, the models with

Slope as the outcome variable did not converge in the whole data and Korean subsets, and the model with Speech Pause Ratio (PR) as the outcome variable did not exist because there were no pauses in the Korean subset. As a result, there were 48 models in total. To understand individual differences in the production of vocal emotions (Matsumoto, 2006), we included Speaker as the random effect in all models. For the model selection, we adopted the backward elimination method, such that each model started with the maximal structure including fixed predictors, random effects, and random slopes for the predictors. In case of non-convergence, models were simplified by removing the random slopes with the smallest variance (Barr et al., 2013). To choose the best-fitting model, we used ANOVA to compare the model fit and selected the one with the lowest AIC score.

To address Hypotheses 4a and 4b, we built Support Vector Machines (SVMs) to classify emotions based on acoustic parameters in either the Dutch or Korean subset (without Gender as a predictor), and then applied the resulting SVM models once on the same language dataset they were trained on (“in-group”, with appropriate cross-validation) and a second time on the emotional portrayals of the other language dataset (“out-group”, no cross-validation needed). To eliminate individual and gender differences between speakers, we transformed the raw acoustic parameters to speaker-individual z -scores (Banse & Scherer, 1996). This method yielded four SVM models, with a confusion matrix for each model.

Finally, to address Hypothesis 5, we compared the recognition accuracy produced by SVM models with confusion matrices obtained from human listeners. If the confusion structure (and by implication also the accuracy) is similar between the machine and human listeners, both probably use the same acoustic cues to recognize vocal emotions, while discrepancies indicate potential differences in cue utilization.

4.3.1 Effects of emotion, speaker language, and gender (Hypotheses 1-3)

The first research question examined the effect of Emotion, Speaker Language, and Gender on acoustic parameters. We hypothesized that each acoustic parameter would differ depending on emotion, speaker language, and gender.

As shown in Tables 4.3A-B-C, neither Slope nor PR was affected by Emotion, Speaker Language, or Gender. Therefore, we excluded Slope and PR, keeping 15 acoustic parameters for the subsequent analyses. To better illustrate the differences of the 15 acoustic parameters between Dutch and Korean, we created radar plots after z -normalizing the data within each individual speaker (Figure 4.2). We applied normalization to abstract from speaker-dependent

differences. For instance, female voices have twice the F0 of male voices, and there are also large differences in mean pitch (and variability) between voices of the same gender. These differences were handled in the linear mixed-effects models by including random effects for Speaker. However, such adjustments are difficult to visualize in the radar plots. Moreover, we want to ensure all parameters are represented on comparable scales and units, allowing for direct visual comparison of the effect sizes across different parameters. Notably, speaker normalization (within-speaker-transformation or Lobanov normalization, Lobanov, 1971) is a standard procedure in phonetic studies (e.g., Wang & Van Heuven, 2018).

We summarized the effects found by the linear mixed-effects models in Table 4.3. We discuss them separately for the five parameter groups.

Table 4.3. A. Summary of linear mixed-effects regression analysis of effects and interactions of *Emotion* (E), *Speaker Language* (SL), and *Gender* (G) on 17 acoustic variables. B-C. Main effects and interactions are separated by SL. Significant effects and interactions are identified by asterisks. *: $p < .05$, **: $p < .01$, ***: $p < .001$.

| A. All recordings combined | | | | | | | |
|----------------------------|-----|-----|-----|-----|------|------|--------|
| Acoustic parameter | E | SL | G | E×G | E×SL | SL×G | E×SL×G |
| F0-M | *** | *** | *** | | *** | | |
| F0-SD | *** | | | | | | |
| F0-min | *** | * | *** | | *** | | |
| F0-max | *** | *** | *** | | *** | | |
| Slope | | | | | | | |
| Sync | * | | * | | | | |
| Int-M | *** | | | | *** | | |
| Int-SD | *** | *** | | | *** | | |
| Tilt | ** | ** | | *** | *** | | ** |
| F1 | *** | | | * | | *** | *** |
| F2 | | *** | ** | | *** | | |
| F3 | * | *** | | * | *** | | ** |
| AR | *** | *** | * | | *** | | |
| PR | | | | | | | |
| PPQ5 | * | *** | ** | | *** | * | *** |
| APQ5 | * | * | * | ** | *** | * | *** |
| HNR | *** | ** | ** | | *** | * | *** |

| Acoustic parameter | B. Dutch recordings | | | C. Korean recordings | | |
|--------------------|---------------------|-----|-----|----------------------|-----|-----|
| | E | G | E×G | E | G | E×G |
| F0-M | *** | *** | | *** | ** | |
| F0-SD | *** | * | | *** | | ** |
| F0-min | *** | ** | | *** | *** | *** |
| F0-max | *** | ** | | *** | ** | |
| Slope | | | | | | |
| Sync | * | | | * | | |
| Int-M | *** | | | *** | | ** |
| Int-SD | *** | * | | *** | | |
| Tilt | *** | * | *** | *** | | |
| F1 | ** | ** | *** | *** | *** | *** |
| F2 | *** | * | | *** | * | |
| F3 | ** | | ** | *** | | ** |
| AR | | | * | *** | | |
| PR | | | | | | |
| PPQ | ** | | ** | *** | ** | |
| APQ | *** | | *** | ** | *** | |
| HNR | *** | | *** | *** | * | |

Pitch parameters

- **F0-M:** We found a significant main effect of Emotion in each dataset. In all data, F0-M was lower for Pride, Irritation, Relief, Sadness, and Tenderness than for Fear, Anger, and Joy. Furthermore, there was a significant main effect of Speaker Language on F0-M in all data, with Dutch speakers exhibiting a higher F0-M than Korean speakers. Additionally, the analyses across the three datasets consistently revealed a significant main effect of Gender on F0-M, such that F0-M was higher in female than in male actors. Importantly, the two-way interaction between Emotion and Speaker Language reached statistical significance, indicating that the influence of Emotion on F0-M differed by Speaker Language. Specifically, Fear and Anger were related to higher F0-M in Dutch actors than in Korean actors. However, Pride, Sadness, and Tenderness displayed slightly higher F0-M in Korean than in Dutch, whereas Irritation and Relief exhibited similar F0-M in both languages.
- **F0-SD:** We observed a significant main effect of Emotion on F0-SD in each dataset. In all data, the variability in F0 was significantly greater for Pride, Anger, Irritation, and Joy than in Fear. This pattern was found in the Dutch recordings, where the pitch variability was also greater in Pride, Anger, Irritation, and Joy than in Fear. In the Korean recordings, the variability in F0 was greater in Anger and Irritation than in Fear. However, there was no significant main effect of Speaker Language on F0-SD in all data, indicating that F0-SD did not differ between Dutch and Korean actors. Additionally, the analyses revealed no significant effect of Gender on F0-SD across each dataset, showing that the variability in F0 did not differ significantly between female and male actors. Furthermore, there was a significant two-way interaction between Emotion (Tenderness) and Gender. More importantly, we observed a significant three-way interaction between Emotion, Speaker Language, and Gender. Specifically, the variability in F0 for Pride and Tenderness varied more in males than in females in the Dutch recordings, whereas the reverse was seen in the Korean recordings. F0-SD for Fear was similar between males and females in the Dutch recordings, whereas it varied more in males than in females in Korean.
- **F0-min:** There was a significant main effect of Emotion on F0-min in each dataset. Specifically, F0-min was consistently lower for emotions—Joy, Anger, Sadness, Relief, Irritation, Pride, and Tenderness than for Fear. Furthermore, the analyses revealed a significant main effect of Speaker Language on F0-min, with Dutch actors displaying higher F0-min than Korean actors. Also, we found a significant main effect of Gender on F0-min across all datasets, with females showing higher F0-min than males. Importantly, there was a significant two-way interaction between Emotion

and Speaker Language, such that F0 min was higher in Anger, Sadness, Relief, Irritation, and Pride produced in Korean than in Dutch, whereas it was slightly higher in Tenderness uttered in Dutch than in Korean.

- **F0-max:** There was a significant main effect of Emotion on F0-max in each dataset. Furthermore, the analysis yielded a significant main effect of Speaker Language on F0-max in all data, with Dutch actors displaying higher F0-max than Korean speakers. Additionally, we observed a significant main effect of Gender on F0-max across all datasets, with female actors producing higher F0-max than their male counterparts. Importantly, there was a two-way interaction between Emotion and Speaker Language. Specifically, F0-max for Pride, Irritation, and Relief was higher in Dutch than in Korean, while F0-max for Sadness was higher in Korean than in Dutch.
- **Sync:** There was a significant main effect of Emotion on Sync in all data. However, the effect of Speaker Language on Sync did not reach statistical significance. Furthermore, the analyses revealed a significant main effect of Gender on Sync in all data, with higher sync in females than in males, indicating that the pitch crossover from higher to lower (or vice versa) occurs later in the utterance in females than in males. However, this effect was not significant in the subset with Dutch recordings.

Amplitude parameters

- **Int-M:** There was a significant main effect of Emotion on Int-M. In all data, compared to Fear, the Int-M was significantly higher for Joy and Anger, while it was lower for Pride, Relief, Irritation, Sadness, and Tenderness. Specifically, in the Dutch recordings, Int-M for Fear was higher than for Pride, Relief, Irritation, Sadness, and Tenderness. In the Korean recordings, Int-M for Joy and Anger was higher than for Fear. However, the model yielded no significant main effect of Speaker Language on Int-M, as it was similar in both Dutch and Korean recordings. However, there was no significant main effect of Gender in any dataset, indicating that the loudness of vocal emotions did not differ between females and males. Additionally, there was a significant two-way interaction between Emotion (Pride, Anger, Tenderness, Relief, Sadness, and Irritation) and Speaker Language. Notably, we found a significant three-way interaction between Emotion, Speaker Language, and Gender, indicating that the two-way interaction between Emotion and Speaker Language was modulated by Gender. Specifically, in the Dutch recordings, Int-M of Fear was higher in females than in males, whereas in the Korean recordings, it was slightly higher in males than in females. However, the Int-M of Sadness was higher in females than in males in Dutch, while it was higher in males than in females in Korean.

- **Int-SD:** There was a significant main effect of Emotion on Int-SD across each dataset, revealing that the pattern of intensity deviations varied across emotions. Particularly, intensity variability was significantly greater in Anger than in Fear, while Fear displayed greater intensity variability than Tenderness. Furthermore, there was a significant main effect of Speaker Language on Int-SD in all data, such that the intensity deviation was greater in Dutch than in Korean. Additionally, the effect of Gender on Int-SD reached statistical significance only in the subset with Dutch recordings, where female actors exhibited larger intensity deviations than male actors. Notably, there was a significant two-way interaction between Emotion (Tenderness, Relief, and Sadness) and Speaker Language. Specifically, Int-SD of Tenderness was higher in Dutch than in Korean, whereas Int-SD of Relief and Sadness was higher in Korean than in Dutch.
- **Tilt:** There was a significant main effect of Emotion on Tilt across each subset. Moreover, the analyses showed a significant main effect of Speaker Language on Tilt in all data, with more negative slopes in Korean than in Dutch. Additionally, we observed a significant main effect of Gender on Tilt in the subset of Dutch recordings, indicating that males had a more negative slope than females. A steeper negative slope is associated with a relaxed, sonorous voice, while a flatter or even positive tilt, which is featured by high vocal effort (stress), shrillness, and shouting. Further analyses revealed a significant two-way interaction between Emotion and Gender in all data and in the subset with Dutch recordings. Additionally, there was a significant two-way interaction between Emotion (Joy, Pride, Anger, Tenderness, Relief, Sadness, and Irritation) and Speaker Language in all data. Importantly, there was a significant three-way interaction between Emotion (Pride and Irritation), Speaker Language, and Gender. Specifically, Tilt for Fear was more negative in females than in males in the Dutch recordings, while it was similar between males and females in the Korean recordings. Also, Tilt for Pride was more negative in males than in females in Dutch, while it was more negative in females than in males in Korean. Tilt for Irritation was more negative in males than in females in both recordings, but steeper in Dutch than in Korean.

Spectral parameters

- **F1:** We found a significant main effect of Emotion on F1 in each dataset. However, there was no significant main effect of Speaker Language on F1, as F1 did not differ significantly between Dutch and Korean. Moreover, there was a significant main effect of Gender on F1 in both the Dutch and Korean subsets. In the Dutch recordings, females displayed higher F1 than males, whereas in the Korean recordings, males exhibited higher F1 than

females. Additionally, there was a significant two-way interaction between Speaker Language and Gender, with Korean male actors showing higher F1 than Korean female actors, while Dutch female actors had higher F1 than Dutch male actors. Moreover, we found a three-way interaction between Emotion (Pride, Anger, Tenderness, Relief, and Sadness), Speaker Language, and Gender. Specifically, in Dutch recordings, there was a significant two-way interaction between Emotion (Pride, Anger, Tenderness, Relief, Sadness, and Irritation) and Gender. In Korean recordings, there was a significant two-way interaction between Emotion (Anger, Tenderness, Relief, and Sadness) and Gender.

- **F2:** Although there was no significant main effect of Emotion on F2 in all data, there was a significant main effect of Emotion on F2 in the Dutch and Korean subsets. Moreover, there was a significant main effect of Speaker Language on F2 in all data, suggesting that F2 was higher in Korean than in Dutch. Also, the analyses yielded a significant main effect of Gender on F2 in each subset, such that F2 was higher in male than in female voice actors. Additionally, there was a significant two-way interaction between Emotion and Speaker Language in all data. Specifically, F2 of Fear and Pride was higher in Korean than in Dutch, while F2 of Tenderness and Irritation was higher in Dutch than in Korean, and F2 of Relief and Sadness was similar in Dutch and Korean.
- **F3:** We observed a significant main effect of Emotion on F3 in each dataset. Moreover, there was a significant main effect of Speaker Language on F3 in all data, such that F3 was higher in Korean than in Dutch. However, the analyses revealed no significant main effect of Gender on F3 in each dataset, indicating that F3 was similar between female and male actors. Additionally, there were significant two-way interactions between Emotion and Gender across each dataset, and between Emotion and Speaker Language in all datasets. Importantly, the three-way interaction between Emotion, Speaker Language, and Gender reached statistical significance, showing that F3 of Fear was slightly higher in females than in males in Dutch, while it was higher in males than in females in Korean. Also, F3 of Anger and Tenderness was similar in females and males in Dutch, whereas it was slightly higher in females than in males in Korean. However, F3 of Sadness was slightly higher in males than in females in Dutch, whereas it was similar in females and males in Korean.

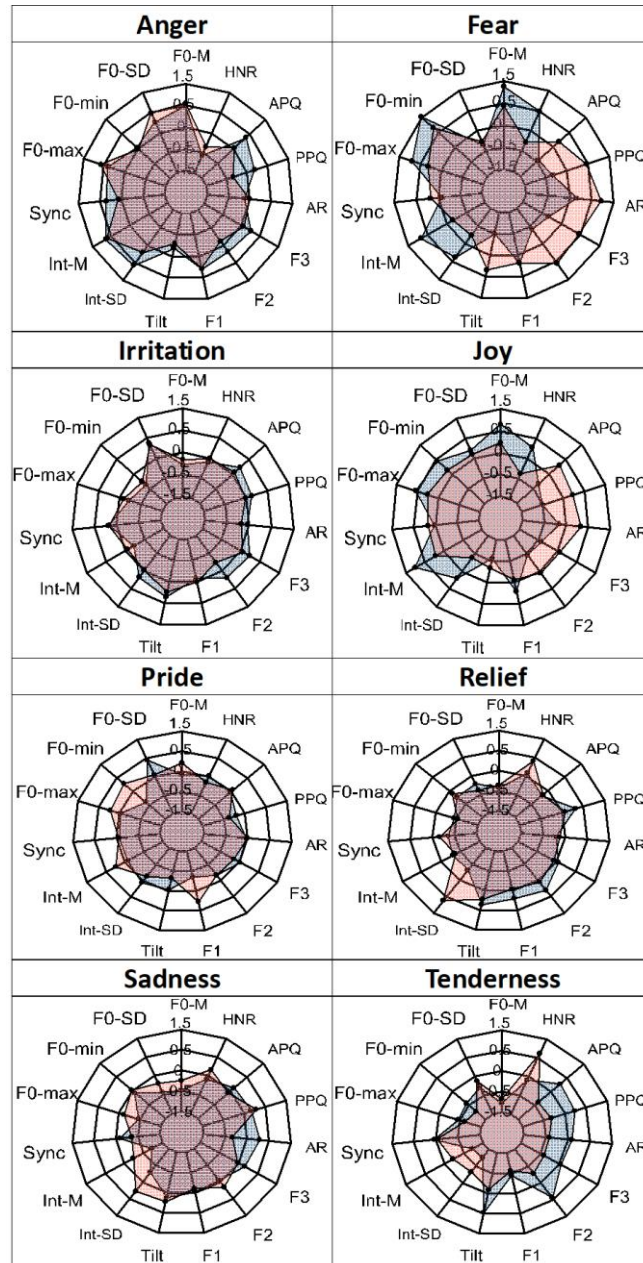


Figure 4.2. Radar plots for eight emotions portrayed by Dutch (blue-shaded polygons) and Korean (red-shaded) voice actors. The means of 15 acoustic parameters are plotted along radial axes after within-speaker z -transformation. For abbreviations of parameter names, see Table 4.2.

Duration parameters

- **AR:** There was a significant main effect of Emotion on AR in all data and in the subset with Korean recordings. Moreover, there was a significant main effect of Speaker Language on AR, which was much faster in Korean than in Dutch. Additionally, the effect of Gender on AR reached statistical significance only in all data, with males speaking faster than females. Also, the analyses yielded a significant two-way interaction between Emotion (Pride, Anger, Tenderness, Relief, Sadness, and Irritation) and Speaker Language in all data, and a significant two-way interaction between Emotion (Anger) and Gender in the subset with Dutch recordings.

Laryngeal parameters

- **PPQ:** There was a significant main effect of Emotion on PPQ across each dataset. Moreover, there was a significant main effect of Speaker Language on PPQ in all data, with more obvious fluctuations observed in Korean than in Dutch, showing vocal vibrations caused by linguistic differences. Additionally, the analyses revealed a significant main effect of Gender in all data and in the subset with Korean recordings, with more instabilities in males than in females. Also, significant two-way interactions were found between Speaker Language and Gender, between Emotion and Speaker Language in all data, and between Emotion and Gender in the subset with Dutch recordings. Particularly, a significant three-way interaction was observed between Emotion (Tenderness, Relief, and Sadness), Speaker Language, and Gender. The PPQ of Sadness, Tenderness, and Relief was higher in males than in males in both recordings, whereas it was remarkably higher in Fear in males than in females in Korean, with slightly higher PPQ in males than in females in Dutch.
- **APQ:** There was a significant main effect of Emotion on APQ in each dataset. Furthermore, there was a significant main effect of Speaker Language on APQ in all data, which varied more greatly in Korean than in Dutch. Also, the analyses showed a significant main effect of Gender on APQ in all data and in the subset with Korean recordings, with greater variations in males than in females. Additionally, there were significant two-way interactions between Speaker Language and Gender, between Emotion and Gender, between Emotion and Speaker Language in all data, and between Emotion and Gender in the subset with Dutch recordings. Importantly, the three-way interaction between Emotion, Speaker Language, and Gender reached statistical significance in all data. Specifically, APQ of Tenderness, Sadness, Relief, Irritation, and Pride was higher in males than in females in Dutch and Korean, whereas APQ of

Fear was higher in males than in females in Korean, with similar APQ in males and females in Dutch.

- **HNR:** We observed a significant main effect of Emotion on HNR in each dataset. Moreover, there was a significant main effect of Speaker Language on HNR in all data, indicating that the voice was clearer in Dutch than in Korean. Also, the analyses revealed a significant main effect of Gender on HNR in all data and in Korean recordings, with females displaying higher HNR than males, suggesting that the voice was clearer in females than in males. Furthermore, the analyses yielded significant two-way interactions between Speaker Language and Gender, and between Emotion and Speaker Language in all data, as well as between Emotion and Gender in Dutch recordings. Particularly, a significant three-way interaction was observed between Emotion, Speaker Language, and Gender in all data. Specifically, HNR of Pride, Anger, Tenderness, Relief, and Irritation was higher in females than in males in both Dutch and Korean, whereas HNR of Sadness was significantly higher in females than in males in Dutch. However, HNR of Fear was similar in females and males in Dutch, while it was much higher in females than in males in Korean. These findings underscore the intricate relationship between emotion, speaker language, and gender on HNR, highlighting the need to consider these factors in voice quality studies.

Collectively, these findings revealed that the 15 acoustic parameters (Slope and PR were not included in the set) varied across Emotion, Speaker Language, and Gender. Among these factors, Emotion has the strongest influence on acoustic parameters, especially on pitch, amplitude, and laryngeal acoustic parameters. Furthermore, acoustic parameters, particularly laryngeal parameters, were modulated by speaker language, with Dutch actors producing more stable and clearer voice patterns than Korean actors. Additionally, Gender mainly affected pitch and laryngeal parameters, with females exhibiting larger pitch variations but a more stable and clearer voice quality than males.

As shown in the radar plot (Figure 4.2), for most panels, the discrepancy between Dutch and Korean is relatively small (i.e., the polygons are almost the same for both languages). However, some emotions, Fear, Joy, and Tenderness, have clearly different polygons for Dutch and Korean.

- Fear: Korean actors emphasized Tilt, F2, F3, AR, PPQ, and APQ, whereas Dutch actors focused on F0-M, F0-min, F0-max, Int-M, Int-SD, and HNR.
- Joy: Differences for Joy between Dutch and Korean resemble those seen in Fear but concern fewer acoustic parameters. Specifically, Korean actors

no longer emphasized Tilt and F3, but still emphasized F2, AR, PPQ, and APQ. Dutch actors continue to emphasize F0-M, F0-min, F0-max, Int-M, Int-SD, and HNR.

- **Tenderness:** Discrepancies between Dutch and Korean seem a bit the opposite of what is seen in Fear and Joy. Specifically, Dutch actors paid attention to the spectral and laryngeal parameters, while Korean actors emphasized intensity (but not pitch) and HNR.

4.3.2 Vocal emotion recognition by Support Vector Machine (SVM) (Hypothesis 4)

The fourth research question asked whether vocal emotions can be predicted from acoustic parameters, and to compare the performance of machine classifiers for Dutch and Korean in both within-group and between-group conditions. We hypothesized that vocal emotions can be reliably predicted from acoustic parameters, and that recognition accuracy would be higher for both speaker languages in the within-group condition than in the between-group condition. We decided to adopt the SVM technology as the machine learning environment. The Support Vector Machine (SVM) is a machine learning algorithm that can be used for classification and regression analyses (Kecman, 2005). SVM can be applied to both linear and non-linear classification by using the kernel function (Noble, 2006). Four types of kernels are frequently used: linear, radial, polynomial, and sigmoid (Kavzoglu & Colkesen, 2009). In this study, we adopted the radial kernel, as it is suitable for both linearly and non-linearly distributed data.

The analyses of the linear mixed-effects models demonstrated that Slope and PR did not significantly vary across emotion, speaker language, or gender. Therefore, we excluded these two parameters, leaving 15 acoustic parameters. These acoustic parameters were *z*-normalized within speakers to eliminate individual differences in mean and standard deviation on the production of vocal emotions, ensuring that these irrelevant speaker differences did not affect the models' performance. Fifteen *z*-normalized acoustic parameters were then used as predictors in all Support Vector Machine (SVM) models to classify the tokens of eight emotions. The SVM models were built on two datasets: either the Dutch subset ($n = 128$) or the Korean subset ($n = 128$). These SVM models used leave-one-out cross-validation (LOOCV), which omits one data point each time for testing, and trains the recognizer on all other data points. All SVM models were created once with and once without Gender as an added predictor to evaluate the influence of gender on emotion classification across the three datasets. The output of the SVM models revealed that Gender did not play a role in the classification accuracy.

Therefore, in the following sections, we excluded Gender in the presentation of the SVM models.

Recognition accuracy was higher for the Dutch data (0.60, see Table 4.4) than for the Korean data (0.38). These results reveal that the optimal SVM model with 15 acoustic parameters performed best on the Dutch data, while the Korean model performed less well than the Dutch model. The differences in within-group recognition accuracy between these models demonstrate that while the performance of the model is affected by language-specific features, other factors, such as variability within and between individual voice actors—possibly caused by cultural differences and/or acting traditions, may play a role as well. These potential effects require further investigation.

Next, we applied the Korean model to the Dutch data, resulting in an out-group accuracy of 0.26, which was much lower than the in-group Dutch identification (0.60). Second, we applied the Dutch model to the Korean data, resulting in an out-group accuracy of 0.23, which was lower than the in-group Korean accuracy (0.38). Clearly, the results of both models were much lower when applied out-group than in-group, indicating that the separate models based on the Korean or Dutch recordings do not generalize well to the other language. This demonstrates that the acoustic parameters employed for emotion recognition differ significantly between Dutch and Korean, leading to lower recognition accuracy in out-group testing.

Third, we used Matthews Correlation Coefficient (MCC) to evaluate the performance of classification models based on the full confusion matrices (Chicco & Jurman, 2020; Păvăloi & Muscă, 2015; Silva et al., 2016) rather than on only the correct identification scores. The MCCs are ϕ association coefficients (between -1 and $+1$, computed on a 2×2 matrix defined by true vs false positives vs negatives, and tested for significance against a chi-square distribution with $df = 1$) computed on the observed number of correct responses and confusions found in the matrix and the theoretically perfect (error-free) scores (i.e., 100% along the main diagonal, and 0% in all off-diagonal cells). Again, we found that our SVM classifiers perform better when trained and tested on the same (within-group) data (Dutch: $\phi = 0.56$; Korean: $\phi = 0.38$) than across data (between-group) (Dutch model on Korean data: $\phi = 0.13$; Korean model on Dutch data: $\phi = 0.16$).

Together, the above results suggest that the acoustic features employed for emotion recognition differ between Dutch and Korean, and that the models do not perform well across language and culture. The large decrease in recognition accuracy in the between-group condition suggests that the

acoustic parameters are predominantly language-specific, displaying different patterns across languages that cannot be generalized well to cross-cultural/language vocal emotion recognition. However, there must still be a language/culture-universal component, as evidenced by the cross-language test scores (0.26 and 0.23 correct for Dutch and Korean, respectively) that remain approximately twice the chance level (= 0.125). This result was confirmed by a binomial test ($p < .001$, 95% CI [0.16, 0.32]).

Table 4.4. Emotion recognition accuracy (proportion correct) and Matthews Correlation Coefficient (MCC for correct identifications and confusions) obtained for Dutch and Korean emotion portrayals. Dutch and Korean SVM models (radial kernel function, 15 acoustic predictors) and human listeners were tested in-group (with Leave-One-Out cross-validation for machine recognition) and out-group.

| Listener/SVM model | Machine (MCC) | Humans (MCC) |
|---|----------------------|---------------------|
| Dutch model/listeners tested in-group | 0.60 (0.56) | 0.51 (0.44) |
| Dutch model/listener tested with Korean data | 0.23 (0.13) | 0.42 (0.34) |
| Korean model/listeners tested in-group | 0.38 (0.30) | 0.46 (0.39) |
| Korean model/listeners tested with Dutch data | 0.26 (0.16) | 0.39 (0.31) |

4.3.3 Comparison of vocal emotion recognition by machines and by human listeners (Hypothesis 5)

The fifth research question asked to what extent machine classifiers and human listeners differ in their use of acoustic parameters when identifying vocal emotions. We hypothesized that machine classifiers and human listeners would use acoustic parameters in a similar way when identifying vocal emotions (Hypothesis 5).

Machine classifiers, such as SVMs, may find accidental combinations of features that uniquely characterize an object but do not generalize to other tokens of the same object type. Therefore, cross-validation is needed to avoid spurious correct identification of objects. To reliably assess the performance of our SVM models, we tried different types of cross-validation. Ultimately, we chose Leave-one-out cross-validation (LOOCV), since it did not differ much from other forms of cross-validation and was easiest to implement. LOOCV is not needed for out-group SVM identification, as the training and test data do not overlap. The human identification data were collected in the experiment reported in Liang et al. (2025; Chapter 2). In that paper, only the accuracy scores of the identifications were presented; here, we analyze the

complete human confusion tables (which can be found in Table I1, Appendix I).

We observe that both listener groups recognized vocal emotions more accurately in their native language than in the unknown language, confirming the in-group advantage found in the previous study (Liang et al., 2025; Chapter 2). Both machines and human listeners recognized vocal emotions more accurately in the Dutch data than in the Korean data, although machines scored higher in the Dutch data than human listeners (Dutch model on Dutch recordings: 60% correct; Dutch listeners on Dutch recordings: 51%; Korean listeners on Dutch recordings: 39%). However, human listeners performed better in the Korean data than machines (Dutch listeners on Korean recordings: $\phi = 0.42$; Korean listeners on Korean recordings: $\phi = 0.46$; Korean model on Korean recordings: $\phi = 0.38$). The better performance of the Dutch model on Dutch data demonstrates that machine classifiers are optimized for the data they are trained on but perform less well cross-culturally. Compared to machines, human listeners maintain better performance across cultures, indicating that human listeners are better at generalizing their recognition ability across cultures. Therefore, it is necessary to improve machine classifiers' generalizability across cultures.

Additionally, the Matthews Correlation Coefficient was found to be higher for in-group than out-group listeners (Dutch listeners on Dutch recordings: $\phi = 0.44$; Korean listeners on Korean recordings: $\phi = 0.39$), while human listeners underperformed relative to the machine learning models.

Some emotions tend to be misclassified more often than others (see Table I1 in Appendix I). For instance, Anger and Irritation were symmetrically confused with each other by machine learning models and human listeners alike, revealing that these two categories shared similar vocal characteristics and are hard to differentiate—both by men and by machines. Additionally, Fear was frequently misinterpreted as Sadness, especially by human listeners.

To evaluate the difference in performance between machine classifiers and human listeners, we compared the recognition accuracy of each emotion (see Table 4.5).

Table 4.5. Recognition accuracy (% correct) of the intended emotion (in rows) as identified by SVM model (left part) and by human listeners (right part) for four combinations (in columns) of speaker language (test data) and listener language (training data).

| Trained on/ listeners | Identification by SVM | | | | Identification by human listeners | | | | Mean |
|--------------------------|-----------------------|--------|--------|--------|-----------------------------------|--------|--------|--------|------|
| | Dutch | Korean | Dutch | Korean | Dutch | Korean | Dutch | Korean | |
| Tested on/ speakers | Dutch | Korean | Korean | Dutch | Dutch | Korean | Korean | Dutch | |
| Anger | 100 | 94 | 69 | 81 | 63 | 35 | 39 | 67 | 69 |
| Fear | 100 | 69 | 31 | 25 | 41 | 52 | 56 | 26 | 50 |
| Irritation | 19 | 0 | 06 | 0 | 61 | 63 | 53 | 27 | 29 |
| Joy | 19 | 13 | 0 | 0 | 41 | 22 | 22 | 21 | 17 |
| Pride | 75 | 6 | 25 | 13 | 25 | 26 | 8 | 23 | 25 |
| Relief | 13 | 6 | 0 | 19 | 54 | 44 | 48 | 39 | 28 |
| Sadness | 88 | 44 | 38 | 19 | 85 | 83 | 76 | 86 | 65 |
| Tenderness | 69 | 69 | 19 | 5 | 34 | 40 | 31 | 16 | 41 |
| Mean | 60 | 38 | 24 | 26 | 51 | 46 | 42 | 38 | |

Machine classifiers outperformed human listeners on most vocal emotions, except Irritation and Sadness. Specifically, machine classifiers consistently achieved higher classification rates for Anger (94-100%) and Fear (69-100%), depending on the specific condition (see Table 4.5), than human listeners did (35-63% and 41-52%, respectively). In contrast to this, human listeners recognized particular emotions, Irritation (61-63%), Joy (22-41%), and Relief (44-54%) more accurately than machine classifiers. Furthermore, SVM models trained on Dutch data performed better in-group in Dutch than out-group in Korean, while models trained on Korean data performed better on Korean emotions than on Dutch emotions. This indicates that recognition accuracy decreased when testing cross-culturally (out-group) by machine classifiers. This mimics the results obtained earlier for human listeners.

To better present the similarities and differences of performance between machine classifiers and human listeners, we computed a hierarchical clustering dendrogram (Figure 4.3), which groups items according to their similarities or differences (distance) calculated by Ward's method (Contreras & Murtagh, 2015; Großwendt & Schmidt, 2019; Vichi et al., 2022). Ward's method minimizes the distance within each cluster. The distances were computed on the full confusion matrices of the eight emotions (64 cells per matrix), based on the frequencies rather than the proportions in Table 4.5. It should be pointed out that the y-axis plots the relative rather than the absolute distance between each model (see the distance matrix in Table I2, Appendix I). The chi-square distance matrix was computed by the *vegdist* function from

the *vegan* package in R (Oksanen et al., 2024). The Chi-square distance is suitable for frequency or proportion data (Mohanavalli & Jaisakthi, 2015). Each value shows the dissimilarity/distance between SVMs or human listeners, and the diagonal is 0 (a row compared to itself) (see Table I2, Appendix I). The relative distance ranges from 0 to 2.0. Zero indicates the smallest distance between the models with the greatest similarities, while 2.0 is the largest distance, i.e., between models with the greatest differences.

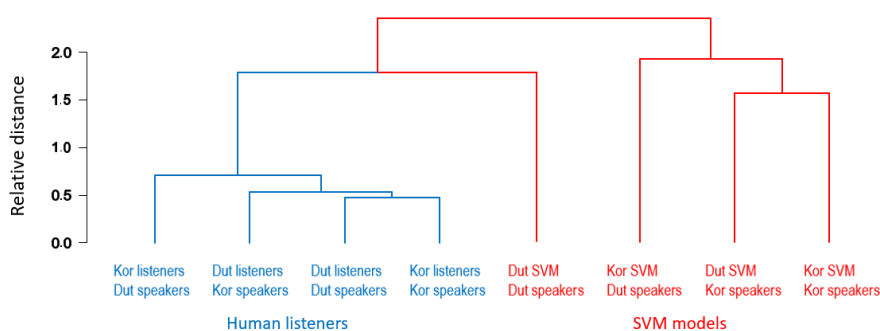


Figure 4.3. Hierarchical clustering dendrogram (Ward’s method) computed on all pairs of confusion matrices obtained by in-group and out-group testing of emotion recognition by SVM-simulated and human Dutch and Korean listeners.

The dendrogram reveals two main clusters. One is predominantly human, the other is exclusively SVM. Human listeners (Dutch and Korean listeners) are clustered together, separated from other SVM models, except SVM Dutch model on Dutch data. This demonstrates that Dutch and Korean listeners tended to use the same acoustic cues, resulting in a small distance. All human emotion identifications, whether in-group or out-group, resemble one another (relative distances are always below 1.0). Within the human cluster, the in-group conditions (Dutch listeners listening to Dutch recordings and Korean listeners listening to Korean recordings) are clustered at the lowest level of the hierarchy, indicating the greatest similarity. Then, the out-group conditions (Korean listeners listening to Dutch recordings and Dutch listeners listening to Korean recordings) are added to the human cluster.

Within the SVM cluster, one in-group condition (SVM model trained and tested on Korean), and two out-group conditions (SVM model trained either Dutch or Korean and tested on the other language) are grouped together, indicating that these three models probably rely on similar acoustic cues to identify vocal emotions, regardless of the language difference. However, the

in-group SVM model trained and tested on Dutch displays the least similarity within the SVM cluster and is positioned further from the other machine classifiers and human listeners, suggesting that it may use different acoustic cues in vocal emotion recognition. This difference can be attributed to the specific acoustic cues used by the Dutch model, highlighting language-specific influence on recognition performance.

4.4 Discussion

This study acoustically analyzed a total number of 256 emotional portrayals produced in two typologically different languages, i.e., Dutch and Korean, with 17 acoustic cues divided into five categories: pitch-, amplitude-, spectral-, duration-, and perturbation-related parameters. Our results corroborated previous findings and shed new light on acoustic patterns of vocal emotions, particularly in a less-studied language, i.e., Korean.

The first aim of this study was to examine the influence of emotion, speaker language, and gender on acoustic parameters. As predicted by Hypotheses 1-3, the selected acoustic parameters exhibit distinct patterns depending on emotion, speaker language, and gender.

Different vocal emotions exhibited distinct emotion-specific acoustic patterns. For instance, emotions like Fear, Anger, and Joy were produced with higher pitch parameters (F0-M, F0-min, and F0-max) and higher amplitude parameters (Int-M and Int-SD) than Sadness, Tenderness, Relief, Irritation, and Pride. This indicates a higher pitch level, wider pitch range, and larger intensity variations in Fear, Anger, and Joy than in the other emotions. Furthermore, Fear and Joy were spoken at a faster rate than the other six emotions. Additionally, Fear showed the lowest APQ and PPQ among the eight emotions, indicating that its voice quality was unstable and noisier than the others. Collectively, these findings align with Juslin and Laukka's (2003) meta-analysis, revealing that pitch, intensity, and speech rate are crucial acoustic parameters that successfully distinguish vocal emotions. However, in the present study, we found that laryngeal parameters provided additional distinctions between vocal emotions.

It is well-established that languages play a role in shaping vocal emotions (Scherer, 2001). In this study, we focused on two typologically different languages, where Dutch has quantity-sensitive word stress and a stress-timed rhythm, while Korean has phrasal stress and is syllable-timed (see § 1.3), leading to prosodic differences that can affect the production of vocal

emotions. Our data demonstrated that vocal emotions displayed language-dependent acoustic patterns. Notably, temporal discrepancies were observed between Dutch and Korean. Specifically, articulation rate (AR) was significantly higher in Korean than in Dutch, and there were no speech pauses in Korean, leading to a fast and more fluent speech flow in Korean. The Dutch speakers demonstrated a slower speech rate, occasionally accompanied by pauses. These differences highlight temporal distinctions in the portrayals by actors from the two languages when expressing vocal emotions. Moreover, Dutch speakers had higher F0-M, F0-min, F0-max, and Int-SD than Korean speakers. This indicates greater pitch and intensity variations with wider pitch range and greater loudness variation, which may be attributed to prosodic-linguistic differences between these two languages. Also, Dutch speakers exhibited higher HNR than Korean speakers, while Korean speakers demonstrated higher PPQ and APQ than Dutch speakers, suggesting a more stable and clearer voice quality in Dutch than in Korean. Together, these differences in acoustic parameters probably reflect language-dependent acoustic patterns in Dutch versus Korean speech in general (and by implication also of vocal emotions), which contribute to the in-group advantage, which is in line with previous findings (Banse & Scherer, 1996; Juslin & Laukka, 2003).

Specifically, pitch- and amplitude-related acoustic parameters are generally higher for female than male actors, although males displayed a more relaxed and sonorous voice than females, with greater and more rapid pitch change. Although Dutch and Korean actors exhibited parallel acoustic patterns in pitch-related parameters with subtle differences, Korean actors displayed higher mean intensity than their Dutch counterparts, particularly in male actors. These findings align with previous studies that vocal emotions are affected by physiological differences between females and males (Klatt & Klatt, 1990; Patel et al., 2011; Scherer, 1986).

Together, our observations are in line with previous findings that vocal emotions display universal and language-specific acoustic characteristics (Banse & Scherer, 1996; Scherer et al., 1991). The universal patterns of emotional prosody enable listeners to recognize vocal emotions above chance, even in unknown languages (Juslin & Laukka, 2003; Laukka et al., 2016; Paulmann & Uskul, 2014). Meanwhile, vocal emotions exhibit language-specific features affected by language and culture, supporting the language distance hypothesis (Scherer et al., 2001). This leads to the in-group advantage, such that individuals recognize emotions more accurately in their native language than in an unknown language (Laukka & Elenkin, 2021).

The second aim was to examine whether vocal emotions can be accurately classified by Support Vector Machine (SVM) models based on acoustic parameters, and to compare the performance of machine classifiers for Dutch and Korean in both in-group and out-group conditions (Hypothesis 4). Our findings demonstrated that classification rates can be reliably predicted from a constellation of acoustic parameters, which aligns with previous findings (e.g., Banse & Scherer, 1996). 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 level in both conditions. However, the SVM model performed better on the Dutch data than on the Korean data. The poor performance on the Korean dataset may be partly affected by language-specific factors, although other factors (e.g., data distribution) may have introduced noise, reducing the machine's ability to classify vocal emotions. Moreover, applying the Dutch SVM model to the Korean subset produced much lower accuracy than the results obtained when models were trained and tested within the same language. These findings further illustrate that SVM models can accurately classify vocal emotions from acoustic parameters within a given language, while their performance is hampered when applied across languages, highlighting the substantial influence of language-specific factors on vocal emotion recognition, supporting the dialect theory of emotion (Elfenbein et al., 2007; Elfenbein & Ambady, 2002b). Meanwhile, the cross-language recognition accuracy was twice as high as the chance level, indicating the presence of universal factors in cross-cultural emotion recognition. Together, these findings corroborate previous studies that cross-cultural emotion recognition is shaped by universal and language-specific factors (Elfenbein, 2013; Elfenbein & Ambady, 2002b; see also Keltner et al., 2019 for a review).

The third aim was to assess to what extent machine classifiers and human listeners use the same acoustic parameters to identify vocal emotions (Hypothesis 5). If they rely on the same measures to the same extent, then—apparently—the machine mimics human behavior, which would make the machine a cognitively realistic analog. In most cases, however, the machine performs better and uses different features than the human listener. This will happen especially when the train and test sets are small and contain the same elements. The machine will then set up one specific model for each item in the training set, which will find an exact match in the (identical) test set. This problem can be substantially reduced by applying cross-validation, i.e., by making sure that the test item is not contained in the training set. If, despite cross-validated modeling, the machine does (much) better than the human listener, it would mean that there are subtle systematic features in the acoustic signal that reliably signal a particular emotion, but the human listener fails to

pick these up (probably because they are too small to be audible, i.e., these features must be below threshold). Alternatively, the machine may perform more poorly than the human listener. In that case, the machine has not been given access to the full set of relevant acoustic properties that the human listener uses—the researcher should rethink the set of relevant features.

The results revealed that machine learning models outperformed human listeners in the in-group condition on the Dutch data. The better performance of machines may be attributed to their exclusive dependence on acoustic parameters, which are unaffected by cultural and linguistic factors. However, although human listeners performed less successfully than machine classifiers in the in-group condition, they obtained higher recognition accuracy than machine classifiers in out-group conditions. This suggests that machine classifiers may ignore some important acoustic parameters in identifying vocal emotions when trained and tested on different datasets. Moreover, we observed that the recognition accuracy of particular emotions, i.e., Joy, Sadness, and Irritation, was much lower than that of human listeners, indicating that machine classifiers may fail to capture key acoustic parameters of specific emotions or were not given as many acoustic parameters as possible.

4.5 Conclusion

Emotional prosody plays a pivotal role in social communication (Pell & Kotz, 2021). It is characterized by a number of acoustic parameters, such as fundamental frequency, intensity, formants, tempo, and vocal instability or perturbation (Banse & Scherer, 1996). Although there is ample literature on acoustic parameters of vocal emotions, few studies have focused on non-European languages such as Korean. Consequently, few studies have compared vocal emotions produced in languages with radically different typologies. To bridge this gap, we acoustically analyzed eight vocal emotions, including some of the lesser-studied ones (e.g., Pride and Tenderness). To the best of our knowledge, this is the first empirical study on vocal emotions produced in two typologically different cultures and languages with a large number of acoustic parameters, including pitch-, amplitude-, spectrum-, duration-, and perturbation-related cues. Findings from this study provide compelling evidence that vocal emotions are distinguished by different patterns of acoustic parameters, which are further modulated by the language of the speaker. Vocal emotions are influenced by linguistic factors, such as prosodic features shaped by different languages. Additionally, different

emotions exhibit distinctive patterns of vocal features (Banse & Scherer, 1996; Johnson et al., 1986; Juslin & Laukka, 2003).

The second aim was to examine whether vocal emotions can be accurately classified by Support Vector Machine (SVM) models. Our results revealed that SVM models accurately classify vocal emotions based on a configuration of acoustic parameters, which is consistent with previous findings (Fragopanagos & Taylor, 2005; Luengo et al., 2005). Particularly, the model on the Dutch data performed best (83% correct within-group identification), so that the conclusion follows that it captured most of the important parameters for classification.

The third aim was to examine to what extent machine classifiers mimic human listeners. The findings revealed that machine classifiers obtained significantly higher recognition accuracy than human listeners, especially higher for in-group than out-group listeners. This discrepancy may be attributed to the absence of cultural and linguistic bias in machine classifiers. Despite the overall superior performance of the machine classifiers, they performed less successfully than human listeners in between-group conditions, suggesting that they may adopt a different approach in identifying vocal emotions when trained and tested on linguistically different data. Furthermore, we observed that the classification rates of particular emotions were much lower than those of human listeners, indicating that machine classifiers may have overlooked some crucial acoustic parameters that are distinct for specific emotions, or—more likely—may not have been given access to all relevant acoustic parameters that the human listener employs.

Our findings provide supporting evidence for the universality and language-specific hypotheses in emotion recognition (Elfenbein, 2013; Elfenbein & Ambady, 2002b). However, it remains unknown to what extent each acoustic parameter contributes to emotion recognition. Further studies should examine the relative contributions of each acoustic parameter to recognition accuracy and might include, perhaps, supplementary acoustic parameters. Moreover, in this study, we examined acoustic parameters for each of the eight emotions separately. Follow-up studies might investigate the relationships between acoustic parameters and more general emotional dimensions, such as arousal, valence, basicness, intensity, and potency (Laukka et al., 2005). The ultimate aim is a comprehensive framework to better understand the contributions and interactions of multiple factors that affect the production and perception of vocal emotions.

Chapter Five

Recognizing vocal emotions in unfamiliar languages²³

Abstract

This study investigates cross-cultural and cross-linguistic vocal emotion recognition by testing American English and French listeners (Listener Language) listening to vocal emotions produced in Dutch and Korean (Speaker Language). Both listener groups recognized all emotions above chance, except for Pride in Korean by American English listeners (.08), thereby supporting the Universality hypothesis (Elfenbein, 2013). However, we did not find consistent patterns for Cultural Proximity (Elfenbein & Ambady, 2003), as both groups of listeners performed similarly on average in both recordings. Likewise, the Prosodic Proximity hypothesis was not supported by our data, since French listeners did not outperform American English listeners in recognizing emotions in Korean, although French and Korean obviously share basic prosodic structures. Instead, the effect of culture and language was emotion-specific, varying across Speaker Language and/or Listener Language. Additionally, despite the significant impact of the emotional dimensions of arousal, valence, and basicness on recognition accuracy, these dimensions did not consistently predict accuracy, since individual emotions deviated from the general patterns of accuracy due to emotion-specific features. Together, these findings provide a more nuanced perspective on the relative contributions of universality, culture, language, and emotional dimensions to cross-cultural emotion recognition, especially for listeners unfamiliar with the target language.

Keywords: prosodic similarities, vocal emotion recognition, culture, language

²³ This article is a substantially revised and expanded version of Liang et al. (2023). We used the same database, but we changed and refined the testing of the statistical models in a fundamental way, and we included emotion both as a fixed effect and as a random effect (the latter in testing emotional dimensions). By expanding the framework of analysis, we provide more and better evidence to test our hypotheses.

https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/253.pdf
The revised version has been submitted. Liang, Y., van Heuven, V., & van Hout, R. (submitted). Recognizing vocal emotions in unfamiliar languages: universal patterns versus cultural and prosodic proximity.

5.1 Introduction

Emotions are prevalent in daily life, being fundamental to human interaction (Jensen & Pedersen, 2016). Research on cross-cultural emotion production and perception dates back more than a century. As a pioneer in affective science, Charles Darwin proposed that the production and perception of emotions are biologically determined and universal, being inherited from human genes (1872, reprinted in 1998). Building upon Darwin's work, numerous studies have been carried out to resolve the controversial issue of whether cross-cultural emotion recognition is universal or shaped by culture- and/or language-specific factors (Ekman, 1972; Ekman et al., 1969; Elfenbein & Ambady, 2003b; Jiang et al., 2015; Paulmann & Uskul., 2014; Sauter et al., 2010; Scherer et al., 2001; see also Laukka & Elfenbein, 2021 for a review). To date, an increasing amount of evidence demonstrates that cross-cultural emotion recognition is an interplay between universal and culture- and/or language-specific factors (e.g., Ekman, 1972; Elfenbein & Ambady, 2002a). Many studies focus on facial emotion recognition. Our study is on vocal emotion recognition. No other sources of information are available to listeners but the voice. To investigate universality, we selected two groups of listeners to recognize vocal emotions in two languages unfamiliar to them. We selected two groups, American-English and French, listening to Dutch and Korean vocal emotions. Our basic hypotheses originate from three theoretical frameworks: the Universality, Cultural Proximity, and Language Distance hypotheses.

5.1.1 Universality hypothesis

The Universality hypothesis posits that emotional expressions are universally recognized by individuals across cultures and languages. Early studies demonstrated striking similarities in how individuals from unrelated and greatly different cultures expressed emotions with their facial expressions and recognized them in others (Ekman et al., 1969). Along this line, other studies on facial expressions have replicated this finding, providing supporting evidence that certain emotions can be accurately recognized across cultures (for a meta-analysis, see Elfenbein & Ambady, 2002b). Further, similar findings are obtained in the vocal domain. For instance, Van Bezooijen (1984) examined recognition of ten emotions produced in Dutch in Taiwanese Chinese and Japanese listeners, who did not have any knowledge of Dutch, compared to native Dutch speakers. The results showed that the recognition accuracy of both Japanese and Taiwanese Chinese listeners was above chance. More recently, Laukka and Elfenbein (2021) conducted a meta-analysis of 37 studies on cross-cultural vocal emotion recognition and found that both

linguistic (Juslin & Laukka, 2003; Laukka et al., 2016; Paulmann & Uskul, 2014; Pell, Monetta, et al., 2009) and non-linguistic (Cordaro et al., 2016; Laukka et al., 2013; Sauter et al., 2010a, 2015b; Sauter & Scott, 2007) vocal emotions can be identified above chance across cultures.

5.1.2 The Cultural Proximity hypothesis

The Cultural Proximity hypothesis proposed by Effenbein and Ambady (2003), highlights the pivotal role of culture in emotion recognition. Individuals from a similar cultural background can identify emotions more accurately than those from a totally different one, which is regarded as evidence for the existence of cultural “dialects”. Early empirical studies on cross-cultural emotion recognition, which started with facial emotion recognition, found that specific facial expressions (e.g., happiness, sadness, anger, fear, surprise, disgust, interest) are universally recognized, although the recognition accuracy varied in different cultures (Ekman & Friesen, 1971; Ekman, 1970, 1972; Ekman et al., 1969). For example, to examine the effect of cultural differences on emotion recognition, Laukka et al. (2016) tested emotion recognition in English among listeners from five different English-speaking countries (America, Australia, India, Kenya, and Singapore), including Western and non-Western countries. Cultural norms may affect the acoustic features with which emotions are produced and perceived, varying across countries. By including speakers and listeners from these different English-speaking countries, it was ensured that the stimuli reflected both linguistic and cultural diversity. The listeners were more accurate in recognizing emotions produced by speakers from their own culture than by those from a different one.

With respect to the Cultural proximity hypothesis, Korea stands out. Asian culture emphasizes collectivism and restraint in emotional expressions, while Western culture underscores individualism and overt expression of emotion (Keung, 2003). Hofstede (2001) distinguishes six cultural dimensions: power distance, individualism vs. collectivism, masculinity vs. femininity, uncertainty avoidance, long-term vs. short-term orientation, and indulgence vs. restraint. Based on the scores of these six dimensions of the Netherlands, South Korea, America, and France (see Table J1 in Appendix J), we calculated the overall cultural distance using the Euclidean distance formula (see Table J2 in Appendix J). America, France, and the Netherlands are closely related in terms of cultural distance, while South Korea is culturally more distant. However, France and South Korea are culturally closer than America and the Netherlands.

5.1.3 The Language Distance hypothesis

The Language Distance hypothesis introduced by Scherer et al. (2001) emphasizes that it is much easier for listeners to decode emotions produced in a language typologically similar to their native language than to a different one. In a pioneering study, Scherer et al. (2001) investigated the influence of language distance on emotion recognition. They presented 30 pseudo-utterances expressing five basic emotions (anger, fear, joy, sadness, and neutrality) produced in German to listeners from nine different countries (Germany, Switzerland, the United Kingdom, the Netherlands, America, Italy, France, Spain, and Indonesia). While recognition accuracy in each listener group was above chance, German listeners had the highest accuracy, followed by Dutch and English listeners, which is consistent with the Language Distance hypothesis. In contrast, Indonesian listeners, whose native language was the only non-Indo-European language, had the lowest accuracy. In this study, Scherer et al. (2001) regarded language similarity in terms of language family. However, language similarity should not be defined only by the traditional typological distinctions of (morpho)syntax, semantics, and lexicon. Instead, in the context of vocal emotions, it should be understood as well through shared or common prosodic features (e.g., rhythm, pitch range, pitch contour) that may have a direct impact on the production and perception of vocal emotions (Thompson & Balkwill, 2006). These prosodic features may facilitate or interfere with the perception of vocal emotions.

Along this line, Pell, Paulmann et al. (2009) tested 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 revealed that Argentine Spanish listeners had above-chance recognition accuracy in all languages, but with the highest scores in their native language. However, the differences in recognition accuracy may be attributed to the cultural effect, since participants were from different countries. Further, Pell, Monetta et al. (2009) examined correlations between acoustic patterns of vocal emotions and recognition accuracy. They found that acoustic parameters, especially mean fundamental frequency (f_0), f_0 range, and speech rate, accounted for 70-80% of the differences in recognition accuracy. These acoustic parameters provide listeners with essential cues to decode vocal emotions (Thompson & Balkwill, 2006; Wilson & Wharton, 2006), as vocal emotions are encoded via variations and interplays of acoustic parameters (e.g., f_0 , intensity, duration, rhythm, perturbation of voice quality) in speech (Banse & Scherer, 1996; Juslin & Laukka, 2003). For instance, sadness is related to low f_0 and slow speech rate, while anger and happiness exhibit higher mean f_0 with fast speaking rate

(Banse & Scherer, 1996; Juslin & Laukka, 2003). Therefore, when we explain recognition accuracy of vocal emotions with the Language Distance hypothesis, we in fact focus on prosodic cues rather than other linguistic parameters, although Scherer et al. (2001) did not specify what aspects of language distance may play a role. The close relationship between culture and language poses a challenge in splitting their specific effects on emotion recognition. For example, in the study conducted by Scherer et al. (2001), 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.

With respect to language typology, Dutch, English, and French belong to the Indo-European language family, with Dutch and English classified as Germanic languages, and French as a Romance language. Korean belongs to the much smaller Koreanic language family. So, language typologically, Korean stands out, whereas it stands out culturally as well.

5.1.4 The role of prosodic proximity in vocal emotion recognition

Rhythm, a basic feature of prosody, is a fundamental aspect of speech and can be defined as a sequence of auditorily distinct acoustic phenomena that repeats regularly over time (Clarke, 1999). The rhythm of speech is affected—among other things—by the duration of vowels and consonants, each arguably exerting a different influence on affective meaning. For instance, vowels are considered to carry more affective meaning than consonants (Majid, 2012). As a result, listeners are more affected emotionally by vowel duration than by consonant duration. In our stimuli, the pseudo-sentence /nuto hɔm sɛpikɑŋ/ was created with a roughly equal number of vowels (/u/, /o/, /ɔ/, /ɛ/, /i/, and /ɑ/) and single consonants (/n/, /t/, /h/, /m/, /s/, /p/, /k/, and /ŋ/). Given the use of a pseudo-sentence, the differences in prosodic structure seem more important than language typological differences.

Regarding their prosodic structure, the languages involved have to be categorized differently than in a language typological way (see Table 5.1). In Dutch and American English, the foot is the main grouping element in rhythm, and there is lexical stress (e.g., Bertan, 1999; Gussenhoven, 2005). In French (Jun & Fougeron, 2002) and Korean (Arvaniti, 2012), on the other hand, the phrase is the main prosodic grouping element, while these two languages have no lexical stress. There are Intonational Phrases (IP), composed of Accentual Phrases (AP) in French and Korean, and every phrase boundary is signaled by a final rising pitch movement and lengthening (Jun, 2006; Jun & Fougeron, 2002; Kim et al., 2008). In Korean, stress is not applied at the word level;

instead, it is realized at the phrase level (Lee, 1990). In Korean, there is one (and only one) stress in each AP, and its placement follows specific rules. Stress typically falls on either the first or the second syllable of the AP, depending on the number and weight of syllables, as well as on the location of the phrase within the sentence (Jun, 1995). Syllable weight is determined by the structure of the syllable, such that heavy syllables with coda consonants or long vowels attract stress.

In contrast, the stress in French is independent of syllable weight and always goes to the last syllable of the AP (a trailing schwa is not considered a syllable). The final stress rule underscores the fundamental difference between French and Korean, where the latter allows for greater variation in stress placement within the accentual phrase. In terms of prosodic systems, French and Korean are rather close: both have phrasal stress rather than word stress, and both languages are claimed to have a syllable-timed rhythm, in contrast to Dutch and English, which have word stress and stressed-timed rhythm. This selection of languages allows us to investigate whether prosodic proximity facilitates cross-cultural vocal emotion recognition. Table 5.1 summarizes the cultural and linguistic proximities.

Table 5.1. Cultural (bottom triangle) and linguistic (top triangle) proximities between the four groups of languages/cultures involved. Linguistic proximity distinguishes typological and prosodic proximity. “+”: close, “-”: distant.

| | American English | Dutch | | French | | Korean | |
|----------|------------------|----------|---------|----------|---------|----------|---------|
| | | Typology | Prosody | Typology | Prosody | Typology | Prosody |
| Am. Eng. | | + | + | + | - | - | - |
| Dutch | + | | | + | - | - | - |
| French | + | | + | | | - | + |
| Korean | - | | - | | - | | |

5.1.5 Predicting recognition accuracy based on emotional dimensions

Universal, cultural, and linguistic factors are commonly linked to discrete emotions. Vocal emotions are recognized above chance, also in other languages, but these accurate recognitions might be triggered by more general underlying factors or dimensions. Three such dimensions mentioned in relation to emotion are arousal, valence, and basicness.

Arousal describes perceivers' physiological reactions caused by emotions and can be categorized into low-arousal and high-arousal emotions (Russell & Barrett, 1999). An individual's level of arousal affects their ability to make decisions or judgments (Clark et al., 1984; Lane et al., 1999; Mourão-Miranda et al., 2003; Smith et al., 2011). Arousal modifies the vocal features of vocal emotions, such that high-arousal emotions are usually expressed with a higher pitch and longer duration than low-arousal emotions (Breitenstein et al., 2001). Low-arousal emotions are recognized more accurately than high-arousal emotions (Liang et al., 2025; Chapter 2).

Valence, on the other hand, is an individual's subjective experience influenced by emotions and can be divided into either positive or negative emotions. Both positive and negative emotions can be conveyed via a series of vocal signals (Cowen et al., 2019; Laukka & Elfenbein, 2021). Negative emotions are more accurately recognized than positive ones (Laukka et al., 2016; Laukka & Elfenbein, 2021; Sauter et al., 2010; Scherer et al., 2011), which is likely due to the fact that while positive emotions are tied to social relationships, negative emotions are directly linked to danger and survival.

Moreover, basicness is another pivotal characteristic of emotions. The emotion theory distinguishes six emotions (anger, disgust, fear, happiness, sadness, and surprise) that can be universally recognized (Ekman, 1992). Compared to non-basic emotions, basic emotions have well-recognized signals (e.g., physiology, antecedent events, etc.) in all cultures (Ekman, 1999). Therefore, basic emotions are recognized more accurately than non-basic emotions within and across cultures (Liang et al., 2025; Sauter et al., 2010).

5.1.6 Research questions and hypotheses

The overarching goal of this study is to explain vocal emotion recognition in two listener groups without prior knowledge of the languages they are listening to, Dutch and Korean. We want to test hypotheses based on Universality, Cultural Proximity, and Prosodic Proximity. In addition, we want to test the role of three more general emotional dimensions in vocal emotion recognition.

First, we examine the above-chance accuracy in unfamiliar languages in vocal emotion recognition by American English and French listeners. According to the Universality hypothesis, we hypothesize that both groups of listeners can identify emotions above chance, even in an unfamiliar language (Hypothesis 1). Second, we examine whether both groups of listeners will recognize emotions more accurately in Dutch than in Korean. Based on the cultural

proximity hypothesis, we hypothesize that both groups of listeners will recognize vocal emotions more accurately in Dutch than in Korean, since (American) English, French, and Dutch are from the same language family (Indo-European Germanic language) and are culturally distant from Korean (Hypothesis 2). Third, we examine whether French listeners will outperform American English listeners in Korean. Building upon the prosodic proximity hypothesis, we hypothesize that French listeners will identify Korean vocal emotions more accurately than American English listeners, as French is prosodically closer to Korean than to Dutch (Hypothesis 3).

Fourth, we examine whether recognition accuracy can be reliably predicted by emotional dimensions (arousal, valence, and basicness). First, we examine the influence of arousal on emotion recognition. Based on previous studies (Ekman, 1999; Laukka et al., 2016; Liang et al., 2025; Sauter et al., 2010), recognition accuracy will be higher for low-arousal than high-arousal emotions (Hypothesis 4). Second, we examine the influence of valence on emotion recognition. According to prior studies (Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2011), we assume that recognition accuracy will be higher for negative than positive emotions (Hypothesis 5). Third, we examine the influence of basicness on emotion recognition. Based on prior findings (Ekman, 1992, 1999; Elfenbein & Ambady, 2002), we predict that recognition accuracy will be higher for basic than non-basic emotions (Hypothesis 6).

5.2 Method

5.2.1 Materials

For the stimuli, we used all 256 vocal expressions from the Demo/Koremo (Dutch emotion/Korean emotion) corpus (Broersma et al., 2025).²⁴ To avoid any semantic cues to emotion recognition, the corpus uses a pseudo-sentence /nuto hɔm sepikɑŋ/ which is phonologically compatible in both Dutch and Korean.²⁵ The pseudo-sentence 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 (Ekman,

²⁴ The scenarios and corpus are publicly available via Radboud University at <https://doi.org/10.34973/5kg3-9852>

²⁵ The Korean speakers used the vowel sounds [a] and [o] that are slightly different from the vowel sounds [ɑ] and [ɔ] used by the Dutch speakers.

1992b). Moreover, the stimuli are balanced on two dimensions that are crucial in the understanding of emotions (Russell, 1980), namely arousal (high vs. low) and valence (positive vs. negative; Table 5.2). The stimuli consisted of 256 portrayals (8 emotions \times 2 tokens per emotion \times 8 speakers \times 2 languages) in total. For more information regarding the corpus and the recording procedure, refer to Liang et al. (2005, Chapter 2).

5.2.2 Participants

Two groups of participants: 25 native American English listeners (19 females, 6 males, age: $M = 20.6$, $SD = 1.94$) who were students at Northwestern University, Chicago; and 30 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. Participants were given either course credits or a small payment as a reward for their participation.

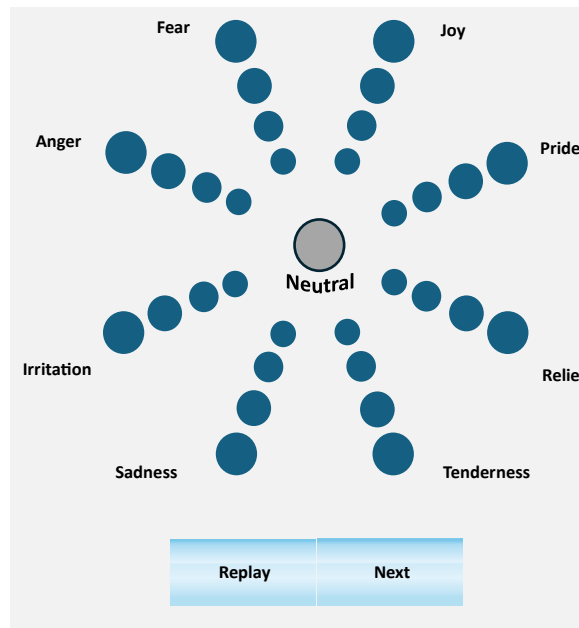


Figure 5.1. The emotion wheel in English. Translation in Dutch and Korean, Joy: “Blijdschap”, “행복”; Pride: “Trots”, “자랑스러움”; Relief: “Opluchting”, “안도감”; Tenderness: “Vertederling”, “애정”; Sadness: “Verdriet”, “슬픔”; Irritation: “Irritatie”, “짜증”; Anger: “Woede”, “분노”; Fear: “Angst”, “공포”; Neutral: “Neutral”, “중립”.

Table 5.2. The eight emotions included in this study in a valence-by-arousal grid (from Goudbeek & Broersma, 2010b: 2212), with basic emotions denoted with “*”.

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

5.2.3 Procedure

Each participant was tested individually in a sound-attenuated room at their university. On each trial, an emotion wheel (Figure 5.1) was displayed on the computer screen in front of the participant, exhibiting eight emotions written in the participants’ native language (English or French), each with four circles in different sizes representing different intensities. Participants could choose Neutral if they believed the recording expressed no emotions. They responded by clicking on one of the circles on the emotion wheel. Participants listened to the recordings via high-quality headphones. There was no time constraint for their responses, as they could listen to each recording as many times as they preferred.

The experiment began with a block of 128 Korean stimuli, followed by a block of 128 Dutch stimuli. Before each block, participants were informed about its language, and they completed eight practice trials. The experiment was conducted using JATOS software (Lange et al., 2015) on a standard laboratory computer and took around 35-45 minutes. In this study, we only analyzed categorical responses (i.e., recognition accuracy, but not intensity ratings).

5.3 Results

We performed the data analyses in R (R Core Team, 2022) with a series of logistic mixed-effects models run with the *lme4* package (Bates et al., 2015). The outcome variable for each model was Accuracy. For hypotheses 1 to 3, we ran a model in the entire dataset, including three fixed predictors (Speaker Language, Listener Language, and Emotion) and two random intercepts (Speaker and Listener). Also, we added the relevant random slopes based on

the model fit obtained from ANOVA comparisons (for more details, see subsections 5.3.1 to 5.3.3).

For hypotheses 4 to 6, we ran three separate models to assess the effects of Arousal, Valence, and Basicness on recognition accuracy. Each model included two fixed predictors: Speaker Language and Listener Language. Each of these models included one additional fixed predictor representing one of the three emotional dimensions: Arousal (high vs. low), Valence (negative vs. positive emotions), and Basicness (basic vs. non-basic emotions). Additionally, we included three random intercepts: Listener, Speaker, and Emotion, accounting for variation within these three factors. Next, we incorporated the random slopes according to ANOVA comparisons of model fit and selected the model with the lowest AIC. In particular, we examined how accuracy varied across emotions by focusing on the random intercept for Emotion. In each model, the mean of the emotions is defined as zero within each category, in which each emotion's effect reflects its deviation from the mean. If the emotion perfectly plays the role as a random effect, its confidence interval would include zero. However, emotions with confidence intervals excluding zero may reveal a bias in the model involved. For all models, we adopted contrast coding.

The recognition accuracy varied across the eight emotions, as is clear from Figure 5.2 (see also Appendix K). Sadness consistently had the highest accuracy across American English and French listeners in both language recordings, ranging from 0.68 to 0.76. Pride had the lowest accuracy by all listeners across Dutch and Korean recordings, especially in Korean recordings by American English listeners (0.08). Overall, emotions like Sadness and Anger were more recognizable than other emotions like Tenderness and Pride. In many cases, there is an overlap in the recognition accuracies of the two listener groups.

5.3.1 Above-chance recognition accuracy by both groups of listeners (Hypothesis 1)

The first research question examined the above-chance accuracy in Dutch and Korean by both groups of listeners. We hypothesized that both groups of listeners would recognize vocal emotions above chance in both Dutch and Korean (Hypothesis 1). We observed that the recognition accuracy of each emotion was above chance of .11 (based on a 9-alternative forced-choice task) in Dutch recordings by both groups of listeners (see Figure 5.2; the exact figures can be found in Model 1, Table 5.3; the confusion matrices for across languages and listener groups can be found in Table L1 and L2 in Appendix L). Only the accuracy of Pride was not significantly above chance level (.11) in Korean recordings, whereas other emotions varied in recognition accuracy, with Sadness being high and Pride low (see Figure 5.2). Specifically, the Pride confidence interval for American English listeners did not cross the chance level, whereas the confidence interval for French listeners slightly crossed the chance level. Such a difference illustrates that there are emotion-dependent distinctions between the two listener groups. A clear example is Anger, where American English listeners performed worse than French listeners in both languages. Sometimes, there is an interaction. The French performed worse for Relief, with a clear distinction from the American-English in recognizing this emotion in Korean. These variations in speaker language, listener language, and emotions do not contradict hypothesis 1 of an overall positive effect of overall above-chance recognition. Aligning with the Universality hypothesis, both listener groups were able to identify vocal emotions above chance, even in an unfamiliar language.

Table 5.3. Summary of results of the logistic mixed-effects model analyses for Hypotheses 1-3.

| Model 1: Analysis with Accuracy by both listener groups (Hypotheses 1-2) | | | | | |
|---|-----------------|--------------|------|--------|--------|
| Random effects | <i>Variance</i> | | | | |
| Listener (Intercept) | 0.25 | | | | |
| Speaker (Intercept) | 0.30 | | | | |
| Fixed effects | β | $Exp(\beta)$ | SE | z | p |
| (Intercept) | 1.07 | 2.92 | 0.10 | 10.71 | < .001 |
| Speaker Language (SL) | -0.28 | 0.76 | 0.19 | -1.47 | .141 |
| Listener Language (LL) | -0.17 | 0.84 | 0.13 | -1.33 | .183 |
| Emotion (Anger) | -1.04 | 0.35 | 0.08 | -13.73 | < .001 |
| Emotion (Fear) | -1.05 | 0.35 | 0.07 | -14.21 | < .001 |
| Emotion (Irritation) | -1.34 | 0.26 | 0.07 | -18.09 | < .001 |
| Emotion (Joy) | -1.75 | 0.17 | 0.08 | -22.35 | < .001 |
| Emotion (Pride) | -2.75 | 0.06 | 0.09 | -30.58 | < .001 |
| Emotion (Relief) | -1.48 | 0.23 | 0.07 | -19.69 | < .001 |
| Emotion (Tenderness) | -2.06 | 0.13 | 0.08 | -26.33 | < .001 |
| SL \times LL | -0.22 | 0.80 | 0.22 | -1.00 | .318 |
| SL \times E (Anger) | -0.83 | 0.44 | 0.15 | -5.50 | < .001 |
| SL \times E (Fear) | 0.57 | 1.77 | 0.15 | 3.87 | < .001 |
| SL \times E (Irritation) | 0.31 | 1.36 | 0.15 | 2.10 | < .05 |
| SL \times E (Joy) | -0.96 | 0.38 | 0.16 | -6.14 | < .001 |
| SL \times E (Pride) | -0.86 | 0.42 | 0.18 | -4.81 | < .001 |
| SL \times E (Relief) | 0.04 | 1.04 | 0.15 | 0.23 | .815 |
| SL \times E (Tenderness) | 0.56 | 1.75 | 0.16 | 3.61 | < .001 |
| LL \times E (Anger) | 0.97 | 2.64 | 0.15 | 6.42 | < .001 |
| LL \times E (Fear) | 0.52 | 1.68 | 0.15 | 3.50 | < .001 |
| LL \times E (Irritation) | -0.16 | 0.85 | 0.15 | -1.11 | .269 |
| LL \times E (Joy) | 0.67 | 1.95 | 0.16 | 4.30 | < .001 |
| LL \times E (Pride) | 0.61 | 1.84 | 0.18 | 3.38 | < .001 |
| LL \times E (Relief) | -0.40 | 0.67 | 0.15 | -2.65 | < .01 |
| LL \times E (Tenderness) | 0.47 | 1.60 | 0.16 | 2.98 | < .01 |
| SL \times LL \times E (Anger) | 0.08 | 1.08 | 0.30 | 0.27 | .788 |
| SL \times LL \times E (Fear) | 0.05 | 1.05 | 0.30 | 0.18 | .856 |
| SL \times LL \times E (Irritation) | 0.10 | 1.11 | 0.30 | 0.34 | .737 |
| SL \times LL \times E (Joy) | 0.58 | 1.79 | 0.31 | 1.86 | .062 |
| SL \times LL \times E (Pride) | 0.14 | 1.15 | 0.36 | 0.38 | .705 |
| SL \times LL \times E (Relief) | -0.23 | 0.79 | 0.30 | -0.76 | .447 |
| SL \times LL \times E (Tenderness) | 0.88 | 2.42 | 0.31 | 2.83 | < .01 |

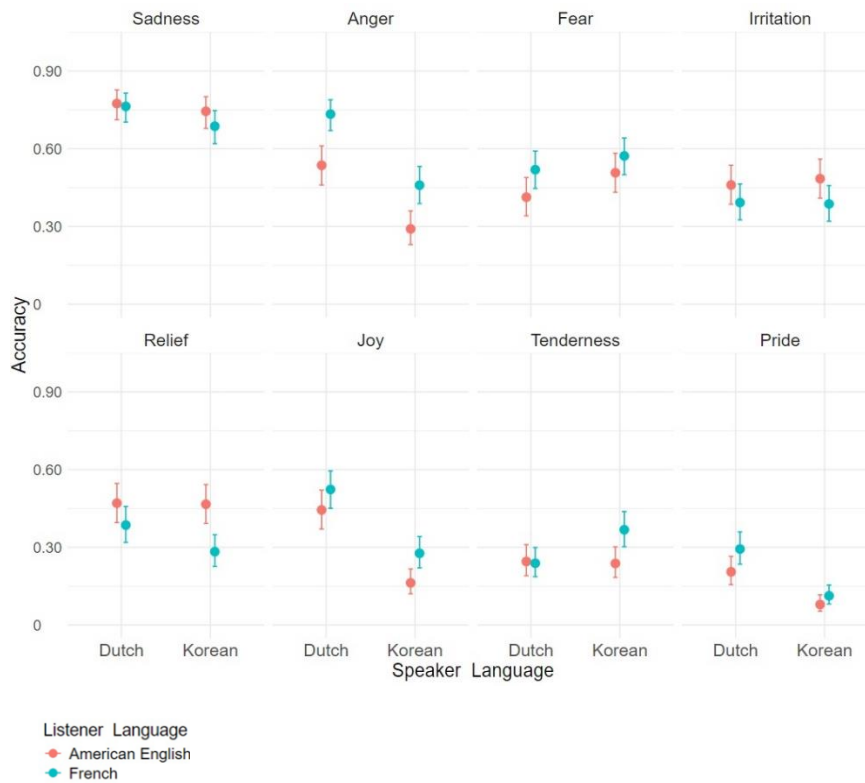


Figure 5.2. Recognition accuracy in Dutch and Korean by both groups of listeners

5.3.2 Recognition accuracy in Dutch recordings by both groups of listeners (Hypothesis 2)

The second hypothesis is about cultural proximity. We hypothesized that both groups of listeners would recognize vocal emotions more accurately in Dutch than in Korean. This hypothesis was tested in Model 1 (Table 5.3). The best-fitting model included three fixed predictors (Speaker Language, Listener Language, and Emotion) and two random intercepts (Speaker and Listener), leading to convergence.

The results revealed a significant main effect of Emotion, such that all emotions had lower accuracy scores than Sadness. This signals substantial differences between emotions. However, there was no significant main effect of Speaker Language or Listener Language on recognition accuracy. Figure 5.2 may seem to show that vocal emotions are slightly better recognized in

Dutch than in Korean, but individual emotions are too different, in particular in their interaction with Listener Language (see Model 1, Table 5.3). That means there are differences related to the two groups of listeners, but they are bound to particular emotions. The strongest interaction is found for Anger, as observed earlier on the basis of Figure 5.2.

There was even a significant three-way interaction between Speaker Language, Listener Language, and Emotion (Tenderness). Specifically, French listeners recognized Tenderness more accurately in Korean than in Dutch, whereas American English listeners recognized Tenderness similarly in both Dutch and Korean.

The lack of a significant main effect of Speaker Language on accuracy indicates that both groups of listeners identified vocal emotions in both Dutch and Korean on an equal footing. The cultural proximity effect depends on the specific emotions involved. There is no general cultural proximity effect over all emotions. There is no overall evidence supporting hypothesis 2.

5.3.3 Recognition accuracy in Korean recordings by both groups of listeners (Hypothesis 3)

The third research hypothesis predicts that French listeners would outperform American English listeners in recognizing emotions in Korean recordings. However, there was no significant main effect of Listener Language on accuracy (see Model 1, Table 5.3), indicating that both groups of listeners recognized vocal emotions with the same accuracy on average. On the other hand, there were many significant interaction effects between Listener Language and Emotion (see Model 1, Table 5.3). One strong interaction is between Relief and Listener Language, as French listeners perform worse for Korean (see also Figure 5.2). Listener groups vary in their performance, depending on the emotion involved, but there is no overall support for Hypothesis 3.

5.3.4 The effect of Arousal on accuracy (Hypothesis 4)

The fourth hypothesis addresses the effect of Arousal on accuracy. We hypothesized that listeners would recognize low-arousal emotions more accurately than high-arousal emotions, with variations across emotions. The best-fitting model (Model 2 in Table 5.4) consisted of three fixed predictors: Speaker Language, Listener Language, and Arousal. It included three random intercepts (Listener, Speaker, and Emotion), as well as random by-listener slopes for Speaker Language, random by-speaker slopes for Listener

Language, and random by-listener/speaker slopes for Arousal. The results demonstrated a significant main effect of Speaker Language on accuracy, as accuracy was higher in Dutch than in Korean ($\Delta = .39$), providing evidence supporting the cultural proximity hypothesis for arousal. However, there was no significant main effect of Arousal on accuracy. This finding contradicts earlier research that low-arousal emotions were recognized more accurately than high-arousal emotions (Liang et al., 2025; Chapter 2). However, there was a significant two-way interaction between Speaker Language and Arousal, indicating that listeners recognized low-arousal emotions more accurately than high-arousal emotions in Dutch, whereas they recognized high-arousal and low-arousal emotions similarly in Korean. This interaction is visualized in Figure 5.3. Moreover, there was a significant two-way interaction between Listener Language and Arousal, suggesting that American English listeners recognized low-arousal emotions more accurately than high-arousal emotions, as can be seen in Figure 5.3. In contrast, French listeners recognized low-arousal and high-arousal emotions equally well. Together, these findings indicate that the influence of Arousal on accuracy differs depending on the language of speakers and listeners. There was no three-way interaction.

Table 5.4. Summary of results of the logistic mixed-effects model analyses for Hypothesis 4.

| Model 2: Analysis with Arousal (Hypothesis 4) | | | | | |
|--|-----------------|--------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | | |
| Listener (Intercept) | 0.06 | | | | |
| By-Arousal | 0.24 | | | | |
| Listener (Intercept) | 0.01 | | | | |
| Speaker Language | 0.002 | | | | |
| Speaker (Intercept) | 0.08 | | | | |
| By-Arousal | 0.37 | | | | |
| Speaker (Intercept) | 0.02 | | | | |
| By-Listener Language | 0.02 | | | | |
| Emotion (Intercept) | 0.55 | | | | |
| Fixed effects | β | $exp(\beta)$ | <i>SE</i> | <i>z</i> | <i>p</i> |
| (Intercept) | -0.36 | 0.70 | 0.28 | -1.31 | .190 |
| Speaker Language (SL) | -0.39 | 0.68 | 0.16 | -2.43 | < .05 |
| Listener Language (LL) | 0.14 | 1.15 | 0.09 | 1.54 | .123 |
| Arousal (A) | 0.40 | 1.49 | 0.55 | 0.72 | .471 |
| SL \times LL | -0.04 | 0.96 | 0.11 | -0.40 | .692 |
| SL \times A | 0.67 | 1.95 | 0.31 | 2.15 | < .05 |
| LL \times A | -0.74 | 0.48 | 0.15 | -4.87 | < .001 |
| SL \times LL \times A | -0.02 | 0.98 | 0.15 | -0.12 | .904 |

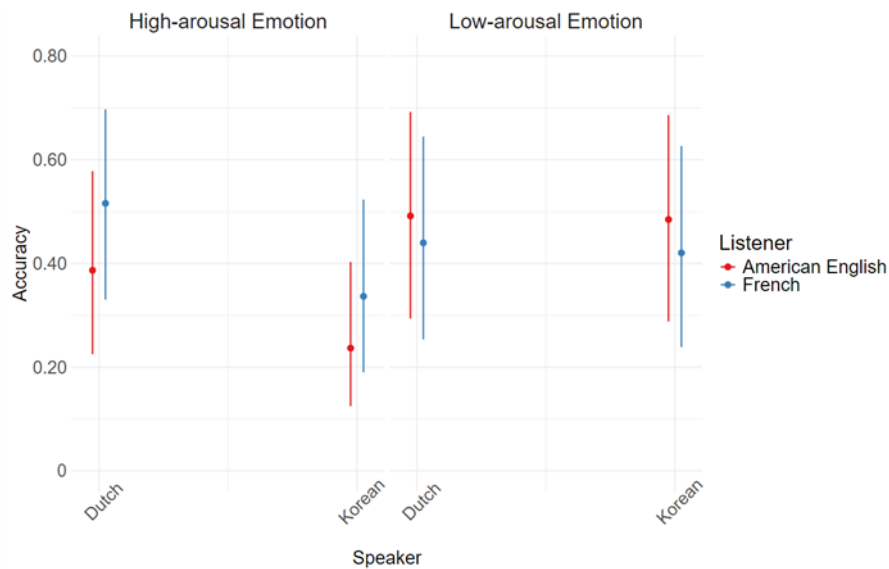


Figure 5.3. Recognition accuracy for high-arousal and low-arousal emotions in Dutch and Korean recordings by American English and French listeners (marginal R^2 : 0.032, conditional R^2 : 0.230), and their confidence intervals (2SE).

To examine the success of including arousal as a fixed variable, we visualized the random intercepts across the eight emotions. Figure 5.4 shows the values of the intercepts for the low-arousal versus the high-arousal emotions. The discrepancy between high-arousal and low-arousal emotions was indicated by the main effect of arousal ($\Delta = .40$). If emotions are a random effect within the arousal category, their confidence intervals would include the zero value. However, Sadness, Tenderness, Anger, Fear, and Pride violate the pattern predicted by arousal. Specifically, Sadness is extremely high, whereas Pride is extremely low. Their deviations are far above the .40 effect of Arousal. These findings demonstrate that the binary classification in high-arousal versus low-arousal does not account for a reliable dichotomy of the recognition accuracy across emotions. Hypothesis 4 has to be rejected.

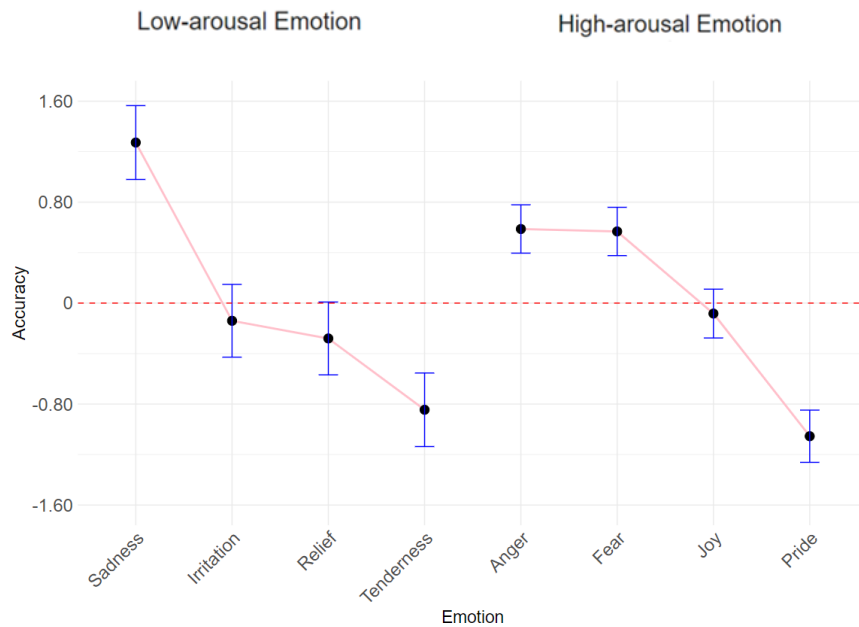


Figure 5.4. Intercepts of recognition accuracy of low-arousal (Sadness, Irritation, Relief, Tenderness) and high-arousal emotions (Anger, Fear, Joy, Pride) with their confidence interval (2SE). The order of emotions from left to right is listed as follows: (1) the four low-arousal emotions are presented left, and the four high-arousal emotions are presented right; (2) within each group (low-arousal or high-arousal), emotions are listed according to the size of their intercepts in the accuracy (from high to low).

5.3.5 The effect of Valence on accuracy (Hypothesis 5)

The fifth hypothesis concerned the effect of Valence on emotion recognition. We hypothesized that listeners would recognize negative emotions more accurately than positive emotions. The best-fitting model (Model 3 in Table 5.5) included three fixed predictors: Speaker Language, Listener Language, and Valence. Additionally, we added three random intercepts (Listener, Speaker, and Emotion), as well as random by-speaker slopes for Listener Language, and random by-listener/speaker slopes for Valence. The output revealed a significant main effect of Speaker Language on accuracy, displaying that recognition accuracy was higher in Dutch than in Korean ($\Delta = .36$), supporting the cultural proximity hypothesis for Valence. Further, there was a significant main effect of Valence on accuracy, exhibiting higher

accuracy in negative than in positive emotions ($\Delta = 1.15$). The distinction can be seen in Figure 5.5. There was also a main effect of Speaker Language, indicating that the French listeners scored higher than their American English counterparts. This can be seen in Figure 5.5 as well. In all four conditions, the French scored higher than the American English. There were no two-way and three-way interactions.

Table 5.5. Summary of results of the logistic mixed-effects model analyses for Hypothesis 4.

| Model 3: Analysis with Valence (Hypothesis 5) | | | | | |
|--|-----------------|--------------|-----------|----------|----------------|
| Random effects | <i>Variance</i> | | | | |
| Listener (Intercept) | 0.0002 | | | | |
| By-Valence | 0.13 | | | | |
| Listener (Intercept) | 0.06 | | | | |
| Speaker (Intercept) | 0.08 | | | | |
| By-Valence | 0.36 | | | | |
| Speaker (Intercept) | 0.04 | | | | |
| Listener Language | 0.02 | | | | |
| Emotion (Intercept) | 0.24 | | | | |
| Fixed effects | β | $exp(\beta)$ | <i>SE</i> | <i>z</i> | <i>p</i> |
| (Intercept) | -0.37 | 0.69 | 0.20 | -1.88 | .060 |
| Speaker Language (SL) | -0.36 | 0.70 | 0.17 | -2.03 | <.05 |
| Listener Language (LL) | 0.14 | 1.15 | 0.09 | 1.62 | .105 |
| Valence (V) | -1.15 | 0.32 | 0.38 | -3.00 | <.01 |
| SL \times LL | -0.06 | 0.94 | 0.11 | -0.56 | .573 |
| SL \times V | -0.22 | 0.80 | 0.31 | -0.72 | .469 |
| LL \times V | -0.04 | 0.96 | 0.12 | -0.37 | .715 |
| SL \times LL \times V | 0.19 | 1.21 | 0.15 | 1.28 | .200 |

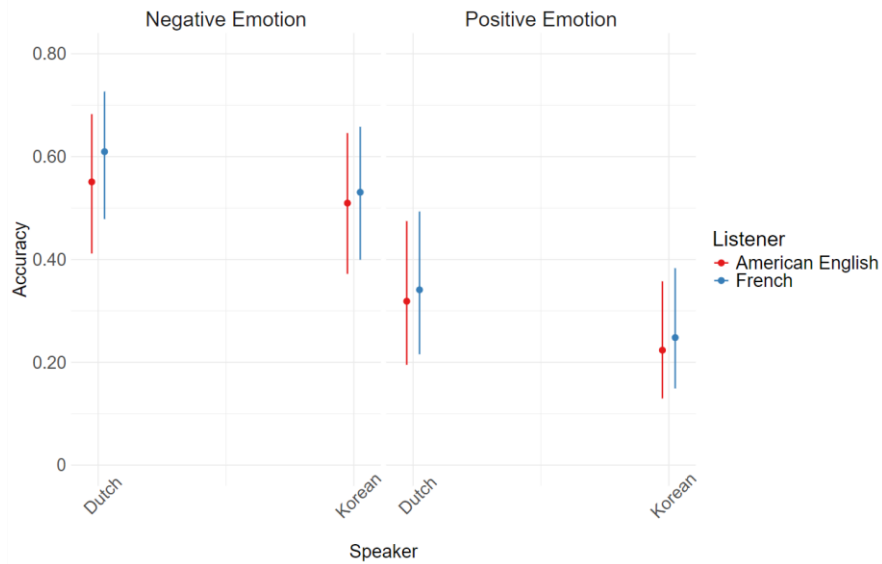


Figure 5.5. Recognition accuracy for negative and positive emotions in Dutch and Korean recordings by American English and French listeners (marginal R^2 : 0.090, conditional R^2 : 0.198), and their confidence intervals (2SE).

The success of the model is again tested by visualizing the random intercepts of the emotions. These intercepts and their confidence intervals can be found in Figure 5.6. Several emotions (Sadness, Irritation, Relief, and Pride) deviate substantially from the effects predicted by the fixed effects. These results demonstrate that, although accuracy is affected by valence, confirming hypothesis 5, the binary split in negative versus positive does not lead to a clear dichotomy of the eight emotions in accuracy, in fact contradicting hypothesis 5.

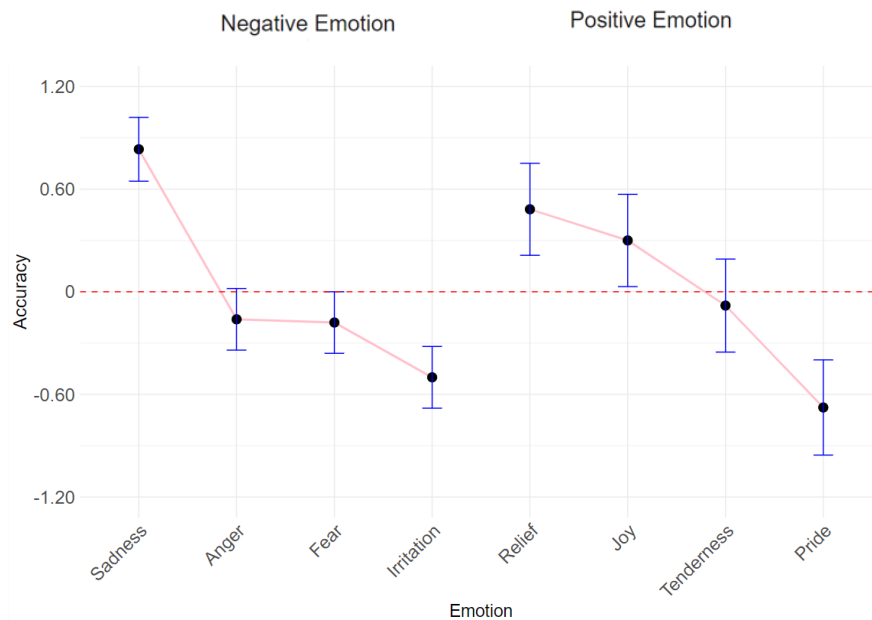


Figure 5.6. Intercepts of recognition accuracy of negative (Sadness, Anger, Fear, Irritation) and positive emotions (Relief, Joy, Tenderness, Pride), with their confidence interval (2SE). The order of emotions from left to right is listed as follows: (1) the four negative emotions are presented left, and the four positive emotions are presented right; (2) with each group (negative or positive), emotions are listed according to the size of their intercepts in the accuracy (from high to low).

5.3.6 The effect of Basicness on accuracy (Hypothesis 6)

The sixth hypothesis focuses on the influence of Basicness on emotion recognition. We hypothesized that listeners would recognize basic emotions more accurately than non-basic emotions. The best-fitting model (Model 4 in Table 5.6) included three fixed predictors: Speaker Language, Listener Language, and Basicness. We included three random intercepts (Listener, Speaker, and Emotion), as well as random by-listener slopes for Speaker Language, random by-speaker slopes for Listener Language, and random by-listener/speaker slopes for Basicness. The results demonstrated a significant main effect of Speaker Language on accuracy, such that both groups of listeners recognized emotions more accurately in Dutch than in Korean ($\Delta = .37$), as shown in Figure 5.7, providing evidence for the cultural proximity

hypothesis for Basicness. Also, there was a significant main effect of Basicness on accuracy, indicating that listeners recognized basic emotions more accurately than non-basic emotions ($\Delta = .97$), as shown again in Figure 5.7. Moreover, there was a significant two-way interaction between Listener Language and Basicness: while both listener groups recognized basic emotions more accurately than non-basic emotions, the difference was more pronounced for French listeners than for American English listeners. There was no three-way interaction between Speaker Language, Listener Language, and Basicness, indicating that the two-way interaction between Listener Language and Basicness was not modulated by Speaker Language.

Table 5.6. Summary of results of the logistic mixed-effects model analyses for Hypothesis 4.

| Model 4: Analysis with Basicness (Hypothesis 6) | | | | | |
|--|-----------------|--------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | | |
| Listener (Intercept) | 0.01 | | | | |
| By-Basicness | 0.13 | | | | |
| Listener (Intercept) | 0.06 | | | | |
| Speaker Language | 0.001 | | | | |
| Speaker (Intercept) | 0.06 | | | | |
| By-Basicness | 0.27 | | | | |
| Speaker (Intercept) | 0.04 | | | | |
| Listener Language | 0.02 | | | | |
| Emotion (Intercept) | 0.32 | | | | |
| Fixed effects | β | $exp(\beta)$ | <i>SE</i> | <i>z</i> | <i>p</i> |
| (Intercept) | -0.36 | 0.70 | 0.22 | -1.64 | .101 |
| Speaker Language (SL) | -0.37 | 0.69 | 0.16 | -2.28 | < .05 |
| Listener Language (LL) | 0.13 | 1.14 | 0.09 | 1.46 | .145 |
| Basicness (B) | -0.97 | 0.38 | 0.43 | -2.26 | < .05 |
| SL \times LL | -0.08 | 0.92 | 0.11 | -0.72 | .471 |
| SL \times B | 0.37 | 1.45 | 0.27 | 1.36 | .173 |
| LL \times B | -0.48 | 0.62 | 0.12 | -3.84 | < .001 |
| SL \times LL \times B | 0.03 | 1.03 | 0.15 | 0.21 | .837 |

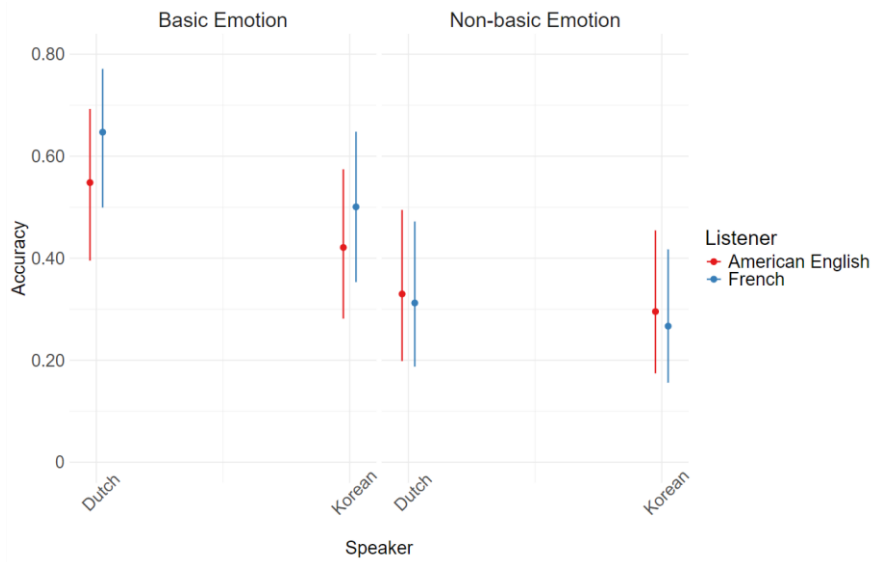


Figure 5.7. Recognition accuracy for basic and non-basic emotions in Dutch and Korean recordings by American English and French listeners (marginal R^2 : 0.075, conditional R^2 : 0.197), and their confidence intervals (2SE).

To evaluate the differences in accuracy across emotions, we visualized their random intercepts in Figure 5.8. The difference between basic and non-basic emotions was captured by the main effect of basicness ($\Delta = .97$). However, some emotions (Sadness, Joy, Irritation, Relief, and Pride) clearly violate the consistent pattern of recognition of basic/non-basic emotions. Specifically, Sadness is notably high, while Pride is particularly low. Undoubtedly, it is a violation of the overall effect of basicness. Therefore, the binary split between basic versus non-basic does not result in a transparent dichotomy of the eight emotions. Hypothesis 6 has to be rejected as the effect of basicness is not general enough.

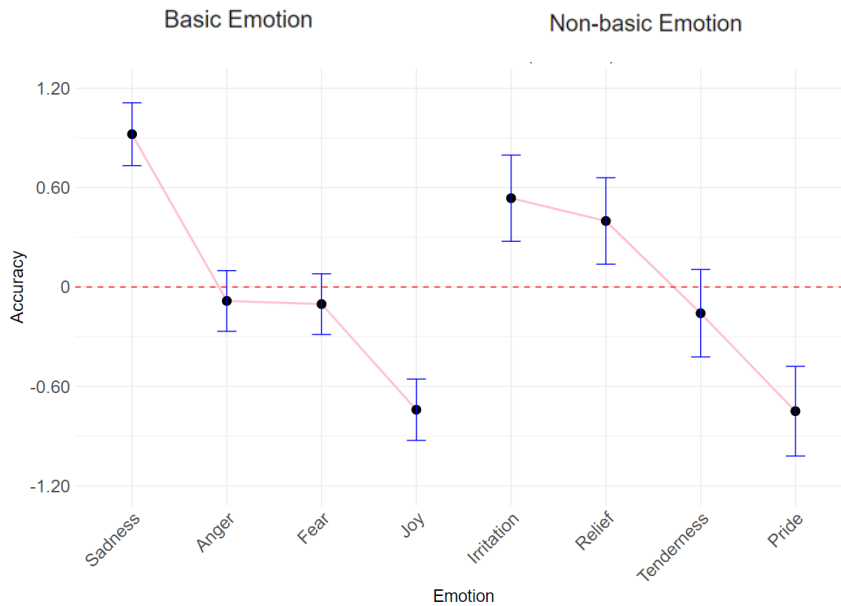


Figure 5.8. Intercepts of recognition accuracy of basic (Sadness, Anger, Fear, Joy) and non-basic emotions (Irritation, Relief, Tenderness, Pride), with their confidence interval (2SE). The order of emotions from left to right is listed as follows: (1) the four basic emotions are presented left, and the four non-basic emotions are presented right; (2) with each group (basic or non-basic), emotions are listed according to the size of their intercepts in the accuracy (from high to low).

5.3.7 Comparing the dimensional and discrete emotional effects

Taken together, our dimensional results replicate effects found in previous studies (Laukka & Elenfeldt, 2021; Liang et al., 2025; Sauter et al., 2015): higher accuracy for (i) negative and (ii) basic emotions than for (i) positive and (ii) non-basic emotions. Contrary to an earlier finding by Liang et al. (2025; Chapter 2), there was no significant main effect of arousal on accuracy. However, the dimensional effects were marked by exceptions and do not seem to apply as an overall effect, transcending individual emotions.

We summarized more details of the statistical results of the three binary emotional dimensions in Table 5.7. There is a discrepancy between the marginal and conditional R^2 , whereas they are both not high, indicating that the three dimensions were not successful as general predictors of accuracy. Interestingly, they produce significant effects involving Speaker Language

and Listener Language. These effects are spurious as more general effects and seem to be related to individual emotions.

Table 5.7. Summary of the analysis of the dimensions of Arousal, Valence, and Basicness. Violations are intercepts of emotions with a confidence interval not including zero value in the wrong direction. D = Dimension. “+” indicates a significant effect, “-” indicates no significant effect.

| Dimension | R ² | | Vio. | D | SL | LL | SL×LL | SL×D | LL×D |
|-----------|----------------|-------------|------|---|----|----|-------|------|------|
| | Marginal | Conditional | | | | | | | |
| Arousal | 0.032 | 0.230 | 4 | - | + | - | - | + | + |
| Valence | 0.090 | 0.198 | 5 | + | + | - | - | - | - |
| Basicness | 0.075 | 0.197 | 4 | + | + | - | - | - | + |

We decided to perform an analysis per emotion to uncover in detail the differences in the effects of Speaker Language and Listener Language. The outcomes are summarized in Table 5.8.

Table 5.8. Significant effects of Speaker Language (SL), Listener Language (LL), and their interaction (SL × LL) on recognition accuracy of emotions (“+” indicates a significant effect, “-” indicates no significant effect).

| Emotion | SL | LL | SL × LL |
|------------|----|----|---------|
| Sadness | - | - | - |
| Fear | - | - | - |
| Joy | + | - | - |
| Irritation | - | + | - |
| Anger | + | + | - |
| Pride | + | + | - |
| Relief | - | + | + |
| Tenderness | - | - | + |

Table 5.8 shows that the significant effects of Speaker Language, Listener Language, and their interaction on recognition accuracy vary across the eight emotions. The recognition of Fear and Sadness was not affected by Speaker Language or Listener Language at all. Joy, Irritation, Anger, and Pride showed independent effects of Speaker Language and/or Listener Language. A Speaker Language effect points to a cultural proximity effect. The score for

Korean is lower than for Dutch. A Listener Language effect may point to a prosodic proximity effect, but the American English score higher for Irritation and Relief, contradicting the effect of prosodic proximity. Relief and Tenderness show a difference for the two language groups for Korean, pointing to a potential language effect.

5.4 Discussion

The primary goal of this study was to examine the contributions of universal, cultural, and linguistic factors in vocal emotion recognition in an unknown language. Their contributions are put forward by three different theoretical perspectives on vocal emotion recognition based on universality, cultural proximity, and prosodic proximity. The combination of two evidently different languages, Dutch and Korean, and two listener groups with a different language background, American English and French, seems to offer a promising testing ground to unravel the relative contribution of various factors.

The first aim was to examine above-chance recognition accuracy in both groups of listeners. Consistent with the Universality hypothesis (Elfenbein, 2013; Elfenbein & Ambady, 2002a), American English and French listeners recognized all emotions above chance in both Dutch and Korean recordings, except for the Pride by American English listeners, who identified Pride below chance level (.08). This is a strong outcome, as American English and French listeners who had no prior knowledge of Dutch or Korean were able to recognize vocal emotions above chance. The Universality hypothesis, however, does not explain the rather substantial variation in accuracy between the emotions. The one exception of Pride, moreover, shows that violations may occur of the universality claim.

The second aim was to investigate the role of Cultural Proximity in emotion recognition. The results revealed that there is no overall main effect for the two speaker languages involved. The expectation was that Korean was harder for both groups across all emotions. This outcome contradicts earlier findings that cultural similarity leads to higher recognition accuracy in a cross-cultural setting (Elfenbein & Ambady, 2002, 2003; Scherer et al., 2001). The strong interactions between speaker language and emotion reveal that Cultural Proximity may play a role in individual emotions, meaning that we need to look more locally at differences for individual emotions to investigate cultural proximity.

The third aim was to explore the role of prosodic proximity in vocal emotion recognition. Building upon the Language Distance hypothesis (Scherer et al., 2001). We argued that French listeners would outperform American English listeners in Korean recordings, since French and Korean share basic prosodic features (Jun & Fougeron, 2002). Notably, Scherer et al. (2001) defined language similarity in terms of language typology, without specifying which exact aspect(s) play(s) a role. However, in the vocal domain, language distance should be considered in terms of common prosodic characteristics, like rhythm, pitch range, or pitch contour, which may directly affect the interpretation and expression of vocal emotions. The simple nonsense sentence spoken by the Dutch and Korean speakers can only give prosodic information, as there are no morphosyntactic or lexical-semantic cues. The hypothesis of a positive effect of prosodic similarity was not supported by our data, as French listeners did not recognize vocal emotions overall better than American English listeners. Again, there were differences between the two listener groups, but these were dependent on the vocal emotion involved.

The fourth aim was to examine whether vocal language recognition can be predicted by general dimensions underlying the similarity structure of emotions: Arousal, Valence, and Basicness. These dimensions may be more successful in tracing the effects found in the vocal emotions investigated with respect to speaker language and listener language. First, however, we did not find a significant main effect of arousal on accuracy, which contradicts previous findings (Liang et al., 2025; Chapter 2). The pattern for arousal was mixed, with significant interactions between arousal and speaker language/listener language. Second, we found that negative emotions were recognized more accurately than positive emotions, which aligns with prior findings (Laukka & Elfенbein, 2021; Liang et al., 2025). Third, basic emotions were recognized more accurately than non-basic emotions, in line with earlier findings that basic emotions are more easily decoded than non-basic emotions across cultures (Liang et al., 2025; Sauter et al., 2010). Further, we examined the role of arousal (high vs. low), valence (negative vs. positive), and basicness (basic vs. non-basic) as a binary dichotomy in accuracy. However, our data did not reveal consistent patterns of recognition accuracy for these dimensions due to violations of particular emotions. Notably, across these three dimensions, Sadness scores extremely high, while Pride scores exceptionally low, revealing that recognition accuracy cannot be reliably predicted based merely on these three dimensions. Therefore, the interpretation of vocal emotions should take individual emotions and emotion-specific characteristics into consideration.

Individual emotions exhibit emotion-specific features, as the accuracy of emotions varied across Speaker Language, Listener Language, and their interaction. For instance, the perception of Fear and Sadness was unaffected by either Speaker Language or Listener Language, supporting the Universality hypothesis (Elfenbein, 2013) and basic emotion theory (Ekman, 1992b). However, recognition of Joy, Irritation, Anger, and Pride was affected by either Speaker Language or Listener Language, suggesting that these emotions were influenced by either cultural or linguistic factors.

5.5 Conclusion

This study provides additional evidence to the literature on cross-cultural and cross-language emotion recognition by exploring the relative contributions of the Universality hypothesis, Cultural Proximity, Prosodic Proximity, and emotional dimensions (Arousal, Valence, and Basicness) to vocal emotion recognition. Specifically, we tested how well American English and French listeners, without prior exposure to Dutch or Korean, recognized vocal emotions produced in these two unfamiliar languages.

In line with the Universality hypothesis, both groups of listeners identified most emotions above chance in unfamiliar languages, except for Pride in Korean by American English listeners. Our data support the Universality hypothesis, although there are emotion-dependent variations in recognition. However, we did not find a general effect of Cultural Proximity on emotion recognition, as recognition accuracy varied across emotions, which was further modulated by Speaker Language or Listener Language. Notably, we discovered no evidence supporting the effect of Prosodic Proximity on cross-cultural/language vocal emotion recognition, as French listeners did not outperform American English listeners in Korean recordings. The results revealed that prosodic similarity alone did not consistently affect emotion recognition. Similar to Cultural Proximity, its influence was emotion-specific, which was further influenced by Speaker Language or Listener Language. Therefore, it remains unknown how prosodic structures affect vocal emotion recognition.

Further, our results demonstrated that while Arousal did not have a significant impact on recognition accuracy, Valence and Basicness did display significant influence. However, none of these dimensions consistently predicted recognition accuracy across cultures, since some particular emotions like Sadness and Pride violated the general patterns. These findings indicate that 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 recognition accuracy.

In conclusion, our results manifest that while universal principles exist in vocal emotion recognition, they are modulated by culture, language, emotional dimension, and emotion-specific factors. To better predict recognition accuracy, follow-up studies should develop a comprehensive framework incorporating variations caused by universality, culture, language, emotional dimensions, and emotions to understand their roles in shaping the production and perception of vocal emotions in cross-cultural settings.

Chapter Six

Conclusion and discussion

6.1 Introduction

This dissertation investigated cross-language (Dutch and Korean) vocal emotion recognition from multiple perspectives, 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) patterns of acoustic parameters across emotion, speaker language, and gender; 4) the role of cultural and prosodic similarity in vocal emotion recognition by American English and French listeners. Each perspective was addressed in a separate chapter, with four main research questions:

Chapter 2: Do Dutch and Korean listeners recognize vocal emotions above chance in Dutch and Korean, and is there an in-group advantage in vocal emotion recognition?

Chapter 3: Is there an in-group bias in intensity ratings of Dutch and Korean vocal emotions by Dutch and Korean listeners?

Chapter 4: How do acoustic parameters of vocal emotions vary across emotions, speaker language, and gender in Dutch and Korean?

Chapter 5: Is cross-cultural/language vocal emotion recognition in unfamiliar languages affected by Universality, Cultural Proximity, Prosodic Proximity, and emotional dimensions?

In addressing these research questions, I used affectively balanced corpora—Demo (Dutch emotion) and Koremo (Korean emotion). Notably, since previous studies on cross-cultural emotion recognition have either used “one-to-many” (one listener group) or “many-to-one” (one language) designs, this project adopted a “four-by-two” design, with four listener groups from typologically different cultures and languages and two typologically different languages. Moreover, I took not only a categorical approach to emotions (for

a similar approach, see Laukka, 2003), but also a dimensional approach (for a similar approach, see Laukka et al., 2005) into consideration.

In this final chapter, I will first summarize the research sub-questions, chapter by chapter, and discuss the results of the various experiments and analyses to see if and how these sub-questions and the main research questions were answered (§ 6.2). The next section (§ 6.3) discusses the contribution that the present series of studies makes to the literature on the signaling and perception of vocal emotion, and identifies the gaps in our knowledge that can now be filled in. In § 6.4, I formulate the general conclusions that can be drawn from the dissertation. The chapter ends by discussing the limitations inherent to the experimental choices that were made in the project (§ 6.5), followed by suggestions for future research (§ 6.6).

6.2 Main findings

6.2.1 Investigating cross-cultural vocal emotion recognition with an affectively balanced design

The first study, described in Chapter 2, aimed to examine the recognition accuracy of vocal emotions in a cross-cultural setting by two groups of listeners whose culture and language are typologically different. Dutch listeners with no knowledge of Korean and Korean listeners with no knowledge of Dutch participated in the first study. They were asked to identify the vocal emotions they heard in the stimuli produced in a pseudo-sentence /nuto hɔm sɛpikaŋ/. They were given a choice from eight different emotions, i.e., the basic emotions anger, fear, joy, sadness, and the secondary (non-basic) emotions irritation, pride, relief, and tenderness. Four emotions are characterized by high arousal (anger, joy, fear, pride), the other four by low arousal (irritation, relief, sadness, tenderness). The eight emotions can also be split by valence, yielding a subset of four positive (joy, pride, relief, tenderness) and four negative (anger, fear, irritation, sadness) emotions. Listeners who identify emotional portrayals produced by speakers of their own language, i.e., Dutch listeners identifying emotions produced by Dutch speakers and Korean listeners identifying emotions by Korean speakers, respond in a so-called in-group mode. Listeners responding to emotions produced by speakers of the language they are not familiar with respond in a so-called out-group mode. This study examined four sub-questions:

- 1) Is there an in-group advantage in cross-cultural emotion recognition?

- 2) Is there a difference in recognition accuracy between high-arousal and low-arousal emotions?
- 3) Is there a difference in recognition accuracy between positive and negative emotions?
- 4) Is there a difference in recognition accuracy between basic and non-basic emotions?

The results revealed that both listener groups, Dutch and Korean, recognized the eight emotions significantly above chance (chance level = 11%), not only in their own language (in-group mode) but also in the unknown language (out-group mode). The recognition accuracy for Dutch listeners in Dutch and Korean recordings was 47% and 38%, respectively, and that for Korean listeners in Dutch and Korean recordings was 36% and 43%, respectively.²⁶ Both listener groups displayed the predicted in-group advantage, such that listeners identified vocal emotions more accurately in their native language than in the unknown language, supporting the idea that cross-cultural emotion recognition relies on both universal and culture-/language-specific factors (Elfenbein, 2013; Elfenbein & Ambady, 2002b), as well as the dialect theory (Elfenbein & Ambady, 2002b; Elfenbein et al., 2007). In terms of the three binary splits of the eight emotions, recognition accuracy was higher for the subsets containing low-arousal, negative, and basic emotions than for the high-arousal, positive, and non-basic counterpart quadruplets. These regularities were found both within (in-group mode) and across (out-group mode) cultures/languages.

6.2.2 Interpreting the intensity of vocal emotions across cultures

The second study, reported in Chapter 3, investigated the intensity ratings of vocal emotions by Dutch and Korean listeners, which were obtained but not analyzed in the first study. In the first study, listeners not only indicated the emotion they heard, but also rated the intensity of the target emotion. In Chapter 3, we examined the rating of emotional intensity in listeners' native language (in-group mode) and in the unknown language (out-group mode), targeting three sub-questions:

- 1) Do accurate trials receive higher intensity ratings than inaccurate ones?
- 2) Is there an in-group bias for intensity ratings cross-culturally, specifically, are in-group intensity ratings higher than out-group ratings?

²⁶ In Chapter 2, we used 11% as the chance level for the recognition accuracy, since we included Neutrality as a ninth response category. In Chapter 4, however, we treated "Neutral" responses as missing data to afford better comparison of the classification accuracy between machines and humans; consequently, we used 12.5% as the chance level in the latter chapter.

- 3) Is there an association between intensity ratings and the categories defined by the three binary splits of arousal, valence, and basicness?

The results demonstrated that intensity ratings were higher for accurate than for inaccurate trials, for Dutch as well as for Korean respondents. Anger received the highest intensity ratings by both listener groups across accurate and inaccurate trials. However, we did not find the predicted in-group bias for intensity ratings in the data. With respect to the three binary splits, although they differed across individual emotions, the results revealed a strong association between intensity ratings and arousal, valence, as well as basicness, such that intensity ratings were higher for high-arousal than for low-arousal emotions (with Anger being extremely high and Pride being extremely low), higher for negative than for positive emotions (with Anger and Joy being the highest and Irritation being the lowest), and higher for basic than for non-basic emotions, with Anger being the highest—both within and across the Dutch-Korean language barrier.

6.2.3 Classifying emotions from acoustic parameters

The third study (Chapter 4) acoustically measured each stimulus according to a total number of 17 acoustic cues, which were divided into five categories: 1) pitch-related, 2) amplitude-related, 3) spectrum-related, 4) duration-related, and 5) perturbation-related. This study aimed to address the following three sub-questions:

- 1) How can the acoustic patterns of each of the eight vocal emotions be characterized across speaker language and gender?
- 2) How accurately can the eight emotions be classified based on acoustic cues, and to what extent does the classification improve when the languages (Dutch, Korean) are analyzed separately?
- 3) To what extent does the machine learning classifier adopted for the purpose (Support Vector Machine, SVM) mimic the (in-group and out-group) performance of human listeners in emotion classification?

The results showed that 1) each of the eight emotions is characterized by a different acoustic profile and that the parameters constituting the profiles differ across speaker language and gender, 2) vocal emotions can be reliably classified by SVM via an optimized configuration of acoustic parameters, 3) the classification rates improved when the data was separated by speaker language, 4) the machine learning classifiers outperformed human listeners, but 5) the overall order of difficulty of the identification tasks was the same for human listeners and machine classifiers, such that Dutch emotions were

better identified than Korean emotions while in-group identification was consistently better than out-group identification.

6.2.4 Universal patterns, Cultural Proximity, Linguistic Proximity, and emotional dimensions in cross-cultural vocal emotion recognition

The fourth study (Chapter 5) investigated vocal emotion recognition by American-English and French listeners. This study aimed to examine to what extent Universality, Cultural Proximity, Linguistic Proximity, and emotional dimensions affect vocal emotion recognition across cultures. This study asked six sub-questions:

- 1) Is the recognition accuracy above chance by both groups of listeners?
- 2) Do American English and French listeners recognize vocal emotions more accurately in Dutch than in Korean?
- 3) Do French listeners perform better than American English listeners on the Korean recordings?
- 4) Is the recognition accuracy higher in low-arousal than high-arousal emotions?
- 5) Is the recognition accuracy higher in negative than positive emotions?
- 6) Is the recognition accuracy higher in basic than non-basic emotions?

The results revealed that both new listener groups reached above-chance recognition accuracy for all emotions (chance level = 11%) in both types of recordings, which is consistent with the Universality hypothesis. However, we did not find the significant main effect of Speaker Language, although both new listener groups achieved slightly higher recognition accuracy with Dutch than with Korean recordings, which contradicts the Cultural Proximity hypothesis (Elfenbein & Ambady, 2003a). However, the strong correlations between Speaker Language and emotion indicate that culture may affect the perception of individual emotions. Furthermore, the results did not support Linguistic Proximity, since French listeners did not outperform American English listeners in Korean recordings, although French and Korean share similar prosodic features (Jun & Fougeron, 2002). Finally, the recognition accuracy was higher for negative than for positive, and higher for basic than non-basic emotions, which is consistent with earlier findings (Laukka & Elfenbein, 2021; Liang et al., 2025; Sauter et al., 2015). However, counter to our prediction, there was no significant main effect of arousal, as both listener groups recognized high-arousal and low-arousal emotions similarly. Notably, we found that the relative influence of emotional dimensions (arousal, valence, and basicness) varied across the eight emotions, with Sadness being extremely

high, and Pride being particularly low, unaffected by either Speaker Language or Listener Language.

6.3 Discussion

6.3.1 Adapting the dimensional approach

Classic emotion theory recognizes six basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992b). In this set, there are more negative than positive emotions, with happiness being positive, while surprise can be either a positive or a negative emotion. In the present study, there are four basic (anger, fear, joy/happiness, sadness) and four non-basic emotions (irritation, pride, relief, tenderness). The eight emotions involved in this study are equally divided between positive/negative (valence), high/low (arousal), and basicness (basic, compound) (see Table 1.1), but in a non-orthogonal fashion, which has made it impossible to come up with a straightforward factorial analysis of the effects of this three-way binary split.

It would seem doable, however, to come up with a set of eight emotions such that the arousal, valence, and basicness factors can be combined orthogonally in a factorial design by replacing some of the categories as indicated in Table 6.1.

Table 6.1. An alternative set of eight emotions in an orthogonal design defined by Arousal (high, low), Valence (positive, negative), and Basicness (basic, complex). Basic emotions are additionally specified by an asterisk.

| Valence | | Positive | | Negative | |
|-----------|------|-----------|---------|----------|------------|
| Basicness | | Basic | Complex | Basic | Complex |
| Arousal | High | Joy* | Pride | Anger* | Contempt |
| | Low | Surprise* | Relief | Sadness* | Irritation |

In Table 6.1, six of the original eight emotion types have been maintained. To obtain a completely balanced (orthogonal) design, however, basic Fear was replaced by non-basic (complex) Contempt, while non-basic Tenderness was replaced by basic Surprise. These basic-complex exchanges were made within the original Arousal \times Valence subsets.

The set of emotions in Table 6.1 does not mean that individual emotions have their own status and that their recognition is not an addition of underlying dimensions. It is necessary to sort out how emotions are embedded in a more general pattern of connectivity.

6.3.2 Cluster analysis: The cross-cultural perspective: separating and confusing emotions

Cluster analysis is a statistical method used to group items together that share similarities, where closer/clustered emotions are more similar than those farther apart (Albornoz et al., 2011; Kurematsu et al., 2010). In Chapter 2, the Korean and Dutch listeners had to recognize the emotions in their own language. The analysis focused on the recognition accuracies, not further analyzing the confusion matrices. We present all confusion matrices in Table II in Appendix I. These confusion matrices may be used to investigate which emotions are more similar and which are more dissimilar. This is especially relevant in comparing the Dutch and Korean data when listeners evaluate the emotional expressions in their own language. Is the similarity structure of the emotions in the two languages the same?

We computed two hierarchical clustering dendrograms on the basis of the frequencies in the confusion matrices, i.e., one for the Dutch listeners' in-group and one for the Korean listeners' in-group, using the *vegdist* function from the *vegan* package in R (Oksanen et al., 2024). First, we calculated the distance between each pair of emotions in terms of a chi-square distance, where more similar emotions have smaller distances. Second, we used these chi-square distances to construct the clustering diagrams, applying Ward's method (Contreras & Murtagh, 2015; Vichi et al., 2022). The distances between the pairs of emotions can be found in the distance matrices in Appendix M.

The two resulting dendrograms are connected in Figure 6.1. The position along the y-axis in the plots represents the distances. Higher distance values represent greater dissimilarities. Zero is the smallest distance (= complete similarity) between any two emotions in the analysis, while 3.0 is assigned to the distance between the two topmost nodes in the tree.

The dendrograms show that Dutch and Korean listeners display similar recognition patterns when identifying vocal emotions produced in their native language, consistent with the universality hypothesis. For both listener/speaker groups, Joy and Pride, Fear and Sadness, as well as Anger and Irritation are clustered pairwise at the lowest level, revealing that the members

of these pairs share similar vocalization patterns. Joy and Pride are farthest from Anger and Irritation, indicating that these two pairs of vocal emotions share the least similarities and are most distinct from each other. For Dutch listeners, Tenderness and Relief are clustered together, forming the first branch with the Joy + Pride cluster. In contrast, for Korean listeners, Relief first joins the Joy + Pride cluster, then forms the second cluster with Tenderness. Nevertheless, the Joy-Pride-Relief-Tenderness branch forms the lowest-level main cluster in both language groups. It appears that the Dutch and Korean listeners identified the vocal emotions in largely the same way.

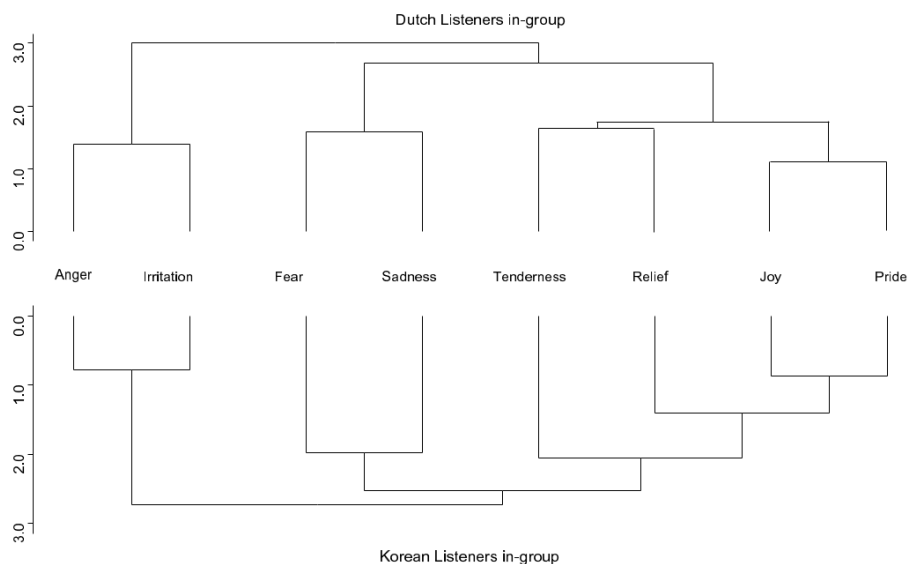


Figure 6.1. Hierarchical clustering dendrograms (Ward’s method) for in-group identification of vocal emotions by Dutch (upper tree) and Korean (lower tree) listeners.

6.3.3 In-group advantage

As shown in the dendrograms, Dutch and Korean listeners display similar/symmetrical recognition patterns when identifying vocal emotions produced in their native language, consistent with the universality hypothesis. The recognition discrepancy between Dutch and Korean listeners can be partly explained by culturally specific “display rules”, such that the expression and perception of emotions is shaped by social and cultural norms (Ekman & Friesen, 1969). Although the overall trends in recognition between these two

listener groups display similarity, the subtle differences in the perception of non-basic emotions like Tenderness and Pride demonstrate cultural norms. For instance, Korean conventions emphasize emotional restraint, whereas Dutch norms encourage more overt emotional expression. These cultural differences can partly account for the differences in recognition accuracy. The hierarchical ordering of the emotions shows many resemblances. For both languages, Joy and Pride, Fear and Sadness, as well as Anger and Irritation, are clustered pairwise, within the same branch. The ordering of Tenderness and Relief is different between the two languages, but nevertheless comparable.

Lower values on the y-axis mean more mutual confusion between the two emotions involved and the same confusion patterns. The pair of Anger/Irritation is more confusing in Korean than in Dutch, whereas the situation is the other way around for Fear and Sadness. These differences in similarity reflect differences between the two languages, but more remarkable is the same structure of the dendrograms. These outcomes strongly support the universality hypothesis, but show at the same time that there are differences between languages in expressing vocal emotions. This conclusion is supported by the outcomes of the French and American listeners.

6.3.4 Prosodic structure and vocal emotion recognition

As shown in previous studies, prosodic structures, such as stress and temporal patterns, play a pivotal role in vocal emotion recognition, especially in a cross-cultural setting. However, in Chapter 5, we did not find a contribution of prosodic structure to vocal emotion recognition. Possibly, the short stimulus sentence used in the study, with only simple syllable structures CV(C), did not provide sufficient opportunities for prosodic differences between Dutch and Korean to show up. Thus, further studies should examine the extent to which the variations of prosodic structure affect vocal emotion recognition. Moreover, further studies may also investigate the interaction between word stress and rhythm and other aspects of prosodic structure, such as word tones and sentence melody, to examine how the joint effects affect vocal emotion recognition.

6.3.5 Acoustic parameters identifying vocal emotions

In Chapter 4, 17 acoustic parameters were included. However, it might have been better to include even more acoustic parameters to examine the influence of acoustic parameters on emotion perception. Although vocal emotions are recognized accurately above chance across cultures (Laukka et al., 2016), it is

challenging to list all reliable acoustic cues that differentiate emotions (Scherer, 1986). In our study, Support Vector Machine (SVM) models were built to examine whether recognition accuracy could be reliably predicted from a configuration of the 17 acoustic parameters. SVMs are very powerful and yield high correct classification rates. Unfortunately, SVM technology provides no automatic way to extract profiles that concisely characterize the recognition categories and does not allow the researcher to establish the relative contributions of each acoustic parameter to recognition accuracy.

Overall, the accuracy of the identification of the eight emotion types in Chapter 4 by the SVM models was not systematically different from that obtained from our in-group human listeners, at least not when proper cross-validation (Leave One Out, LOOCV) was applied to prevent test tokens from being included in the training set of the SVM.

We would argue that, at this moment, there is no better way to identify intended emotions in human speech than by asking a group of human listeners who share the speaker's cultural and linguistic profile. Consequently, we expect in-group human emotion identification to be better than automatic identification of emotion type by properly cross-validated machine learning. Ideally, the performance of the machine should be as good as that of the human listener, but never superior. If the (in-group) human identification of an emotion type is better than that by machine, the human listener must have had access to information that the machine did not have, either because the human brain has set up specialized networks for emotions that are still beyond the possibilities of machine learning, or because the humans have information that was not included in the set of 17 parameters measured in Chapter 4.

One way to decide whether there is useful acoustic information in the signals that humans employ but which we may have missed in our set of 17 extracted parameters would be to delegate the parameter extraction to the machine, and then see whether the machine's recognition of the emotions improves. This can be done by feeding the machine not with the extracted parameters but by giving it direct access to relatively fine-grained spectra-temporal acoustic features, for instance a large number (12 or even 20) of Cepstral Coefficients (after perceptual scaling in Mel), the MFCCs (Mel Frequency Cepstral Coefficients), the currently prevalent front end used for automatic speech and speaker recognition through Deep Neural Networks (DNNs).

At the same time, however, we observed that about half of the set of emotion types in our research were better recognized by SVM than by humans, while the reverse was true of the other half. Specifically, Anger, Fear, and

Tenderness were clearly better identified by SVM than by in-group human listeners, both for Dutch and for Korean speech (see Table 4.6). In light of the above reasoning, this should not have happened. Additional analyses should therefore be carried out to ascertain whether the result may have been due to overly lenient cross-validation. For instance, each voice actor contributed two tokens of each emotion. It would make sense to assume that the same actor portrayed the target emotion in the same way twice. Any idiosyncrasies of an individual voice actor will then be incorporated into the model when it is tested with LOOCV. Yet, in the signaling of vocal emotions, we would expect the members of a linguistic community (including voice actors) to express their emotions in approximately the same way, to avoid misinterpretation of affect. Possibly, then, we should attempt a stricter method of cross-validated testing of the SVM models by leaving out both tokens of the same intended emotion for each speaker in turn. The tokens of each voice actor will then be identified on the basis of the regularities found in the corresponding tokens of the seven different voice actors remaining in the training set.

For all the virtues of the SVM machine learning approach, a drawback of the technique is that it provides no easily interpretable characterization of how the eight emotions can be differentiated. More traditional classification algorithms afford just that. As an illustration of one such method, Table 6.2 presents the results from a series of Linear Discriminant Analyses (LDA, Klecka, 1980) (green = in-group).

Table 6.2. Correct emotion identification (% with LOO cross-validation) by LDA for various combinations of training and test sets (in-group testing in cells with green highlight). Acoustic parameters contributing significantly are listed from left to right in descending order of importance. For the legend of abbreviations, see Table 4.2.

| Language | | LDA (chance = 12.5%) | | | | | | |
|----------|--------|----------------------|---|------|--------|----|--------|----|
| Training | Test | Correct | Parameters (in order of inclusion/importance) | | | | | |
| Dutch | Dutch | 48.4 | Int-M | | F0-min | | Int-SD | HR |
| Dutch | Korean | 23.4 | | | | | | |
| Korean | Korean | 37.5 | | F0-M | F0-min | AR | Int-SD | |
| Korean | Dutch | 35.9 | | | | | | |

The eight target emotions are identified well above chance level, even in the poorest condition, i.e., when the LDA is trained on the Korean speech data and then tested on the Dutch tokens. As with the SVM, the model trained on the Dutch speech data performs better than the Korean model, both when

applied in-group and out-group. Independent of this, the in-group identification is considerably better than the out-group identification. These effects run parallel to the human identification results as well as to the performance of the SVM in Chapter 4. The advantage is that the LDA here identifies just a small set of up to six acoustic parameters that play a role in the discrimination among the eight emotions. Moreover, the rank order of the parameters seems to suggest that pitch-related properties are the best discriminators (mean pitch and bottom pitch), followed by articulation rate, the latter suggesting that some emotions, especially in Korean, are differentiated by faster vs. slower speech. Variability in intensity is a consistent discriminator, followed by two parameters that relate to the behavior of the vocal folds, i.e., differences in breathiness (HNR) and instability of the glottal vibration (trembling voice, PPQ). Note, finally, that the performance of the LDA is not necessarily poorer than that of the SVM. The Dutch-Dutch condition is clearly better in the SVM, but the differences in the other conditions are minor.

6.4 General conclusions

This dissertation investigated cross-cultural vocal emotion recognition by four groups of listeners—Dutch, Korean, American English, and French listeners. Although the native languages of these four groups of listeners differ, they displayed similar recognition patterns in cross-cultural and cross-language vocal emotion recognition. All four listener groups identified vocal emotions above chance, within and across cultures, although American English and French listeners' native language is neither Dutch nor Korean. Dutch and Korean listeners exhibited an in-group advantage when listening to the stimuli. Finally, all vocal emotions were analyzed acoustically in terms of five groups of acoustic parameters (pitch, amplitude, spectral distribution, duration, and laryngeal properties), and these parameters were examined for their relative contributions to recognition using a series of linear mixed-effects models.

6.4.1 The Cultural Proximity hypothesis

Dutch and Korean listeners recognized vocal emotions more accurately in their native language than in the unknown language (i.e., in-group advantage). We further found that American English and French listeners recognized vocal emotions more accurately in the culture that is more closely related to their own Western (rather than Asian) background. These findings are consistent with the Culture Proximity hypothesis (Elfenbein & Ambady, 2003a).

6.4.2 The Prosodic Proximity hypothesis

Dutch and Korean listeners identified vocal emotions more accurately in their native language than in the unknown language, which aligns with the Language Distance hypothesis (Scherer et al., 2001). Moreover, American English and French listeners recognized vocal emotions more accurately in Dutch than in Korean. However, French listeners did not outperform American English listeners when listening to Korean recordings. According to the Prosodic Proximity hypothesis, listeners can recognize vocal emotions more accurately in languages that are typologically similar to their native language than in typologically different ones, especially in prosodic structure. Although French is more similar to Korean in prosodic structure than Dutch, French listeners did not obtain higher recognition accuracy in Korean than in Dutch.

6.4.3 Acoustic parameters of vocal emotions

The vocal emotions studied in this dissertation displayed characteristically different acoustic patterns. For example, high-arousal emotions like Anger, Fear, and Joy exhibited a higher pitch level, wider pitch range, and larger intensity variations than Sadness, Tenderness, Relief, Irritation, and Pride. In terms of articulation rate, Fear and Joy were spoken faster than the other six emotions. Fear showed the most stable voice quality compared to other emotions, as it had the lowest APQ and PPQ. Emotion-specific acoustic patterns varied across speaker language and gender. Nevertheless, the recognition accuracy for Anger, Fear, Joy, and Sadness was consistently higher than other emotions across Dutch and Korean listeners, suggesting that pitch-, amplitude-, and duration-related parameters are pivotal in the perception of vocal emotions, consistent with Juslin and Laukka's (2003) findings.

6.4.4 Dimensionality of vocal emotions

The cluster analysis (dendrograms in Figure 6.1) revealed that vocal emotions were not only recognized in a discrete approach but also showed an underlying dimensional structure. Anger + Irritation and Joy + Pride are the farthest from each other, demonstrating that these two pairs of vocal emotions are easily distinguished from each other, as they share the least vocal similarities. However, high-arousal and positive emotions (Joy-Pride), negative-basic emotions (Fear-Sadness), and other (observer-directed) negative emotions (Anger-Irritation) cluster pairwise, suggesting that the emotions in each pair share similar vocal cues, making them difficult to distinguish.

6.5 Limitations

This dissertation presents a comprehensive investigation of cross-cultural vocal emotion recognition in two typologically different languages—Dutch and Korean, by integrating both discrete and dimensional approaches. In the following subsections, I will raise possible shortcomings of the research carried out in the preceding chapters and suggest ways to remedy the weaknesses in future experiments.

It is important to realize that the various studies were undertaken at rather different points in time. The primary data were collected some 15 years ago, while additional experiments and acoustic analyses were carried out 10 years later. In retrospect, some of the earlier decisions concerning experimental designs, methods of data collection, and analyses seem less than optimal. In the following (sub)sections, therefore, I will identify weaknesses in the various experiments and analyses, and suggest future experiments to remedy such weaknesses and/or answer new questions that can be formulated in the wake of the research reported in this dissertation.

6.5.1 Phonological legitimacy

An important issue in the present series of experiments was the contribution to our understanding of cross-cultural identification of vocal emotions. Specifically, we were interested in the question of whether vocal emotions are better identified when speaker and hearer belong to the same cultural and linguistic community than when the listener has received no prior exposure to the culture and/or language of the speaker. This is the issue of the in-group advantage. To test the in-group advantage hypothesis, the speech stimuli should be phonologically as similar as possible in the spoken language(s) (Matsumoto, 2002), since dissimilar and incompatible stimuli will produce confounds that affect the processing of vocal emotions. For this reason, the stimulus sentence designed for the Demo-Koremo vocal emotion corpus was chosen to be phonologically neutral, i.e., legal in both Dutch and Korean. It should be emphasized that the requirement made by Matsumoto (2002) specifically concerned stimuli to be used for the identification of facially expressed emotions, arguing that exactly the same set of facial muscles should be involved in the emotions portrayed by members of the different cultures under comparison. In retrospect, one may wonder why this requirement would translate to phonological compatibility in the context of signaling and identifying vocal emotions. Presumably, if the observer is alerted to the fact that the person signaling an emotion hails from a different culture—either due to different facial features, hair style and/or skin color in the visual domain, or

to the use of unfamiliar sounds, rhythms and/or melodies in the speech domain—they will realize they cannot judge the signals against the background defined by their own past experience. Korean allows for simple consonant-vowel structures without vowel length, while Dutch has complex syllable structures with vowel length. In the speech domain, therefore, it will be hazardous for a Korean listener (with no length contrast in the vowels) to determine whether the local speech rate goes up or down in Dutch, a language with long and short vowels. Consequently, the hearer will try to evaluate the emotional signals in a more general (universal) mode when confronted with an obvious out-group speaker.

According to Goudbeek and Broersma (2010a, b), the pseudo-sentence we used, /nuto hɔm sepikaŋ/, is phonologically legal in both Dutch and Korean. However, the rhyme [aŋ] is not allowed in Dutch. A tense vowel such as [a] cannot be followed by a coda [ŋ]. Therefore, we suppose that the pseudo-sentence should be interpreted as /nuto hɔm sepikaŋ/, which is indeed how the Dutch voice actors pronounced it consistently. In Korean, vowel qualities [a] and [ɑ] are interchangeable, representing the same phoneme (Shin, 2015). The pseudo-sentence is phonologically illegal in Korean as well, since the vowel [ɔ] does not exist in Korean. However, the vowel qualities [ɔ] and [o] are both acceptable realizations of the Korean back mid vowel phoneme /o/. Therefore, the pseudo-sentence is neither phonologically legal in Dutch nor in Korean, but contains one deviant (but acceptable) vowel in either language.

But even if the pseudo-sentence were phonologically legal in both Dutch and Korean, the details of the pronunciation, rhythm, and melody would immediately reveal the origin of the speaker and alert the listener that they should engage the in-group or out-group listening mode. It seems unrealistic, therefore, to assume that the listeners in our experiments used their normal, i.e., native, language-specific emotion assessment when responding to out-group stimuli.

6.5.2 The eight emotions

Classic emotion theory recognizes six basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992b). In this set, there are more negative than positive emotions, with happiness being positive, while surprise can be either a positive or a negative emotion. In the present study, there are four basic (anger, fear, joy/happiness, sadness) and four non-basic emotions (irritation, pride, relief, tenderness). The eight emotions involved in this study are equally divided between positive/negative (valence), high/low (arousal), and basicness (basic, compound) (see Table 1.1), but in a non-orthogonal

fashion, which has made it awkward to come up with a straightforward factorial analysis of the effects of this three-way binary split.

The present set of eight emotions cannot be arranged such that the arousal, valence, and basicness factors can be combined orthogonally in a factorial design. An orthogonal design may, however, be obtained by replacing some of the categories in the design (see Table 6.1).

6.5.3 The neutral category

In Chapters 2 and 3, the respondents were asked to identify the stimulus emotions as one of eight categories (the spokes of the emotion wheel), and to indicate how strongly they felt the emotion of their choice was expressed by the speaker (by selecting a more or less peripheral position along the corresponding spoke). Only if they could not decide on a particular emotion, were they allowed to select “Neutral” as a ninth response option in the bull’s eye of the emotion wheel—with no intensity at all. We should bear in mind here that the speakers were explicitly instructed to produce one of eight different emotions and to produce each token as convincingly as they could. Speakers were never asked to portray a weaker version of an emotion, and least of all to speak in a neutral tone of voice. Of course, speakers could have been asked to imitate the speaking style of a newsreader, but this was not done when the stimuli were recorded. It is unexpected, therefore, that for some emotion portrayals, native listeners could not decide on one of the eight targeted emotion response categories and chose “Neutral”. In hindsight, we argue such responses should not be interpreted as neutral. Rather, the neutral option in the early chapters signals that the listener could not identify the stimulus as a token of one of the eight target emotions, so that a more adequate characterization of the option would be “none of the above” or “other”—but not necessarily “neutral”.

In total, 7 percent of the responses in Chapters 2 and 5 were “Neutral”. We would now argue that these choices should have been treated as missing responses, rather than as a ninth emotion. In that case, the chance level for the emotion recognition accuracy would have been 1 out of 8 (i.e., 12.5%) rather than 1 out of 9 (11.1%). Since the reports of these experiments were already published, we decided to include the original analyses in the dissertation.

In Chapter 4, however, we used machine learning to identify the eight emotions produced by the Dutch and Korean voice actors while simulating Dutch and Korean listeners. In our machine learning approach, only the stimulus categories were possible response categories, so that “Neutral” could

not be an option. Under these circumstances, the chance level for correct emotion identification is 1 out of 8, i.e., 12.5%, which is the level we used to evaluate the performance of emotion identification by the Support Vector Machine. In the same chapter, we also compared the identification accuracy obtained by SVM and by the human listeners employed in Chapters 2 and 3. To ensure a fair comparison, the “neutral” responses in the human subset were this time treated as missing data, so that for both datasets, machine and human, the chance level was 12.5%.

6.5.4 Acoustic correlates of emotional intensity

One of the goals of the present dissertation was to come to grips with the in-group and out-group perception of vocal expression of emotional intensity. The results obtained in Chapter 3 show that stimuli perceived by the listeners as expressing an emotion with great intensity were also more likely to be identified accurately (i.e., as intended by the speaker) than emotional tokens perceived with low(er) intensity. This effect was found both for in-group and out-group perception of emotions, by Dutch and Korean respondents alike.

It is not uncommon in research on speech perception to ask respondents how confident they are that their response is correct. We suggest that an alternative way to measure the perceived strength of emotional intensity would be to instruct listeners to identify the emotional token, with forced choice from a closed set of alternatives, and then to indicate the level of confidence that they identified the emotion correctly as intended by the speaker. The confidence level is expected to correlate strongly with intensity scores, such as those collected in Chapters 2 and 3.

In Chapter 4, we investigated in substantial detail the possible acoustic correlates of the eight emotional categories targeted in our research. The results of that chapter indicate that each of the eight emotion types has a profile of acoustic characteristics, which distinguishes it from the other seven types, and that the profiles, in spite of a common core, differ considerably between Dutch and Korean. What we did not do is complement this enterprise with a similar investigation of the acoustic correlates of perceived emotional strength.

Emotional intensity, as used in our studies, should not be confused with acoustic intensity, i.e., the physical property of a (speech) sound, expressed in decibels, that is often associated with loudness. If, for instance, a low-arousal emotion such as sadness is perceived as being expressed with great intensity (or strength), it is unlikely to be spoken in a loud voice with a lot of acoustic intensity. It is not clear, at this time, what acoustic properties of an emotional

token drive the observer's perceived strength of the emotion, but it seems a viable enterprise to pursue this matter in future work. I will sketch one possible approach in the following paragraph.

Perceived emotional intensity, in Chapter 3, was estimated by our listeners on a scale from 1 to 4 (= strongest). The emotional intensity of each of the 128 tokens per language (Dutch or Korean), separately for in-group and out-group judgments, can be averaged into a scalar variable, which we can then try to predict from some combination of acoustic measurements such as those reported in Chapter 4. The prediction of perceived emotional intensity may well employ different subsets of acoustic parameters depending on the emotional type. The number of tokens for which we have perceived strength scores is no more than 16 per type per language (8 speakers per language, who produced each type twice). This yields a rather limited dataset, so the attempt would be a pilot test at best.

6.5.5 The impact of stimulus order on recognition accuracy

One limitation of this study is the language order in which the stimuli were presented. In both perception experiments, all listeners were presented with the 128 Korean portrayals before the Dutch ones, with stimuli randomized within each language block. As shown in Appendices N and O, mean accuracy across the 256 stimuli, plotted as a time series, reveals only a slight upward trend over time. For all listeners, accuracy displays a modest drop in the middle followed by a gradual increase, forming a shallow U-shaped pattern. This variation is relatively small compared to the overall range of accuracy scores (as shown by the blue and red points). Importantly, there is no clear discontinuity between the Korean and Dutch blocks, indicating that any effect of sequential order on recognition accuracy was limited.

6.5.6 The ecological validity of stimuli

Another limitation of this study lies in its ecological validity. The stimuli we used were derived from acted speech, which may differ from spontaneous speech produced by people in daily life. Although acted speech provides a controlled way for comparisons across languages and cultures, it may not capture the full variability and authenticity of natural speech. Therefore, further studies should extend this research by incorporating spontaneous or semi-natural speech to explore whether the same cross-cultural patterns align under more ecologically valid conditions.

6.6 Future research

Findings from this study may have important implications for second language acquisition. So far, cross-cultural emotion recognition studies have mainly concentrated on native listeners, i.e., members of the same cultural and linguistic community as the speaker of the emotional utterance. Non-native listeners in cross-cultural and cross-lingual studies of emotion perception were always selected so as to meet the requirement that they had no prior exposure to the non-native language, a precaution that was also taken in the studies reported in this dissertation. From the perspective of an arts faculty, with a rich variety of foreign language programs, it is both a disappointment and a challenge to see that only a few studies have examined the expression and recognition of emotion by second and foreign language learners (Min & Schirmer, 2011; Wu et al., 2022; Zhu, 2013). Findings highlight the role of three dimensions (arousal, valence, and basicness) in cross-cultural emotion recognition.

6.6.1 Second language acquisition

Extending from the current study, we can further investigate emotion recognition in second language learners. Zhu's (2012, 2013) results revealed that Dutch university students specializing in Mandarin Language and Culture were significantly more adept (and almost as successful as native Mandarin listeners) at identifying vocally expressed emotions in Mandarin than Dutch listeners with no prior exposure to Mandarin. Clearly, the advanced Dutch learners of Mandarin had internalized at least part of the language/culture-specific signaling of emotions of Mandarin. It would be a challenge to develop teaching methods and materials to help foreign language learners in general to effectively recognize (and maybe even actively produce) vocal emotions in the foreign language.

6.6.2 Acoustic manipulations of stimuli

Using speech-technological tools, however, it should be feasible to generate stimulus materials that would be perfectly native in terms of the pronunciation, temporal organization, and yet contain the rhythmic and melodic properties of a foreign speaker signaling specific emotions in another language. This would require being able to strip the emotional details from an utterance, thereby reducing the utterance to a newsreader's neutral delivery, and exporting only the emotional characteristics to another language.

6.6.3 Neuroimaging

Moreover, since emotions affect attention, working memory, and cognition (Okon-Singer et al., 2015), further studies on emotions should use neuroimaging techniques, focusing on the dynamic neural networks of emotion processing and language acquisition. Findings in this field not only contribute to our understanding of how acoustic parameters affect the expression and identification of vocal emotions but also have practical applications in Language and Speech Technology, i.e., human-computer interfaces using speech synthesis (Murray & Arnott, 1993) and automatic speech understanding (Hashem et al., 2023).

References

- Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7*(4), 481–490. <https://doi.org/10.1177/002202217674009>
- Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language, 25*(3), 556–570. <https://doi.org/10.1016/j.csl.2010.10.001>
- Anolli, L., Wang, L., Mantovani, F., & De Toni, A. (2008). The voice of emotion in Chinese and Italian young adults. *Journal of Cross-Cultural Psychology, 39*(5), 565–598. <https://doi.org/10.1177/0022022108321178>
- App, B., McIntosh, D. N., Reed, C. L., & Hertenstein, M. J. (2011). Nonverbal channel use in communication of emotion: How may depend on why. *Emotion, 11*(3), 603–617. <https://doi.org/10.1037/a0023164>
- Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*(3), 351–373. <https://doi.org/10.1016/j.wocn.2012.02.003>
- Bachorowski, J. A., & Owren, M. J. (2003). Sounds of emotion: Production and perception of affect-related vocal acoustics. *Annals of the New York Academy of Sciences, 1000*(1), 244–265. <https://doi.org/10.1196/annals.1280.012>
- Bachorowski, J. A., & Owren, M. J. (2008). Vocal expressions of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 196–210). New York: Guildford Press.
- Bailey, W., Nowicki, S., & Cole, S. P. (1998). The ability to decode nonverbal information in African American, African and Afro-Caribbean, and European American adults. *Journal of Black Psychology, 24*(4), 418–431. <https://doi.org/10.1177/00957984980244002>
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. <https://doi.org/10.1037/0022-3514.70.3.614>
- Bänziger, T., Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. *International Conference on Affective Computing and Intelligent Interaction* (pp. 476–487). Heidelberg, Springer. https://doi.org/10.1007/978-3-540-74889-2_42

- Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161–1179.
<https://doi.org/10.1037/a0025827>
- Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication, 46*(3–4), 252–267.
<https://doi.org/10.1016/j.specom.2005.02.016>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.
<https://doi.org/10.1016/j.jml.2012.11.001>
- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion, 12*(4), 579–599.
<https://doi.org/10.1080/026999398379574>
- Barrett, L. F., & Russell, J. A. (2014). *The psychological construction of emotion*. Guilford Publications.
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1).
<https://doi.org/10.18637/jss.v067.i01>
- Baum, K. M., & Nowicki, S. (1998). Perception of emotion: Measuring decoding accuracy of adult prosodic cues varying in intensity. *Journal of Nonverbal Behavior, 22*(2), 89–107.
<https://doi.org/10.1023/A:1022954014365>
- Beier, E., & Zautra, A. J. (1972). The identification of vocal expressions of emotion across cultures. *Journal of Consulting and Clinical Psychology, 39*(1), 166. <https://doi.org/10.1037/h0033170>
- Bertan, A. P. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages A. *Language Design, 2*, 103–130.
- Best, C. T. (1995). A direct realistic view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore: York Press.
- Bhatara, A., Laukka, P., Boll-Avetisyan, N., Granjon, L., Efenbein, H. A., & Bänziger, T. (2016). Second language ability and emotional prosody perception. *PLoS ONE, 11*(6), 1–13.
<https://doi.org/10.1371/journal.pone.0156855>
- Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian facial expressions of emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*(1), 3–21.
<https://doi.org/10.1023/A:1024902500935>

- Boersma, P., & Weenink, D. (1996). Praat, a system for doing phonetics by computer. *Report*, 132. Institute of Phonetic Sciences University of Amsterdam.
- Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. *Glott International*, 5(9–10), 341–347.
- Bonebright, T. L., Thompson, J. L., & Leger, D. W. (1996). Gender stereotypes in the expression and perception of vocal affect. *Sex Roles*, 34(5–6), 429–445. <https://doi.org/10.1007/BF01547811>
- Borchert, P., & Zellmer-Bruhn, D. M. (2010). Reproduced with permission of the copyright owner. Further reproduction prohibited without. *Journal of Allergy and Clinical Immunology*, 130(2), 556. <http://dx.doi.org/10.1016/j.jaci.2012.05.050>
- Brehm, J. W. (1999). The intensity of emotion. *Personality and Social Psychology Review*, 3(1), 2–22. https://doi.org/10.1207/s15327957pspr0301_1
- Breitenstein, C., Van Lancker, D., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion*, 15(1), 57–79. <https://doi.org/10.1080/0269993004200114>
- Broersma, M.E., Goudbeek, M., Choi, J., Konopka, A. (2025). Demo/Koremo corpus for Dutch and Korean emotional speech. Version 1. Radboud University. (dataset). <https://doi.org/10.34973/5kg3-9852>
- Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, 8(1–2), 135–148. <https://doi.org/10.1163/156770908X289242>
- Buller, D. B. (2005). Methods for measuring speech rate. In V. Manusov (Eds.), *The sourcebook of nonverbal measures: Going beyond words* (pp. 317–324). Lawrence Erlbaum Associates.
- Burgoon, J. K., Manusov, V., & Guerrero, L. K. (2016). *Nonverbal communication*. Routledge.
- Cahn, J. E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8, 1–19.
- Carl, M., Icht, M., & Ben-David, B. M. (2022). A cross-linguistic validation of the test for rating emotions in speech: Acoustic analyses of emotional sentences in English, German, and Hebrew. *Journal of Speech, Language, and Hearing Research*, 65(3), 991–1000. https://doi.org/10.1044/2021_JSLHR-21-00205
- Celeghin, A., Diano, M., Bagnis, A., Viola, M., & Tamietto, M. (2017). Basic emotions in human neuroscience: Neuroimaging and beyond. *Frontiers in Psychology*, 8(AUG), 1–13. <https://doi.org/10.3389/fpsyg.2017.01432>

- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.
<https://doi.org/10.1186/s12864-019-6413-7>
- Chronaki, G., Wigelsworth, M., Pell, M. D., & Kotz, S. A. (2018). The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Scientific Reports*, *8*(1), 1–17.
<https://doi.org/10.1038/s41598-018-26889-1>
- Chung, S. J. (1999). Vocal expression and perception of emotion in Korean. *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, 969–972.
https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0969.pdf
- Cimino, C. R. (2002). The neuropsychology of emotion. *The Journal of Nervous and Mental Disease*, *190*(1), 56–57.
<https://doi.org/10.1097/00005053-200201000-00016>
- Clark, M. S., Milberg, S., & Erber, R. (1984). Effects of arousal on judgments of others' emotions. *Journal of Personality and Social Psychology*, *46*(3), 551–560. <https://doi.org/10.1037/0022-3514.46.3.551>
- Clarke, E. F. (1999). Rhythm and timing in music. In D. Deutsch (Ed.), *The psychology of music* (pp. 473–500). Academic Press.
- Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*(6), 773–780. <https://doi.org/10.1006/anbe.2000.1523>
- Contreras, P., & Murtagh, F. (2015). Hierarchical clustering. In *Handbook of cluster analysis* (pp. 103–124). CRC Press.
<https://doi.org/10.1201/b19706>
- Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion*, *16*(1), 117–128.
<https://doi.org/10.1037/emo0000100>
- Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(6), 864–881. <https://doi.org/10.1037/0096-1523.9.6.864>
- Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), 7900–7909.
<https://doi.org/10.1073/pnas.1702247114>
- Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382.
<https://doi.org/10.1038/s41562-019-0533-6>

- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1), 32–80. <https://doi.org/10.1109/79.911197>
- Crystal, D. (1969). *Prosodic systems and intonation in English*. CUP Archive.
- Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed.). London, England: John Murray. (Original work published 1872).
- Dehghani, A., Soltanian-Zadeh, H., & Hossein-Zadeh, G. A. (2023). Neural modulation enhancement using connectivity-based EEG neurofeedback with simultaneous fMRI for emotion regulation. *NeuroImage*, 279, 120320. <https://doi.org/10.1016/j.neuroimage.2023.120320>
- Dromey, C., Silveira, J., & Sandor, P. (2005). Recognition of affective prosody by speakers of English as a first or foreign language. *Speech Communication*, 47(3), 351–359. <https://doi.org/10.1016/j.specom.2004.09.010>
- Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1), 49–98. <https://doi.org/10.1515/semi.1969.1.1.49>
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J.K. Cole (Ed.), *Nebraska Symposium on Motivation* (pp. 207–283). University of Nebraska Press.
- Ekman, P. (1992a). An Argument for Basic Emotions. *Cognition and Emotion*, 6(3–4), 169–200. <https://doi.org/10.1080/02699939208411068>
- Ekman, P. (1992b). Are there basic emotions? *Psychological Review*, 99(3), 550–553. <https://doi.org/10.1037/0033-295X.99.3.550>
- Ekman, P. (1999). Basic emotions. In T. Dalgleish & T. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Wiley.
- Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on Psychological Science*, 11(1), 31–34. <https://doi.org/10.1177/1745691615596992>
- Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4), 364–370. <https://doi.org/10.1177/1754073911410740>
- Ekman, P., Friesen, W. V., O’Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4), 712–717. <https://doi.org/10.1037/0022-3514.53.4.712>
- Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875), 86–88. <https://doi.org/10.1126/science.164.3875.86>

- Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review*, 5(1), 90–96.
<https://doi.org/10.1177/1754073912451332>
- Elfenbein, H. A., & Ambady, N. (2002a). Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, 128(2), 243–249.
<https://doi.org/10.1037/0033-2909.128.2.243>
- Elfenbein, H. A., & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2), 203–235.
<https://doi.org/10.1037/0033-2909.128.2.203>
- Elfenbein, H. A., & Ambady, N. (2003a). Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *Journal of Cross-Cultural Psychology*, 34(1), 92–110.
<https://doi.org/10.1177/0022022102239157>
- Elfenbein, H. A., & Ambady, N. (2003b). Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, 12(5), 159–164. <https://doi.org/10.1111/1467-8721.01252>
- Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion*, 7(1), 131–146.
<https://doi.org/10.1037/1528-3542.7.1.131>
- Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2002). Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion*, 2(1), 75–84.
<https://doi.org/10.1037/1528-3542.2.1.75>
- Elfenbein, H. A., Marsh, A. A., & Ambady, N. (2002). Emotional intelligence and the recognition of emotion from facial expressions. In L. F. Barrett & P. Salovey (Eds.), *The wisdom in feeling: Psychological processes in emotional intelligence* (pp. 37–59). Guilford Press.
- Engelberg, E., & Sjöberg, L. (2004). Emotional intelligence, affect intensity, and social adjustment. *Personality and Individual Differences*, 37(3), 533–542. <https://doi.org/10.1016/j.paid.2003.09.024>
- Ezhilarasi, R., & Minu, R. I. (2012). Automatic emotion recognition and classification. *Procedia Engineering*, 38, 21–26.
<https://doi.org/10.1016/j.proeng.2012.06.004>
- Farrús, M., Hernando, J., & Ejarque, P. (2011). Jitter and shimmer measurements for speaker recognition. *Proceedings of Interspeech 2007* (pp. 778–781). <https://doi.org/10.21437/Interspeech.2007-147>
- Fenk-Oczlon, G., & Pilz, J. (2021). Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication*, 6, Article 4, 1-7. <https://doi.org/10.3389/fcomm.2021.626032>

- Flett, G. L., Boase, P., Mcandrews, M. P. A. T., & Blankstein, K. R. (1986). Affect intensity and the appraisal of emotion. *Psychological Reports*, 20(4), 447–459. [https://doi.org/10.1016/0092-6566\(86\)90125-X](https://doi.org/10.1016/0092-6566(86)90125-X)
- Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, 18(4), 389–405. <https://doi.org/10.1016/j.neunet.2005.03.006>
- Frick, R. W. (1985). Communicating emotion. The role of prosodic features. *Psychological Bulletin*, 97(3), 412–429. <https://doi.org/10.1037/0033-2909.97.3.412>
- Frijda, N. H., Ortony, A., Sonnemans, J., & Clore, G. L. (1992). The complexity of intensity: Issues concerning the structure of emotion intensity. In M. S. Clark (Ed.), *Emotion: Review of personality and social psychology* (Vol. 13, pp. 60-89). Sage.
- Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, 27(4), 211–219. <https://doi.org/10.1177/0963721417746794>
- Gorris, C., Ricci Maccarini, A., Vanoni, F., Poggioli, M., Vaschetto, R., Garzaro, M., & Aluffi Valletti, P. (2020). Acoustic analysis of normal voice patterns in Italian adults by using Praat. *Journal of Voice*, 34(6), 961.e9-961.e18. <https://doi.org/10.1016/j.jvoice.2019.04.016>
- Goudbeek, M. (2010). The Demo / Kemo corpus: Many-to-many. May, 19–21.
- Goudbeek, M., & Broersma, M. (2010a). Language specific effects of emotion on phoneme duration. *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, 2026–2029. https://www.isca-archive.org/interspeech_2010/goudbeek10_interspeech.html
- Goudbeek, M., & Broersma, M. (2010b). The Demo/Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Seventh Conference on International Language Resources and Evaluation* (pp. 2211–2215). <http://www.lrec-conf.org/proceedings/lrec2010/index.html>
- Goudbeek, M., & Scherer, K. R. (2010). Beyond arousal: Valence and potency/ control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3), 1322-1336. <https://doi.org/10.1121/1.3466853>

- Großwendt, A., & Schmidt, M. (2019). Analysis of Ward's Method. *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 2939–2957).
<https://doi.org/10.1137/1.9781611975482.182>
- Gu, S., Wang, F., Cao, C., Wu, E., Tang, Y. Y., & Huang, J. H. (2019). An integrative way for studying neural basis of basic emotions with fMRI. *Frontiers in Neuroscience, 13*(6), 1–12.
<https://doi.org/10.3389/fnins.2019.00628>
- Guerrero, L. K., & La Valley, A. G. (2006). Conflict, emotion, and communication. In J. G. Oetzel & S. Ting-Toomey (Eds.), *The SAGE handbook of conflict communication: Integrating theory, research, and practice* (pp. 69–96). Sage Publications.
- Gussenhoven, C. (1993). The Dutch foot and the chanted call. *Journal of Linguistics, 29*(1), 37–63.
<https://doi.org/https://www.jstor.org/stable/pdf/4176207>
- Gussenhoven, C. (2005). Transcription of Dutch intonation. In S. A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 118–145). Oxford University Press.
- Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33*(3), 149–180. <https://doi.org/10.1007/s10919-009-0070-5>
- Hall, K. D., Amir, O., & Yairi, E. (1999). A longitudinal investigation of speaking rate in preschool children who stutter. *Journal of Speech, Language, and Hearing Research, 42*(6), 1367–1377.
<https://doi.org/10.1044/jslhr.4206.1367>
- Harre, R. (1986). *The social construction of emotions*. Blackwell.
- Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication, 154*, 102974.
<https://doi.org/10.1016/j.specom.2023.102974>
- Hess, U., Blairy, S., & Kleck, R. E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior, 21*(4), 241–257. <https://doi.org/10.1023/A:1024952730333>
- Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific Reports, 11*(1), 1–10. <https://doi.org/10.1038/s41598-021-88431-0>
- Holz, N., Poeppel, D., & Hagoort, P. (2023). Vocal expressions differentially transmit emotion categories and affective intensity. *Journal of Experimental Psychology: General, 152*(10), 1–19.
<https://doi.org/10.1037/xge0001383>

- Huang, C. F., Erickson, D., & Akagi, M. (2008). Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners. *Proceedings European Conference on Noise Control*, 2317–2322. <https://doi.org/10.1121/1.2933803>
- Hudlicka, E. (2008). Modeling the mechanisms of emotion effects on cognition. *AAAI Fall Symposium - Technical Report, FS-08-04*, 82–86. <https://aaai.org/papers/0022-fs08-04-022-modeling-the-mechanisms-of-emotion-effects-on-cognition/>
- Izard, C. E. (1977). *Differential emotions theory*. Plenum.
- Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, 99(3), 561–565. <https://doi.org/10.1037//0033-295x.99.3.561>
- Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, 2(4), 363–370. <https://doi.org/10.1177/1754073910374661>
- Jensen, T. W. (2014). Emotion in languaging: Languaging as affective, adaptive, and flexible behavior in social interaction. *Frontiers in Psychology*, 5, 1–14. <https://doi.org/10.3389/fpsyg.2014.00720>
- Jensen, T. W., & Pedersen, S. B. (2016). Affect and affordances - The role of action and emotion in social interaction. *Cognitive Semiotics*, 9(1), 79–103. <https://doi.org/10.1515/cogsem-2016-0003>
- Jiang, X., Paulmann, S., Robin, J., & Pell, M. D. (2015). More than accuracy: Nonverbal dialects modulate the time course of vocal emotion recognition across cultures. *Journal of Experimental Psychology: Human Perception and Performance*, 41(3), 597–612. <https://doi.org/10.1037/xhp0000043>
- Johnson, W. F., Emde, R. N., Scherer, K. R., & Klinnert, M. D. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, 43(3), 280–283. <https://doi.org/10.1001/archpsyc.1986.01800030098011>
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Encyclopedia of Personality and Individual Differences* (Vol. 2, pp. 220–235). https://doi.org/10.1007/978-3-319-28099-8_562-1
- Jun, S.-A. (1995). A phonetic study of stress in Korean. *The Journal of the Acoustical Society of America*, 98(5), 2893. <https://doi.org/10.1121/1.414317>
- Jun, S.-A. (2005). Prosody in sentence processing: Korean vs. English. *UCLA Working Papers in Phonetics*, 104, 26–45. <https://doi.org/http://phonetics.linguistics.ucla.edu/workpapph/104/3-Jun-WPP104>

- Jun, S.-A. (2006). Intonational phonology of Seoul Korean revisited. *Japanese/ Korean Linguistics, 14*, 15–26. <https://escholarship.org/uc/item/90d6532f>
- Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation: Analysis, modelling and technology* (pp. 209–242). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-4317-2_10
- Jun, S.-A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus, 14*(1), 147–172. <https://doi.org/10.1515/prbs.2002.002>
- Jungilligens, J., Paredes-Echeverri, S., Popkirov, S., Barrett, L. F., & Perez, D. L. (2022). A new science of emotion: implications for functional neurological disorder. *Brain, 145*(8), 2648–2663. <https://doi.org/10.1093/brain/awac204>
- Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., & Fischer, J. (2013). Encoding conditions affect recognition of vocally expressed emotions across cultures. *Frontiers in Psychology, 4*, 1–10. <https://doi.org/10.3389/fpsyg.2013.00111>
- Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1797–1813. <https://doi.org/10.1037/0096-1523.26.6.1797>
- Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381–412. <https://doi.org/10.1037/1528-3542.1.4.381>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770–814. <https://doi.org/10.1037/0033-2909.129.5.770>
- Kalawski, J. P. (2010). Is tenderness a basic emotion? *Motivation and Emotion, 34*(2), 158–167. <https://doi.org/10.1007/s11031-010-9164-y>
- Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation, 11*(5), 352–359. <https://doi.org/10.1016/j.jag.2009.06.002>
- Kecman, V. (2005). Support vector machines – An introduction. In L. Wang (Ed.), *Support vector machines: Theory and applications* (pp. 1–47). Springer. https://doi.org/10.1007/10984697_1
- Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior, 43*(2), 133–160. <https://doi.org/10.1007/s10919-019-00293-3>

- Keshtiari, N., & Kuhlmann, M. (2016). The effects of culture and gender on the recognition of emotional speech: Evidence from Persian speakers living in a collectivist society. *International Journal of Society, Culture & Language*, 4(2), 71–86. https://www.ijscsl.com/article_19785.html
- Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and Speech*, 51(4), 343–359. <https://doi.org/10.1177/0023830908099069>
- Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2), 820–857. <https://doi.org/10.1121/1.398894>
- Klecka, W. R. (1980). *Discriminant analysis, quantitative applications in the social sciences*. Sage.
- Kommattam, P., Jonas, K. J., & Fischer, A. H. (2019). Perceived to feel less: Intensity bias in interethnic emotion perception. *Journal of Experimental Social Psychology*, 84, 103809. <https://doi.org/10.1016/j.jesp.2019.04.007>
- Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *Journal of Abnormal and Social Psychology*, 68(4), 390–396. <https://doi.org/10.1037/h0042473>
- Kuppens, P., Realo, A., & Diener, E. (2008). The role of positive and negative emotions in life satisfaction judgment across nations. *Journal of Personality and Social Psychology*, 95(1), 66–75. <https://doi.org/10.1037/0022-3514.95.1.66>
- Kurematsu, M., Amanuma, S., Hakura, J., & Fujita, H. (2010). An extraction of emotion in human speech using cluster analysis and a regression tree. *Proceedings of the 10th WSEAS International Conference on Applied Computer Science* (pp. 346–351). World Scientific and Engineering Academy and Society. <https://dl.acm.org/doi/10.5555/1895260.1895324>
- Lane, R. D., Chua, P. M. L., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, 37(9), 989–997. [https://doi.org/10.1016/S0028-3932\(99\)00017-2](https://doi.org/10.1016/S0028-3932(99)00017-2)
- Lange, K., Kühn, S., & Filevich, E. (2015). “Just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, 10(6), e0130834. <https://doi.org/10.1371/journal.pone.0130834>
- Larsen, R. J., & Diener, E. (1985). A multitrait-multimethod examination of affect structure: Hedonic level and emotional intensity. *Personality and Individual Differences*, 6(5), 631–636. [https://doi.org/10.1016/0191-8869\(85\)90013-3](https://doi.org/10.1016/0191-8869(85)90013-3)

- Larsen, R. J., & Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in Personality*, 21(1), 1–39. [https://doi.org/10.1016/0092-6566\(87\)90023-7](https://doi.org/10.1016/0092-6566(87)90023-7)
- Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Emotion* (pp. 25–59). Sage.
- Larsen, R. J., Diener, E., & Cropanzano, R. S. (1987). Cognitive operations associated with individual differences in affect intensity. *Journal of Personality and Social Psychology*, 53(4), 767–774. <https://doi.org/10.1037/0022-3514.53.4.767>
- Latinus, M., & Taylor, M. J. (2012). Discriminating male and female voices: Differentiating pitch and gender. *Brain Topography*, 25(2), 194–204. <https://doi.org/10.1007/s10548-011-0207-9>
- Laukka, P. (2004). *Vocal expression of emotion: Discrete-emotions and dimensional accounts* (Doctoral dissertation). Uppsala University. <http://www.diva-portal.org/smash/record.jsf?pid=diva2:165425>
- Laukka, P., & Elfenbein, H. A. (2021). Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review*, 13(1), 3–11. <https://doi.org/10.1017/S0272263120000674>
- Laukka, P., Elfenbein, H. A., Chui, W., Thingujam, N. S., Iraki, F. K., Rockstuhl, T., & Althoff, J. (2010). Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect* (pp. 53–57). European Language Resources Association. <https://www.researchgate.net/publication/291326243>
- Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., Iraki, F. K., Rockstuhl, T., & Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology*, 4, Article 353. <https://doi.org/10.3389/fpsyg.2013.00353>
- Laukka, P., Juslin, P. N., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, 19(5), 633–653. <https://doi.org/10.1080/02699930441000445>
- Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech and Language*, 25(1), 84–104. <https://doi.org/10.1016/j.csl.2010.03.004>
- Laukka, P., Thingujam, N. S., Iraki, F. K., Elfenbein, H. A., Rockstuhl, T., Chui, W., & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, 111(5), 686–705. <https://doi.org/10.1037/pspi0000066>

- Lausen, A., & Schacht, A. (2018). Gender differences in the recognition of vocal emotions. *Frontiers in Psychology, 9*, Article 882. <https://doi.org/10.3389/fpsyg.2018.00882>
- Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.
- Lazarus, R. S., & Smith, C. A. (1988). Knowledge and appraisal in the cognition-emotion relationship. *Cognition and Emotion, 2*(4), 281–300. <https://doi.org/10.1080/02699938808412701>
- Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293–303. <https://doi.org/10.1109/TSA.2004.838534>
- Lee, H. Y. (1990). *The structure of Korean prosody* (Doctoral dissertation). University of London. <https://discovery.ucl.ac.uk/1382398/1/395201.pdf>
- Liang, Y., Choi, J., Broersma, M., Goudbeek, M., & Konopka, A. (2023). Rhythmic Similarity hypothesis for cross-cultural vocal emotion recognition. *Proceedings of the 20th International Congress of Phonetic Sciences, Prague*, 1315–1319. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/253.pdf
- Liang, Y., Goudbeek, M., Konopka, A., Choi, J., & Broersma, M. (2025). Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design. *Language and Speech*. <https://doi.org/10.1177/00238309251318730>
- Lim, R. Y., Lew, W. C. L., & Ang, K. K. (2024). Review of EEG affective recognition with a neuroscience perspective. *Brain Sciences, 14*(4). Article 364. <https://doi.org/10.3390/brainsci14040364>
- Liu, C., Ham, J., Postma, E., Midden, C., Joosten, B., & Goudbeek, M. (2012). How to make a robot smile? Perception of emotional expressions from digitally-extracted facial landmark configurations. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M. A. Williams (Eds.), *Social Robotics: Proceedings of the 4th International Conference on Social Robotics (ICSR 2012)* (pp. 26–34). Springer. https://doi.org/10.1007/978-3-642-34103-8_3
- Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE, 13*(5), Article e0196391. <https://doi.org/10.1371/journal.pone.0196391>
- Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America, 49*(2B), 606–608. <https://doi.org/10.1121/1.1912396>

- Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. *Proceedings of Interspeech 2005* (pp. 493–496). <https://doi.org/10.21437/Interspeech.2005-324>
- Majid, A. (2012). Current emotion research in the language sciences. *Emotion Review*, 4(4), 432–443. <https://doi.org/10.1177/1754073912445827>
- Mandal, M. K. (2008). Cultural in-group advantage in accuracy at recognizing vocal expressions of emotion. *Psychological Studies*, 53, 126–132.
- Maskikit-Essed, R., & Gussenhoven, C. (2016). No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology*, 33(2), 353–389. <https://doi.org/10.1017/S0952675716000154>
- Matsumoto, D. (2002). Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence. *Psychological Bulletin*, 128(2), 236–242. <https://doi.org/10.1037/0033-2909.128.2.236>
- Matsumoto, D. (2006). Culture and cultural worldviews: Do verbal descriptions about culture reflect anything other than verbal descriptions of culture? *Culture and Psychology*, 12(1), 33–62. <https://doi.org/10.1177/1354067X06061592>
- Matsumoto, D., & Ekman, P. (1989). American-Japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion*, 13(2), 143–157. <https://doi.org/10.1007/BF00992959>
- Mehrabian, A. (2017). *Nonverbal communication*. Routledge.
- Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin*, 112(2), 179–204. <https://doi.org/10.1037/0033-2909.112.2.179>
- Min, C. S., & Schirmer, A. (2011). Perceiving verbal and vocal emotions in a second language. *Cognition and Emotion*, 25(8), 1376–1392. <https://doi.org/10.1080/02699931.2010.544865>
- Mohanavalli, S., & Jaisakthi, S. M. (2015). A precise distance metric for mixed data clustering using chi-square statistics. *Research Journal of Applied Sciences, Engineering and Technology*, 10(12), 1441–1444. <https://doi.org/10.19026/rjaset.10.1846>
- Morningstar, M., Gilbert, A. C., Burdo, J., Leis, M., & Dirks, M. A. (2021). Recognition of vocal socioemotional expressions at varying levels of emotional intensity. *Emotion*, 21(7), 1570–1575. <https://doi.org/10.1037/emo0001024>
- Mourão-Miranda, J., Volchan, E., Moll, J., De Oliveira-Souza, R., Oliveira, L., Bramati, I., Gattass, R., & Pessoa, L. (2003). Contributions of stimulus valence and arousal to visual activation during emotional perception. *NeuroImage*, 20(4), 1955–1963. <https://doi.org/10.1016/j.neuroimage.2003.08.011>

- Mozziconacci, S. J. L. (2002). Prosody and emotions. *Speech Prosody 2002*, 1–9. <https://doi.org/10.21437/SpeechProsody.2002-1>
- Mozziconacci, S. J. L., & Hermes, D. J. (1997). A study of intonation patterns in speech expressing emotion or attitude : Production and perception. *IPO Annual Progress Report*, 32, 154–160. <https://www.researchgate.net/profile/Sylvie-Mozziconacci/publication/255647322>
- Mozziconacci, S. J. L., & Hermes, D. J. (1999). Role of intonation patterns in conveying emotion in speech. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 2001–2004. <https://pure.tue.nl/ws/files/1354072/696624203788951.pdf>
- Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2), 1097–1108. <https://doi.org/10.1121/1.405558>
- Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12), 1565–1567. <https://doi.org/10.1038/nbt1206-1565>
- Okon-Singer, H., Hendler, T., Pessoa, L., & Shackman, A. J. (2015). The neurobiology of emotion-cognition interactions: Fundamental questions and strategies for future research. *Frontiers in Human Neuroscience*, 9, Article 58. <https://doi.org/10.3389/fnhum.2015.00058>
- Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O’Har, R., Solymos, P., Stevens, M., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Weedon, J. (2024). *Vegan: Community Ecology Package* (R package version 2.6-8).
- Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). F0-contours in emotional speech. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1, 929–932. <https://doi.org/10.21437/ICPhS.1999-232>
- Pallewela, N., Alahakoon, D., Adikari, A., Pierce, J. E., & Rose, M. L. (2024). Optimizing speech emotion recognition with machine learning-based advanced audio cue analysis. *Technologies*, 12(7), 1–17. <https://doi.org/10.3390/technologies12070111>
- Palomero-Gallagher, N., & Amunts, K. (2022). A short review on emotion processing: A lateralized network of neuronal networks. *Brain Structure and Function*, 227(2), 673–684. <https://doi.org/10.1007/s00429-021-02331-7>
- Panksepp, J. (1998). The preconscious substrates of consciousness: Affective states and the evolutionary origins of the self. *Journal of Consciousness Studies*, 5(5–6), 566–582.

- Papez, J. W. (1937). A proposed mechanism of emotions. *Archives of Neurology & Psychiatry*, 38(4), 725–743.
- Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, 87(1), 93–98.
<https://doi.org/10.1016/j.biopsycho.2011.02.010>
- Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion*, 28(2), 230–244.
<https://doi.org/10.1080/02699931.2013.812033>
- Păvăloi, I., & Muscă, E. (2015). Experimental study in emotion recognition using prosodie features. In *2015 E-Health and Bioengineering Conference (EHB)* (pp. 1–4). IEEE.
<https://doi.org/10.1109/EHB.2015.7391422>
- Pell, M. D., & Kotz, S. A. (2021). Comment: The next frontier: Prosody research gets interpersonal. *Emotion Review*, 13(1), 51–56.
<https://doi.org/10.1177/1754073920954288>
- Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33(2), 107–120. <https://doi.org/10.1007/s10919-008-0065-7>
- Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, 37(4), 417–435.
<https://doi.org/10.1016/j.wocn.2009.07.005>
- Pellegrino, F., Coupe, C., & Marisco, E. (2009). A cross-language perspective on speech information rate. *Language*, 87(3), 539–558.
<https://doi.org/10.2307/23011654>
- Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 539–558.
<https://doi.org/10.1353/lan.2011.0057>
- R Core Team. (2018). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. Vienna. <https://www.r-project.org>
- R Core Team. (2022). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. Vienna. <https://www.r-project.org>
- R Core Team. (2023). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. <https://www.r-project.org>
- Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, 67(3), 525–539.
<https://doi.org/10.1037/0022-3514.67.3.525>

- Renna, M. E., Quintero, J. M., Fresco, D. M., & Mennin, D. S. (2017). Emotion regulation therapy: A mechanism-targeted treatment for disorders of distress. *Frontiers in Psychology, 8*, Article 98. <https://doi.org/10.3389/fpsyg.2017.00098>
- Robb, M. P., Maclagan, M. A., & Chen, Y. (2004). Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics, 18*(1), 1–15. <https://doi.org/10.1080/0269920031000105336>
- Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion, 4*(3), 161–190. <https://doi.org/10.1080/02699939008407122>
- Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Russell, J. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102–141. <https://doi.org/10.1037/0033-2909.115.1.102>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition and Emotion, 23*(7), 1259–1283. <https://doi.org/10.1080/02699930902809375>
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805–819. <https://doi.org/10.1037/0022-3514.76.5.805>
- Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*(3), 273–294. [https://doi.org/10.1016/0092-6566\(77\)90037-X](https://doi.org/10.1016/0092-6566(77)90037-X)
- Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology, 63*(11), 2251–2272. <https://doi.org/10.1080/17470211003721642>
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2015). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America, 112*(23), E3086. <https://doi.org/10.1073/pnas.1508604112>

- Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion*, *31*(3), 192–199. <https://doi.org/10.1007/s11031-007-9065-x>
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. <https://doi.org/10.1016/j.neucom.2017.06.053>
- Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165. <https://doi.org/10.1037/0033-2909.99.2.143>
- Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In I. H. W. & A. M. (Eds.), *Handbook of social psychophysiology* (pp. 165–197). John Wiley & Sons.
- Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). Oxford University Press.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. <https://doi.org/10.1080/02699930902928969>
- Scherer, K. R. (2019). Acoustic patterning of emotion vocalizations. In S. Frühholz & P. Belin (Eds.), *Oxford handbook of voice perception* (pp. 61–91). Oxford University Press.
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92. <https://doi.org/10.1177/0022022101032001009>
- Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*(2), 123–148. <https://doi.org/10.1007/BF00995674>
- Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology*, *46*(6), 401–435. <https://doi.org/10.1080/00207594.2011.626049>
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, *1*(4), 331–346. <https://doi.org/10.1007/BF00992539>

- Scherer, K. R., Pell, M. D., Paulmann, S., Dara, C., Allasseri, A., & Kotz, S. A. (2009). Speech prosody across languages: Perceptual and acoustic analyses. *Proceedings of Interspeech 2009* (pp. 1–4).
- Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (pp. 87–90).
<https://doi.org/10.21437/eurospeech.2001-34>
- Shin, J. (2015). Vowels and consonants. In L. Brown & J. Yeon (Eds.), *The Handbook of Korean Linguistics* (pp. 3–21). Wiley.
<https://doi.org/10.1002/9781118371008.ch1>
- Shioiri, T., Someya, T., Helmeste, D., & Tang, S. W. (1999). Cultural difference in recognition of facial emotional expression: Contrast between Japanese and American raters. *Psychiatry and Clinical Neurosciences*, 53(6), 629–633.
<https://doi.org/10.1046/j.1440-1819.1999.00617.x>
- Shiota, M. N., Campos, B., Oveis, C., Hertenstein, M. J., Simon-Thomas, E., & Keltner, D. (2017). Beyond happiness: Building a science of discrete positive emotions. *American Psychologist*, 72(7), 617.
<https://doi.org/https://doi.org/10.1037/a0040456>
- Shiota, M. N., Campos, B., Keltner, D., & Hertenstein, M. J. (2004). Positive emotion and the regulation of interpersonal relationships. In P. Philippot & R. S. Feldman (Eds.), *The regulation of emotion* (pp. 127–155). Erlbaum.
- Shochi, T., Rilliard, A., & Aubergé, V. (2009). Intercultural perception of English, French. In S. Hancil (Ed.), *The Role of Prosody in Affective Speech* (pp. 31–59). Peter Lang.
- Shuman, V., & Scherer, K. R. (2014). Concepts and structures of emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education* (pp. 13–35). Routledge.
- Silva, V., Soares, F., Esteves, J. S., Figueiredo, J., Leão, C. P., Santos, C., & Pereira, A. P. (2016). Real-time emotions recognition system. In *2016 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)* (pp. 201–206). IEEE.
<https://doi.org/10.1109/ICUMT.2016.7765357>
- Sluijter, A. M. C., & van Heuven, V. J. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996)* (Vol. 2, pp. 630–633).
<https://doi.org/10.21437/icslp.1996-159>

- Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. <https://doi.org/10.1037/0022-3514.48.4.813>
- Smith, S. D., McIver, T. A., Di Nella, M. S. J., & Crease, M. L. (2011). The effects of valence and arousal on the emotional modulation of time perception: Evidence for multiple stages of processing. *Emotion*, *11*(6), 1305–1313. <https://doi.org/10.1037/a0026145>
- Sonnemans, J., & Frijda, N. H. (1994). The structure of subjective emotional intensity. *Cognition and Emotion*, *8*(4), 329–350. <https://doi.org/10.1080/02699939408408945>
- Sonnemans, J., & Frijda, N. H. (1995). The determinants of subjective emotional intensity. *Cognition and Emotion*, *9*(5), 483–506. <https://doi.org/10.1080/02699939508408977>
- Spielberger, C. D., Reheiser, E. C., & Sydeman, S. J. (1995). Measuring the experience, expression, and control of anger. *Issues in Comprehensive Pediatric Nursing*, *18*(3), 207–232.
- Stanislavski, K. (1988). *An actor prepares*. Methuen. (Original work published 1936).
- Thompson, W. F., & Balkwill, L. L. (2006). Decoding speech prosody in five languages. *Semiotica*, *158*, 407–424. <https://doi.org/10.1515/SEM.2006.017>
- Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707. <https://doi.org/10.1121/1.397959>
- Tracy, J. L., & Robins, R. W. (2007). Emerging insights into the nature and function of pride. *Current Directions in Psychological Science*, *16*(3), 147–150. <https://doi.org/10.1111/j.1467-8721.2007.00493.x>
- Trampe, D., Quoidbach, J., & Taquet, M. (2015). Emotions in everyday life. *PLoS ONE*, *10*(12), 1–15. <https://doi.org/10.1371/journal.pone.0145450>
- Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100. <https://doi.org/10.1121/1.399849>
- Tsao, Y. C., & Weismer, G. (1997). Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research*, *40*(4), 858–866. <https://doi.org/10.1044/jslhr.4004.858>
- Turnbull, O. H., & Salas, C. E. (2021). The neuropsychology of emotion and emotion regulation: The role of laterality and hierarchy. *Brain Sciences*, *11*(8), 1075. <https://doi.org/10.3390/brainsci11081075>
- Van Bezooijen, R. (1984). *Characteristics and recognizability of vocal expressions of emotion*. Foris. <https://doi.org/10.1515/9783110850390>

- Van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, *14*(4), 387–406. <https://doi.org/10.1177/0022002183014004001>
- Van Heuven, V. J. (2018). Acoustic correlates and perceptual cues of word and sentence stress: Towards a cross-linguistic perspective. In R. Goedemans, J. Heinz, & H. van der Hulst (Eds.), *The study of word stress and accent: Theories, methods and data* (pp. 15–59). Cambridge University Press. <https://doi.org/10.1017/9781316683101.002>
- Van Heuven, V. J., Roosman, L., & van Zanten, E. (2008). Betawi Malay word prosody. *Lingua*, *118*(9), 1271–1287. <https://doi.org/10.1016/j.lingua.2007.09.005>
- Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, *47*(3), 297–308. <https://doi.org/10.1177/00238309040470030401>
- Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181. <https://doi.org/10.1016/j.specom.2006.04.003>
- Vichi, M., Cavicchia, C., & Groenen, P. J. F. (2022). Hierarchical means clustering. *Journal of Classification*, *39*(3), 553–577. <https://doi.org/10.1007/s00357-022-09419-7>
- Vigil, J. M. (2009). A socio-relational framework of sex differences in the expression of emotion. *Behavioral and Brain Sciences*, *32*(5), 375–390. <https://doi.org/10.1017/S0140525X09991075>
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3–28. <https://doi.org/10.1007/BF00987006>
- Wang, H., & van Heuven, V. J. (2018). Relative contribution of vowel quality and duration to native language identification in foreign-accented English. *Proceedings of the 2nd International Conference on Cryptography, Security and Privacy* (pp. 16–20). ACM. <https://doi.org/10.1145/3199478.3199507>
- Widen, S. C., & Russell, J. A. (2010). Descriptive and prescriptive definitions of emotion. *Emotion Review*, *2*(4), 377–378. <https://doi.org/10.1177/1754073910374667>
- Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4B), 1238–1250. <https://doi.org/10.1121/1.1913238>

- Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, 38(10), 1559–1579.
<https://doi.org/10.1016/j.pragma.2005.04.012>
- Wingenbach, T. S. H., Ashwin, C., & Brosnan, M. (2016). Validation of the Amsterdam dynamic facial expression set: Bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions. *PLoS ONE*, 11(1), 1–28.
<https://doi.org/10.1371/journal.pone.0147112>
- Wright, R. A., Brehm, J. W., Wright, R. A., Solomon, S., Silka, L., & Greenberg, J. (1983). Perceived difficulty, energization, and the magnitude of goal valence. *Psychology Library Editions: Social Psychology*, 23(1), 169–210.
[https://doi.org/10.1016/0022-1031\(83\)90003-3](https://doi.org/10.1016/0022-1031(83)90003-3)
- Wu, C., Zhang, J., & Yuan, Z. (2022). An ERP investigation on the second language and emotion perception: The role of emotion word type. *International Journal of Bilingual Education and Bilingualism*, 25(2), 539–551. <https://doi.org/10.1080/13670050.2019.1703895>
- Yrizarry, N., Matsumoto, D., & Wilson-Cohn, C. (1998). American-Japanese differences in multiscale intensity ratings of universal facial expressions of emotion. *Motivation and Emotion*, 22(4), 315–327.
<https://doi.org/10.1023/A:1021304407227>
- Yumot, E., & Gould, W. J. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, 71(6), 1544–1549. <https://doi.org/10.1121/1.387808>
- Zamek, I., & Zamek, S. (2005). Definitions of jitter measurement terms and relationships. *Proceedings of the International Test Conference 2005* (pp. 25–34). <https://doi.org/10.1109/TEST.2005.1583959>
- Zhang, S., & Pell, M. D. (2022). Cultural differences in vocal expression analysis: Effects of task, language, and stimulus-related factors. *PLoS ONE*, 17(10), e0275915. <https://doi.org/10.1371/journal.pone.0275915>
- Zhu, Y. (2013). *Expression and recognition of emotion in native and foreign speech: the case of Mandarin and Dutch*. LOT.
https://www.lotpublications.nl/Documents/341_fulltext.pdf
- Zsidó, A. N. (2023). The effect of emotional arousal on visual attentional performance: a systematic review. *Psychological Research*, 88(1), 1–24.
<https://doi.org/10.1007/s00426-023-01852-6>

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 hòm sepik

on/, 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 \times 8 actors \times 2 tokens \times 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.

Samenvatting

Sinds Darwins "The Expression of the Emotions in Man and Animals" (1872; herdruk in 1998) heeft emotieonderzoek steeds meer aandacht gekregen in verschillende vakgebieden, zoals biologie, psychologie en taalkunde. Darwin stelde dat de productie en perceptie van emoties biologisch bepaald en universeel zijn, terwijl de sociaal-constructivistische theorie stelde dat emoties cultureel en taalkundig geconstrueerd zijn (Harre, 1986). Volgens de dialecttheorie is emotieherkenning fundamenteel universeel, maar kunnen culturele en taalkundige varianten in expressieve stijlen de nauwkeurigheid van de herkenning beïnvloeden (Elfenbein & Ambady, 2002b). Recentelijk groeit de consensus dat cross-culturele emotieherkenning het resultaat is van interacties tussen universele, culturele en taalkundige factoren (Elfenbein, 2013; Elfenbein, Mandal et al., 2002; Mesquita & Frijda, 1992).

Hoewel eerdere studies hebben aangetoond dat vocale emoties over culturen heen herkend kunnen worden, is de vraag nog steeds in hoeverre de productie en perceptie van emoties universeel zijn of beïnvloed worden door cultuur en taal. Dit proefschrift behandelt deze kwestie door te onderzoeken hoe emoties vocaal worden uitgedrukt en waargenomen in twee typologisch en cultureel verschillende talen: Nederlands en Koreaans. Met behulp van het Demo/Koremo-corpus (Broersma et al., 2025) onderzoekt dit proefschrift de herkenning van vocale emoties in verschillende talen, zowel met discrete als dimensionale benaderingen. Er waren drie perceptie-experimenten met luisteraars en één studie gebaseerd op een uitgebreide akoestische analyse van de stimuli, waarbij de volgende aspecten werden onderzocht: 1) de crossculturele en/of crosslinguïstische herkenning van vocale emoties door Nederlandse en Koreaanse luisteraars; 2) de intensiteitsbeoordelingen van vocale emoties door Nederlandse en Koreaanse luisteraars; 3) de relatieve bijdragen van emoties, spreektaal en geslacht aan akoestische parameters; 4) de impact van cultuur en prosodische gelijkenis op de herkenning van vocale emoties door Amerikaans-Engelse en Franse luisteraars.

Hoofdstuk 1 "Inleiding" presenteert een overzicht van de theoretische en empirische studies naar de productie en perceptie van emoties, waarbij de discrete en dimensionale benaderingen worden besproken. Het introduceert een gebalanceerd "twee-bij-twee"-ontwerp, met sprekers en luisteraars uit twee typologisch verschillende talen: Nederlands en Koreaans. Het omvatte vier basisemoties (angst, boosheid, verdriet, vreugde) en vier niet-basisemoties (irritatie, opluchting, tederheid, trots), gebalanceerd in opwinding

(hoge versus lage opwinding) en valentie (positief/aangenaam versus negatief/onaangenaam). Om semantische aanwijzingen te vermijden, werden alle emoties geproduceerd op een pseudozin /nuto hɔm sepikɑŋ/, die uitspreekbaar is in het Nederlands en het Koreaans, maar in beide talen geen betekenis heeft. De acht emoties werden uitgedrukt door acht Nederlandse en acht Koreaanse stemacteurs, met vier vrouwen en vier mannen per taalgroep. Dit stelde ons in staat om gendergerelateerde verschillen in prosodische expressie van emoties te bestuderen (Klatt & Klatt, 1990). Elke acteur produceerde twee keer dezelfde emoties, wat resulteerde in een totaal van 256 vertolkingen (8 emoties × 8 acteurs × 2 tokens × 2 talen). Ten slotte schetst dit hoofdstuk de onderzoeksvragen die in elk van de volgende hoofdstukken worden behandeld.

Hoofdstuk 2 “Onderzoek naar cross-culturele vocale emotieherkenning met een affectief en taalkundig gebalanceerd ontwerp”, onderzoekt de herkenning van vocale Nederlandse en Koreaanse emoties door Nederlandse en Koreaanse luisteraars. Dit hoofdstuk onderzocht 1) of er een in-group voordeel is bij cross-culturele emotieherkenning; 2) of er een verschil is in herkenningsnauwkeurigheid tussen emoties met hoge en lage opwinding; 3) of er een verschil is in herkenningsnauwkeurigheid tussen positieve en negatieve emoties; 4) of er een verschil is in herkenningsnauwkeurigheid tussen basis- en niet-basisemoties.

De resultaten lieten zien dat beide luisteraarsgroepen vocale emoties boven kans herkenden, zelfs in de onbekende taal. Bovendien vertoonden beide luisteraarsgroepen een in-group voordeel bij het herkennen van vocale emoties, zodat luisteraars vocale emoties die in hun moedertaal worden geproduceerd nauwkeuriger herkenden dan die welke in de onbekende taal werden uitgedrukt. Deze bevindingen ondersteunen het idee dat cross-culturele emotieherkenning het resultaat is van een interactie tussen universele en cultuur-/taalspecifieke factoren (Elfenbein, 2013; Elfenbein & Ambady, 2002b). Bovendien onderzocht deze studie de dimensionale effecten van opwinding, valentie en basiciteit op de herkenning van vocale emoties, binnen en tussen culturen. De herkenningsnauwkeurigheid bleek hoger voor lage opwinding, negatieve en basisemoties dan voor hoge opwinding, positieve en niet-basisemoties, zowel binnen als tussen culturen.

Hoofdstuk 3 “Interpretatie van de intensiteit van vocale emoties tussen culturen” onderzoekt de intensiteitsbeoordelingen van Nederlandse en Koreaanse luisteraars verzameld in Studie 1. Intensiteit is de sterkte van een emotie die door de ontvanger wordt waargenomen (Bänziger & Scherer, 2005; Diener et al., 1985; Larsen & Diener, 1987; Sonnemans & Frijda, 1994). Mensen reageren over het algemeen sterker op emoties met een hogere

intensiteit dan op emoties met een lagere intensiteit. Bovendien geven individuen doorgaans hogere intensiteitsbeoordelingen aan emoties die worden geuit door leden van dezelfde of vergelijkbare cultuur/taalgroep dan door leden van een typologisch andere groep, wat bekend staat als de in-group intensiteitsbias (Kommattam et al., 2019). Dit hoofdstuk onderzocht 1) of accurate trials hogere intensiteitsbeoordelingen krijgen dan onnauwkeurige; 2) of er een in-group bias is voor intensiteitsbeoordelingen tussen culturen; 3) of de drie (dimensionale) binaire splitsingen in opwinding, valentie en basiciteit de intensiteitsbeoordelingen betrouwbaar kunnen voorspellen.

De bevindingen wezen uit dat accurate trials van beide groepen luisteraars hogere intensiteitsbeoordelingen kregen dan inaccurate trials. Boosheid kreeg met name de hoogste intensiteitsbeoordelingen van beide luisteraarsgroepen, zowel in accurate als inaccurate trials. Er werd echter geen in-group bias gevonden in de intensiteitsbeoordelingen. Bovendien waren de intensiteitsbeoordelingen nauw gerelateerd aan opwinding, valentie en basiciteit, zodat de intensiteitsbeoordelingen hoger waren voor emoties met hoge opwinding, negatieve en basisemoties dan voor emoties met lage opwinding, positieve en niet-basisemoties.

Hoofdstuk 4 “Emoties classificeren op basis van akoestische parameters”, analyseerde 17 akoestische parameters, die werden ingedeeld in vijf categorieën gerelateerd aan: 1) toonhoogte, 2) amplitude, 3) spectrale samenstelling, 4) duur en 5) perturbatie (stemonvastheid). Deze studie onderzocht 1) hoe de akoestische patronen verschillen tussen emoties, sprekerstaal en geslacht; 2) hoe nauwkeurig de herkenning van emoties kan worden voorspeld op basis van akoestische parameters, en in hoeverre de classificatie verbetert wanneer de materialen per taal worden gesplitst; 3) in hoeverre herkenning van emoties d.m.v. machine learning (Support Vector Machine, SVM) overeenkomt met de prestaties van menselijke luisteraars.

De resultaten lieten zien dat elke emotie een uniek akoestisch profiel vertoonde, dat varieerde per taal en geslacht. Vocale emoties kunnen betrouwbaar worden onderscheiden met een SVM-classificator door middel van geoptimaliseerde combinaties van akoestische signalen. De classificatienauwkeurigheid verbeterde wanneer de Nederlandse en Koreaanse gegevens afzonderlijk werden geanalyseerd. De verwarringsmatrices suggereren dat er relevante kwalitatieve verschillen zijn tussen de prestaties van SVM en die van menselijke luisteraars, die nader onderzocht moeten worden.

Hoofdstuk 5 “Het herkennen van vocale emoties in onbekende talen”, breidt het onderzoek uit naar cross-linguïstische herkenning door Amerikaans-

Engelse en Franse luisteraars die niet bekend waren met het Nederlands en Koreaans. Deze studie onderzoekt de relatieve invloed van de universaliteits-hypothese, culturele nabijheid, linguïstische nabijheid en emotionele dimensies (opwinding, valentie en basiciteit) op emotieherkenning. Bij het herkennen van vocale emoties herkennen luisteraars emoties gemakkelijker in talen die vergelijkbaar zijn met hun eigen taal. Hoewel emotionele dimensies zoals opwinding, valentie en basiciteit de herkenning beïnvloeden, is hun specifieke impact op de nauwkeurigheid niet volledig begrepen. Deze studie onderzoekt hoe universele, culturele, linguïstische en emotionele factoren de perceptie van vocale emoties beïnvloeden. We selecteerden Amerikaans-Engelse en Franse luisteraars omdat Engels, een taal met klemtoontiming, prosodisch vergelijkbaar is met het Nederlands, terwijl Frans, een taal met lettergreeptiming, vergelijkbaar is met het Koreaans. Door de herkenning-nauwkeurigheid van deze groepen te vergelijken, willen we de invloed van deze factoren verduidelijken. De studie beoordeelde eerst of beide groepen vocale emoties boven verwachting herkenden. Vervolgens vergeleken we hun nauwkeurigheid met Nederlandse opnames. Daarna evalueerden we of Franse luisteraars beter presteerden dan Amerikaans-Engelse luisteraars met Koreaanse opnames, gezien de prosodische overeenkomsten. Tot slot analyseerden we de effecten van opwinding, valentie en basiciteit op de herkenning van vocale emoties.

De resultaten tonen aan dat zowel Amerikaanse als Franse luisteraarsgroepen vocale emoties boven kans herkenden, wat de universaliteitshypothese steunt (Elfenbein, 2013; Elfenbein & Ambady, 2002a). Bovendien geven de resultaten aan dat mensen emoties nauwkeuriger herkennen wanneer ze hun culturele en taalkundige achtergrond delen met de spreker, een bevinding die in lijn is met Elfenbein en Ambady (2003a).

Hoofdstuk 6 “Conclusies en discussie” herhaalt de onderzoeksvragen van elk hoofdstuk en vat de belangrijkste bevindingen van elke studie samen met een discussie en een integratie van de resultaten. Dit hoofdstuk behandelt ook de beperkingen van de empirische studies. Tot slot worden de volgende vijf nieuwe bevindingen en bijdragen aan de bestaande literatuur over emotieonderzoek genoemd.

Ten eerste presenteert onze studie een evenwichtig “twee-bij-twee” gekruist experimenteel design, met sprekers en luisteraars uit twee typologisch verschillende talen en culturen. Dit design maakt een systematische vergelijking mogelijk van herkenningsspatronen binnen en tussen culturen, met name om het in-group voordeel in emotieherkenning te onderzoeken. We repliceerden de eerdere bevinding dat luisteraars vocale emoties boven kans

herkenden, zelfs in een onbekende taal, wat consistent is met de universaliteitshypothese. Bovendien ontdekten we dat beide luisteraarsgroepen een in-group voordeel vertoonden, waardoor ze vocale emoties nauwkeuriger herkenden wanneer ze in hun moedertaal werden uitgesproken dan wanneer ze in de onbekende taal werden uitgesproken.

Ten tweede hebben we zowel discrete als dimensionale benaderingen gebruikt, met vier basis- en vier niet-basisemoties, gebalanceerd naar opwinding en valentie, wat een uitgebreider begrip opleverde van de mechanismen die ten grondslag liggen aan emotieherkenning. Deze studie onderzocht niet alleen de herkenningsnauwkeurigheid van elk van de acht emoties categorisch, maar ook de effecten van drie emotionele dimensies (opwinding, valentie en basisgevoel) op emotieherkenning, binnen en tussen culturen. De herkenningsnauwkeurigheid bleek hoger voor lage-opwinding, negatieve en basisemoties dan voor hoge-opwinding, positieve en niet-basisemoties, zowel binnen als tussen culturen. Voor zover wij weten is dit de eerste studie die de herkenningsnauwkeurigheid van lage-opwinding en hoge-opwinding emoties binnen en tussen culturen direct heeft vergeleken. Deze bevinding verklaart de rol van opwinding bij emotieherkenning.

Ten derde onderzochten we intensiteitsbeoordelingen in verschillende culturen. Intensiteit, als belangrijke dimensie van emoties, heeft relatief weinig aandacht gekregen in eerder onderzoek. We vonden geen in-group bias in intensiteitsbeoordelingen. De resultaten lieten wél zien dat intensiteitsbeoordelingen voornamelijk werden bepaald door het type emotie in plaats van door culturele of taalkundige factoren. Met behulp van de dimensionale benadering onderzochten we de rol van opwinding, valentie en basiciteit in intensiteitsbeoordelingen. We ontdekten dat intensiteitsbeoordelingen over het algemeen hoger waren voor emoties met hoge opwinding, negatieve emoties en basisemoties dan voor emoties met lage opwinding, positieve emoties en niet-basisemoties.

Ten vierde biedt onze studie een uitgebreide analyse van 17 akoestische parameters in twee typologisch verschillende talen, waaronder een ondervertegenwoordigde niet-Indo-Europese taal: Koreaans. We ontdekten dat verschillende vocale emoties universele en emotiespecifieke akoestische patronen vertonen. Met name toonhoogte, intensiteit en spreeknelheid kunnen vocale emoties succesvol differentiëren. We ontdekten ook dat laryngale parameters genuanceerde onderscheidingen tussen vocale emoties mogelijk maken. Bovendien vertonen vocale emoties taalafhankelijke akoestische patronen. De articulatiesnelheid was opmerkelijk hoger in het Koreaans dan in het Nederlands, en er waren geen spraakpauzes in het

Koreaans. Bovendien gebruikten vrouwen en mannen verschillende akoestische eigenschappen, onder andere doordat (relatieve) toonhoogte- en amplitudegerelateerde akoestische parameters over het algemeen hoger zijn voor vrouwen dan voor mannelijke acteurs. Op basis van de akoestische parameters onderzochten we of vocale emoties nauwkeurig kunnen worden geclassificeerd door middel van machine learning (SVM-modellen) en vergeleken we de prestaties van machinale classificatoren voor Nederlands en Koreaans in zowel in-group- als out-group-omstandigheden. De resultaten wezen uit dat classificatienauwkeurigheid betrouwbaar kan worden voorspeld uit combinaties van akoestische parameters. Opvallend is dat machinale classificatie, net als menselijke luisteraars, beter presteerde in de in-group-conditie dan in de out-group-conditie voor zowel Nederlanders als Koreanen, hoewel de classificatiepercentages in beide condities boven kans lagen.

Ten slotte onderzochten we cross-culturele/linguïstische herkenning van vocale emoties door Amerikaans-Engelse en Franse luisteraars, die geen kennis hadden van Nederlands of Koreaans. Deze studie breidt voor de bestaande literatuur uit over cross-culturele/linguïstische herkenning van emoties door de bijdragen van universaliteit, culturele nabijheid, prosodische nabijheid en emotionele dimensies (opwinding, valentie en basiciteit) aan cross-culturele herkenning van vocale emoties te testen. In overeenstemming met de Universaliteitshypothese identificeerden beide luisteraarsgroepen vocale emoties over het algemeen boven kans, zelfs in een onbekende taal. We vonden echter geen algemeen effect van culturele afstand op emotieherkenning, aangezien de herkenningsnauwkeurigheid varieerde per emotie, noch vonden we ondersteuning van het effect van prosodische afstand op vocale emotieherkenning over culturen/talen heen, aangezien de Franse luisteraars het met de Koreaanse opnames niet beter deden dan de Amerikaans-Engelse luisteraars. Belangrijker nog, we ontdekten dat hoewel opwinding, valentie en basiciteit de herkenning van vocale emoties beïnvloeden, sommige emoties deze algemene patronen schonden. Hoewel emotionele dimensies een bruikbaar raamwerk bieden om de nauwkeurigheid van herkenning te begrijpen, verklaren zij de variatie in emotieherkenning tussen culturen en talen vooralsnog onvoldoende.

Appendices

Appendix A. Summary of results of one-sample *t*-test analyses for Hypothesis 1.

| | Mean | <i>t</i> | <i>df</i> | 95% Confidence Interval | |
|------------------|------|----------|-----------|-------------------------|-------|
| | | | | Lower | Upper |
| Dutch listeners | | | | | |
| Dutch speakers | 0.47 | 32.69*** | 30 | 0.45 | 0.49 |
| Korean speakers | 0.38 | 34.77*** | 30 | 0.36 | 0.40 |
| Korean listeners | | | | | |
| Dutch speakers | 0.36 | 22.80*** | 23 | 0.34 | 0.38 |
| Korean speakers | 0.43 | 23.70*** | 23 | 0.40 | 0.46 |

“****” $p < .001$, “***” $p < .01$, “**” $p < .05$, “.” $.05 < p < .10$
 All tests used the Bonferroni-corrected p value of .0125.

Appendix B. Summary of results of one-sample *t*-test analyses for Hypothesis 5.

| | Mean | <i>t</i> | <i>df</i> | 95% CI | |
|---------------------|------|----------|-----------|--------|-------|
| | | | | Lower | Upper |
| Basic emotions | | | | | |
| Dutch listeners | | | | | |
| Dutch speakers | 0.56 | 24.55*** | 30 | 0.52 | 0.60 |
| Korean speakers | 0.47 | 30.81*** | 30 | 0.45 | 0.49 |
| Non-basic emotions | | | | | |
| Dutch listeners | | | | | |
| Dutch speakers | 0.38 | 23.29*** | 30 | 0.36 | 0.40 |
| Korean speakers | 0.29 | 15.40*** | 30 | 0.27 | 0.31 |
| Basic emotions | | | | | |
| Korean listeners | | | | | |
| Dutch speakers | 0.49 | 23.04*** | 23 | 0.46 | 0.52 |
| Korean speakers | 0.46 | 18.08*** | 23 | 0.42 | 0.50 |
| Non-basic emotions: | | | | | |
| Korean listeners | | | | | |
| Dutch speakers | 0.23 | 8.45*** | 23 | 0.20 | 0.26 |
| Korean speakers | 0.39 | 20.63*** | 23 | 0.36 | 0.42 |

“****” $p < .001$, “***” $p < .01$, “**” $p < .05$, “.” $.05 < p < .10$

Statistical significance was established against the Bonferroni-corrected p value of .00625.

Appendix C. Summary of LME models 1 through 10. Significant p -values ($p \leq .05$) in boldface.

| Model 1: Linear mixed effects model for Dutch listeners, all responses | | | | |
|---|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.09 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.09 | | | |
| Residual | 0.92 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.82 | 0.10 | 28.98 | < .001 |
| Speaker Language | -0.64 | 0.16 | -3.91 | < .001 |
| Emotion (Joy) | -0.20 | 0.04 | -4.64 | < .001 |
| Emotion (Pride) | -0.89 | 0.04 | -20.61 | < .001 |
| Emotion (Fear) | -0.27 | 0.04 | -6.28 | < .001 |
| Emotion (Tenderness) | -1.23 | 0.04 | -28.54 | < .001 |
| Emotion (Relief) | -0.96 | 0.04 | -22.27 | < .001 |
| Emotion (Sadness) | -0.69 | 0.04 | -15.99 | < .001 |
| Emotion (Irritation) | -1.07 | 0.04 | -24.77 | < .001 |
| Speaker Language \times Emotion (Joy) | 0.10 | 0.09 | 1.22 | .223 |
| Speaker Language \times Emotion (Pride) | 0.85 | 0.09 | 9.84 | < .001 |
| Speaker Language \times Emotion (Fear) | 0.50 | 0.09 | 5.76 | < .001 |
| Speaker Language \times Emotion (Tenderness) | 0.47 | 0.09 | 5.46 | < .001 |
| Speaker Language \times Emotion (Relief) | 0.53 | 0.09 | 6.21 | < .001 |
| Speaker Language \times Emotion (Sadness) | 1.00 | 0.09 | 11.64 | < .001 |
| Speaker Language \times Emotion (Irritation) | 0.52 | 0.09 | 6.04 | < .001 |

| Model 2: Linear mixed effects model for Korean listeners, all responses | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.15 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.09 | | | |
| Residual | 0.93 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 3.14 | 0.11 | 27.42 | < .001 |
| Speaker Language | -0.15 | 0.17 | -0.90 | .380 |
| Emotion (Joy) | -0.40 | 0.05 | -8.21 | < .001 |
| Emotion (Pride) | -0.99 | 0.05 | -20.16 | < .001 |
| Emotion (Fear) | -0.48 | 0.05 | -9.82 | < .001 |
| Emotion (Tenderness) | -1.23 | 0.05 | -25.02 | < .001 |
| Emotion (Relief) | -1.24 | 0.05 | -25.04 | < .001 |
| Emotion (Sadness) | -0.87 | 0.05 | -17.66 | < .001 |
| Emotion (Irritation) | -1.34 | 0.05 | -27.21 | < .001 |
| Speaker Language \times Emotion (Joy) | -0.38 | 0.10 | -3.88 | < .001 |
| Speaker Language \times Emotion (Pride) | 0.55 | 0.10 | 5.54 | < .001 |
| Speaker Language \times Emotion (Fear) | 0.01 | 0.10 | 0.11 | .916 |
| Speaker Language \times Emotion (Tenderness) | 0.37 | 0.10 | 3.80 | < .001 |
| Speaker Language \times Emotion (Relief) | 0.11 | 0.10 | 1.08 | .279 |
| Speaker Language \times Emotion (Sadness) | 0.45 | 0.10 | 4.57 | < .001 |
| Speaker Language \times Emotion (Irritation) | 0.40 | 0.10 | 4.04 | < .001 |

| Model 3: LME model analyses for Dutch listeners, correct responses | | | | |
|---|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.10 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.08 | | | |
| Residual | 0.67 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 3.19 | 0.10 | 31.84 | < .001 |
| Speaker Language | -0.48 | 0.17 | -2.84 | < .010 |
| Emotion (Joy) | -0.32 | 0.06 | -5.09 | < .001 |
| Emotion (Pride) | -0.88 | 0.09 | -10.21 | < .001 |
| Emotion (Fear) | -0.44 | 0.05 | -8.05 | < .001 |
| Emotion (Tenderness) | -0.96 | 0.07 | -14.60 | < .001 |
| Emotion (Relief) | -0.97 | 0.06 | -17.50 | < .001 |
| Emotion (Sadness) | -0.83 | 0.05 | -17.01 | < .001 |
| Emotion (Irritation) | -1.12 | 0.05 | -20.85 | < .001 |
| Speaker Language \times Emotion (Joy) | -0.09 | 0.13 | -0.70 | .482 |
| Speaker Language \times Emotion (Pride) | 0.74 | 0.17 | 4.27 | < .001 |
| Speaker Language \times Emotion (Fear) | 0.40 | 0.11 | 3.69 | < .001 |
| Speaker Language \times Emotion (Tenderness) | 0.50 | 0.13 | 3.82 | < .001 |
| Speaker Language \times Emotion (Relief) | 0.34 | 0.11 | 3.08 | < .010 |
| Speaker Language \times Emotion (Sadness) | 0.76 | 0.10 | 7.84 | < .001 |
| Speaker Language \times Emotion (Irritation) | 0.50 | 0.11 | 4.65 | < .001 |

| Model 4: LME model analyses for Korean listeners, correct responses | | | | |
|--|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.13 | | | |
| Speaker Language | 0.08 | | | |
| Speaker (Intercept) | 0.06 | | | |
| Residual | 0.69 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 3.28 | 0.10 | 31.36 | < .001 |
| Speaker Language | -0.26 | 0.16 | -1.64 | .109 |
| Emotion (Joy) | -0.32 | 0.08 | -3.91 | < .001 |
| Emotion (Pride) | -0.63 | 0.08 | -7.79 | < .001 |
| Emotion (Fear) | -0.55 | 0.07 | -7.86 | < .001 |
| Emotion (Tenderness) | -0.64 | 0.08 | -7.65 | < .001 |
| Emotion (Relief) | -1.16 | 0.07 | -17.07 | < .001 |
| Emotion (Sadness) | -0.84 | 0.06 | -14.95 | < .001 |
| Emotion (Irritation) | -1.25 | 0.07 | -17.68 | < .001 |
| Speaker Language \times Emotion (Joy) | -0.43 | 0.16 | -2.61 | < .010 |
| Speaker Language \times Emotion (Pride) | 0.90 | 0.16 | 5.62 | < .001 |
| Speaker Language \times Emotion (Fear) | 0.21 | 0.14 | 1.50 | .135 |
| Speaker Language \times Emotion (Tenderness) | 0.49 | 0.17 | 2.94 | < .010 |
| Speaker Language \times Emotion (Relief) | 0.32 | 0.14 | 2.39 | < .050 |
| Speaker Language \times Emotion (Sadness) | 0.61 | 0.11 | 5.36 | < .001 |
| Speaker Language \times Emotion (Irritation) | 0.55 | 0.14 | 3.94 | < .001 |

| Model 5: Linear mixed effects model analyses for Arousal, all responses | | | | |
|--|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.10 | | | |
| Arousal | 0.04 | | | |
| Listener (Intercept) | 0.02 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.04 | | | |
| Arousal | 0.07 | | | |
| Speaker (Intercept) | 0.06 | | | |
| Listener Language | 0.00 | | | |
| Residual | 0.99 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.24 | 0.09 | 23.89 | < .001 |
| Speaker Language | -0.05 | 0.16 | -0.34 | .743 |
| Listener Language | 0.17 | 0.10 | 1.72 | .091 |
| Arousal | -0.67 | 0.08 | -8.95 | < .001 |
| Speaker Language \times Listener Language | 0.18 | 0.06 | 2.76 | < .010 |
| Speaker Language \times Arousal | 0.28 | 0.14 | 2.01 | .064 |
| Listener Language \times Arousal | -0.05 | 0.07 | -0.80 | .425 |
| Speaker Lang. \times Listener Lang. \times Arousal | 0.02 | 0.07 | 0.30 | .766 |
| Model 6: Linear mixed effects model analyses for Arousal, correct responses | | | | |
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.03 | | | |
| Arousal | 0.04 | | | |
| Listener (Intercept) | 0.07 | | | |
| Speaker Language | 0.05 | | | |
| Speaker (Intercept) | 0.01 | | | |
| Arousal | 0.09 | | | |
| Speaker (Intercept) | 0.05 | | | |
| Listener Language | 0.01 | | | |
| Residual | 0.71 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.60 | 0.08 | 32.86 | < .001 |
| Speaker Language | -0.05 | 0.14 | -0.39 | .700 |
| Listener Language | 0.09 | 0.10 | 0.98 | .330 |
| Arousal | -0.68 | 0.08 | -8.36 | < .001 |
| Speaker Language \times Listener Language | 0.10 | 0.10 | 1.07 | .293 |
| Speaker Language \times Arousal | 0.44 | 0.15 | 2.85 | < .050 |
| Listener Language \times Arousal | -0.01 | 0.07 | -0.16 | .872 |
| Speaker Lang. \times Listener Lang. \times Arousal | 0.05 | 0.09 | 0.53 | .595 |

| Model 7: Linear mixed effects model analyses for Valence, all responses | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.04 | | | |
| Listener (Intercept) | 0.08 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.01 | | | |
| Valence | 0.08 | | | |
| Speaker (Intercept) | 0.08 | | | |
| Listener Language | 0.00 | | | |
| Residual | 0.10 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.24 | 0.09 | 25.00 | < .001 |
| Speaker Language | -0.05 | 0.16 | -0.35 | .729 |
| Listener Language | 0.17 | 0.09 | 1.78 | .080 |
| Valence | -0.30 | 0.07 | -4.27 | < .001 |
| Speaker Language \times Listener Language | 0.18 | 0.06 | 2.80 | < .010 |
| Speaker Language \times Valence | -0.03 | 0.14 | -0.24 | .811 |
| Listener Language \times Valence | 0.02 | 0.04 | 0.54 | .588 |
| Speaker Lang. \times Listener Lang. \times Valence | -0.04 | 0.07 | -0.53 | .596 |
| Model 8: Linear mixed effects model analyses for Valence, correct responses | | | | |
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.05 | | | |
| Valence | 0.03 | | | |
| Listener (Intercept) | 0.06 | | | |
| Speaker Language | 0.06 | | | |
| Speaker (Intercept) | 0.00 | | | |
| Valence | 0.07 | | | |
| Speaker (Intercept) | 0.06 | | | |
| Listener Language | 0.01 | | | |
| Residual | 0.84 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.52 | 0.08 | 32.19 | < .001 |
| Speaker Language | -0.06 | 0.13 | -0.43 | .674 |
| Listener Language | 0.09 | 0.10 | 0.95 | .348 |
| Valence | -0.17 | 0.08 | -2.23 | < .050 |
| Speaker Language \times Listener Language | 0.13 | 0.10 | 1.33 | .192 |
| Speaker Language \times Valence | 0.02 | 0.14 | 0.14 | .892 |
| Listener Language \times Valence | 0.07 | 0.07 | 1.06 | .294 |
| Speaker Lang. \times Listener Lang. \times Valence | 0.33 | 0.11 | 3.07 | < .010 |

| Model 9: Linear mixed effects model analyses for Basicness, all responses | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.09 | | | |
| Basicness | 0.04 | | | |
| Listener (Intercept) | 0.03 | | | |
| Speaker Language | 0.04 | | | |
| Speaker (Intercept) | 0.05 | | | |
| Basicness | 0.11 | | | |
| Speaker (Intercept) | 0.04 | | | |
| Listener Language | 0.00 | | | |
| Residual | 0.96 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.24 | 0.09 | 25.12 | < .001 |
| Speaker Language | -0.05 | 0.15 | -0.36 | .727 |
| Listener Language | 0.17 | 0.10 | 1.77 | .082 |
| Basicness | -0.75 | 0.09 | -8.68 | < .001 |
| Speaker Language \times Listener Language | 0.18 | 0.07 | 2.74 | < .010 |
| Speaker Language \times Basicness | 0.26 | 0.17 | 1.60 | .132 |
| Listener Language \times Basicness | -0.02 | 0.06 | -0.26 | .789 |
| Speaker Lang. \times Listener Lang. \times Basicness | 0.15 | 0.07 | 2.18 | < .050 |
| Model 10: LME model analyses for Basicness, correct responses | | | | |
| Random effects | <i>Variance</i> | | | |
| Listener (Intercept) | 0.01 | | | |
| Basicness | 0.02 | | | |
| Listener (Intercept) | 0.10 | | | |
| Speaker Language | 0.05 | | | |
| Speaker (Intercept) | 0.02 | | | |
| Basicness | 0.15 | | | |
| Speaker (Intercept) | 0.04 | | | |
| Listener Language | 0.01 | | | |
| Residual | 0.75 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 2.48 | 0.08 | 31.22 | < .001 |
| Speaker Language | -0.03 | 0.13 | -0.22 | .826 |
| Listener Language | 0.08 | 0.10 | 0.88 | .385 |
| Basicness | -0.54 | 0.10 | -5.36 | < .001 |
| Speaker Language \times Listener Language | 0.19 | 0.09 | 2.09 | < .050 |
| Speaker Language \times Basicness | 0.22 | 0.20 | 1.09 | .295 |
| Listener Language \times Basicness | 0.11 | 0.06 | 1.80 | .078 |
| Speaker Lang. \times Listener Lang. \times Basicness | 0.23 | 0.10 | 2.34 | < .050 |

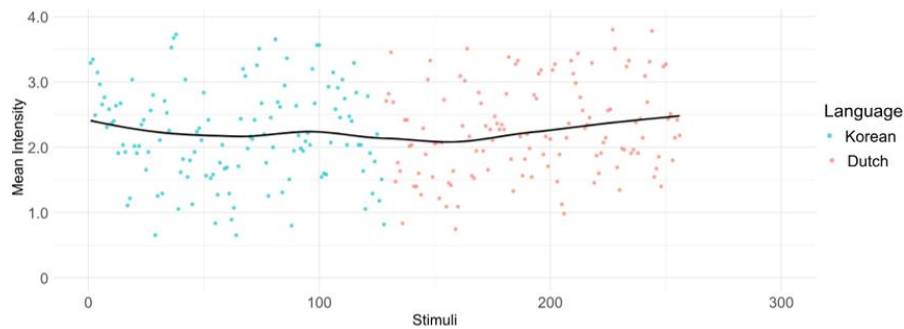
Appendix D. Intercepts for Arousal, Valence, and Basicness for eight emotions, for all responses and for correctly identified emotions only.

| Emotion | Arousal | | Valence | | Basicness | |
|------------|---------|---------|---------|---------|-----------|---------|
| | all | correct | all | correct | all | correct |
| Anger | 0.39 | 0.39 | 0.58 | 0.66 | 0.35 | 0.45 |
| Joy | 0.11 | 0.09 | 0.59 | 0.46 | 0.07 | 0.11 |
| Fear | 0.03 | -0.10 | 0.22 | 0.11 | -0.01 | -0.10 |
| Sadness | 0.30 | 0.13 | -0.19 | -0.24 | -0.41 | -0.46 |
| Pride | -0.53 | -0.38 | -0.05 | -0.03 | 0.17 | 0.18 |
| Relief | -0.01 | -0.08 | -0.20 | -0.33 | 0.03 | -0.07 |
| Irritation | -0.12 | -0.19 | -0.61 | -0.54 | -0.08 | -0.23 |
| Tenderness | -0.16 | 0.15 | -0.35 | -0.11 | -0.12 | 0.12 |

Appendix E. Recognition accuracy (Acc., %) and intensity ratings (Int., 0-4) by Dutch and Korean listeners, for Dutch and Korean speakers.

| Emotion | Dutch listeners | | | | Korean listeners | | | |
|------------|-----------------|------|---------------|------|------------------|------|---------------|------|
| | Dutch speakers | | Kor. speakers | | Dutch speakers | | Kor. speakers | |
| | Acc. | Int. | Acc. | Int. | Acc. | Int. | Acc. | Int. |
| Anger | 62.28 | 3.14 | 37.50 | 2.50 | 64.95 | 3.22 | 33.42 | 3.07 |
| Joy | 42.46 | 2.89 | 22.41 | 2.35 | 20.92 | 3.01 | 21.20 | 2.47 |
| Pride | 20.91 | 1.83 | 8.19 | 2.03 | 20.11 | 1.95 | 24.73 | 2.35 |
| Fear | 40.95 | 2.62 | 53.88 | 2.47 | 26.09 | 2.73 | 51.09 | 2.59 |
| Tenderness | 29.74 | 1.67 | 23.71 | 1.50 | 14.40 | 1.80 | 34.24 | 2.02 |
| Relief | 48.28 | 1.91 | 41.59 | 1.80 | 36.14 | 1.93 | 39.40 | 1.89 |
| Sadness | 80.82 | 1.95 | 73.49 | 2.31 | 82.61 | 2.12 | 81.25 | 2.42 |
| Irritation | 53.45 | 1.81 | 44.40 | 1.69 | 64.95 | 3.22 | 33.42 | 3.07 |

Appendix F. Mean intensity ratings for all data, with a smooth LOESS line to show the pattern over time (from stimulus 1 to stimulus 256).



Appendix G. Summary of LME models 1 through 15 for Hypotheses 1–3 in all data, Dutch data, and Korean data. Significant p -values ($p \leq .05$) in boldface.

| All data | F0 M | | | |
|--|----------|------|--------|--------|
| Random effects | | | | |
| Speaker (Intercept) | Variance | | | |
| Residual | 2.71 | | | |
| Fixed effects | | | | |
| | β | SE | t | p |
| Intercept | 30.04 | 0.69 | 43.64 | < .001 |
| Gender | -7.93 | 0.91 | -8.71 | < .001 |
| Speaker Language | -5.67 | 1.38 | -4.12 | < .001 |
| Emotion (Joy) | -1.34 | 0.78 | -1.72 | 0.087 |
| Emotion (Pride) | -4.50 | 0.78 | -5.76 | < .001 |
| Emotion (Anger) | -0.94 | 0.78 | -1.21 | 0.228 |
| Emotion (Tenderness) | -9.83 | 0.78 | -12.60 | < .001 |
| Emotion (Relief) | -7.59 | 0.78 | -9.73 | < .001 |
| Emotion (Sadness) | -7.42 | 0.78 | -9.51 | < .001 |
| Emotion (Irritation) | -6.04 | 0.78 | -7.74 | < .001 |
| Speaker Language \times Emotion (Joy) | 0.28 | 1.56 | 0.18 | 0.859 |
| Speaker Language \times Emotion (Pride) | 6.49 | 1.56 | 4.16 | < .001 |
| Speaker Language \times Emotion (Anger) | 3.81 | 1.56 | 2.44 | < .050 |
| Speaker Language \times Emotion (Tenderness) | 6.50 | 1.56 | 4.17 | < .001 |
| Speaker Language \times Emotion (Relief) | 5.66 | 1.56 | 3.63 | < .001 |
| Speaker Language \times Emotion (Sadness) | 6.38 | 1.56 | 4.08 | < .001 |
| Speaker Language \times Emotion (Irritation) | 5.77 | 1.56 | 3.69 | < .001 |

| F0-SD | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.29 | | | |
| Residual | 1.29 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 14.30 | 0.24 | 59.35 | < .001 |
| Gender | 0.37 | 0.48 | 0.77 | 0.443 |
| Speaker Language | -0.52 | 0.48 | -1.09 | 0.280 |
| Emotion (Joy) | 0.79 | 0.28 | 2.77 | < .010 |
| Emotion (Pride) | 1.14 | 0.28 | 4.03 | < .001 |
| Emotion (Anger) | 1.36 | 0.28 | 4.80 | < .001 |
| Emotion (Tenderness) | 0.42 | 0.28 | 1.49 | 0.138 |
| Emotion (Relief) | 0.09 | 0.28 | 0.31 | 0.760 |
| Emotion (Sadness) | 0.15 | 0.28 | 0.54 | 0.591 |
| Emotion (Irritation) | 1.38 | 0.28 | 4.88 | < .001 |
| Gender × Speaker Language | 0.79 | 0.96 | 0.82 | 0.417 |
| Gender × Emotion (Joy) | 0.11 | 0.57 | 0.19 | 0.847 |
| Gender × Emotion (Pride) | -0.04 | 0.57 | -0.07 | 0.946 |
| Gender × Emotion (Anger) | -0.09 | 0.57 | -0.17 | 0.867 |
| Gender × Emotion (Tenderness) | -0.69 | 0.57 | -1.22 | 0.226 |
| Gender × Emotion (Relief) | 0.13 | 0.57 | 0.23 | 0.817 |
| Gender × Emotion (Sadness) | -0.71 | 0.57 | -1.25 | 0.212 |
| Gender × Emotion (Irritation) | 0.24 | 0.57 | 0.42 | 0.677 |
| Speaker Language × Emotion (Joy) | -0.39 | 0.57 | -0.70 | 0.488 |
| Speaker Language × Emotion (Pride) | -1.08 | 0.57 | -1.91 | 0.057 |
| Speaker Language × Emotion (Anger) | 0.11 | 0.57 | 0.20 | 0.847 |
| Speaker Language × Emotion (Tenderness) | 0.25 | 0.57 | 0.44 | 0.657 |
| Speaker Language × Emotion (Relief) | -0.37 | 0.57 | -0.65 | 0.513 |
| Speaker Language × Emotion (Sadness) | 0.41 | 0.57 | 0.73 | 0.466 |
| Speaker Language × Emotion (Irritation) | -0.28 | 0.57 | -0.49 | 0.621 |
| Gender × Speaker Language × Emotion (Joy) | 0.16 | 1.13 | 0.14 | 0.891 |
| Gender × Speaker Language × Emotion (Pride) | -2.95 | 1.13 | -2.60 | < .050 |
| Gender × Speaker Language × Emotion (Anger) | -1.74 | 1.13 | -1.53 | 0.127 |
| Gender × Speaker Language × Emotion (Tenderness) | -3.39 | 1.13 | -2.99 | < .010 |
| Gender × Speaker Language × Emotion (Relief) | -0.10 | 1.13 | -0.09 | 0.928 |
| Gender × Speaker Language × Emotion (Sadness) | -0.49 | 1.13 | -0.43 | 0.665 |
| Gender × Speaker Language × Emotion (Irritation) | -0.46 | 1.13 | -0.40 | 0.686 |
| F0-min | | | | |
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 1.94 | | | |
| Residual | 10.21 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 24.74 | 0.66 | 37.27 | < .001 |
| Gender | -7.41 | 0.80 | -9.22 | < .001 |
| Speaker Language | -3.25 | 1.33 | -2.45 | < .050 |
| Emotion (Joy) | -3.08 | 0.80 | -3.86 | < .001 |
| Emotion (Pride) | -6.19 | 0.80 | -7.75 | < .001 |
| Emotion (Anger) | -5.17 | 0.80 | -6.47 | < .001 |
| Emotion (Tenderness) | -8.57 | 0.80 | -10.73 | < .001 |

| | | | | |
|---|----------|-----------|----------|----------|
| Emotion (Relief) | -6.43 | 0.80 | -8.05 | < .001 |
| Emotion (Sadness) | -5.69 | 0.80 | -7.12 | < .001 |
| Emotion (Irritation) | -7.41 | 0.80 | -9.28 | < .001 |
| Speaker Language × Emotion (Joy) | 0.86 | 1.60 | 0.54 | 0.592 |
| Speaker Language × Emotion (Pride) | 7.63 | 1.60 | 4.78 | < .001 |
| Speaker Language × Emotion (Anger) | 3.52 | 1.60 | 2.20 | < .050 |
| Speaker Language × Emotion (Tenderness) | 3.30 | 1.60 | 2.06 | < .050 |
| Speaker Language × Emotion (Relief) | 5.03 | 1.60 | 3.15 | < .010 |
| Speaker Language × Emotion (Sadness) | 4.33 | 1.60 | 2.71 | < .010 |
| Speaker Language × Emotion (Irritation) | 4.98 | 1.60 | 3.12 | < .010 |
| F0-max | | | | |
| Random effects | Variance | | | |
| Speaker (Intercept) | 3.37 | | | |
| Residual | 18.43 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 34.67 | 0.89 | 39.09 | < .001 |
| Gender | -7.51 | 1.06 | -7.07 | < .001 |
| Speaker Language | -6.77 | 1.77 | -3.82 | < .001 |
| Emotion (Joy) | -0.60 | 1.07 | -0.56 | 0.574 |
| Emotion (Pride) | -3.39 | 1.07 | -3.16 | < .010 |
| Emotion (Anger) | -0.06 | 1.07 | -0.05 | 0.959 |
| Emotion (Tenderness) | -8.09 | 1.07 | -7.53 | < .001 |
| Emotion (Relief) | -7.97 | 1.07 | -7.42 | < .001 |
| Emotion (Sadness) | -6.57 | 1.07 | -6.12 | < .001 |
| Emotion (Irritation) | -4.86 | 1.07 | -4.53 | < .001 |
| Speaker Language × Emotion (Joy) | 0.71 | 2.15 | 0.33 | 0.740 |
| Speaker Language × Emotion (Pride) | 5.26 | 2.15 | 2.45 | < .050 |
| Speaker Language × Emotion (Anger) | 3.89 | 2.15 | 1.81 | 0.071 |
| Speaker Language × Emotion (Tenderness) | 3.33 | 2.15 | 1.55 | 0.122 |
| Speaker Language × Emotion (Relief) | 4.28 | 2.15 | 1.99 | < .050 |
| Speaker Language × Emotion (Sadness) | 7.28 | 2.15 | 3.39 | < .001 |
| Speaker Language × Emotion (Irritation) | 5.31 | 2.15 | 2.47 | < .050 |
| Sync | | | | |
| Random effects | Variance | | | |
| Speaker (Intercept) | 0.00005 | | | |
| Residual | 0.03 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.58 | 0.03 | 20.17 | < .001 |
| Gender | -0.05 | 0.02 | -2.50 | < .050 |
| Speaker Language | -0.005 | 0.02 | -0.22 | 0.828 |
| Emotion (Joy) | 0.02 | 0.04 | 0.49 | 0.625 |
| Emotion (Pride) | -0.02 | 0.04 | -0.51 | 0.614 |
| Emotion (Anger) | 0.04 | 0.04 | 0.90 | 0.368 |
| Emotion (Tenderness) | 0.002 | 0.04 | 0.05 | 0.957 |
| Emotion (Relief) | -0.08 | 0.04 | -2.06 | < .050 |
| Emotion (Sadness) | -0.07 | 0.04 | -1.81 | 0.072 |
| Emotion (Irritation) | 0.03 | 0.04 | 0.79 | 0.433 |

| Int-M | | | | |
|--|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 14.87 | | | |
| Residual | 29.23 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 69.11 | 1.36 | 50.91 | < .001 |
| Gender | -2.12 | 2.71 | -0.78 | 0.440 |
| Speaker Language | 0.10 | 2.71 | 0.04 | 0.969 |
| Emotion (Joy) | 2.92 | 1.35 | 2.16 | < .050 |
| Emotion (Pride) | -3.08 | 1.35 | -2.28 | < .050 |
| Emotion (Anger) | 4.49 | 1.35 | 3.33 | < .010 |
| Emotion (Tenderness) | -10.49 | 1.35 | -7.76 | < .001 |
| Emotion (Relief) | -8.31 | 1.35 | -6.15 | < .001 |
| Emotion (Sadness) | -11.01 | 1.35 | -8.15 | < .001 |
| Emotion (Irritation) | -5.29 | 1.35 | -3.91 | < .001 |
| Gender \times Speaker Language | 8.82 | 5.43 | 1.62 | 0.113 |
| Gender \times Emotion (Joy) | 1.44 | 2.70 | 0.53 | 0.596 |
| Gender \times Emotion (Pride) | 0.89 | 2.70 | 0.33 | 0.744 |
| Gender \times Emotion (Anger) | 0.52 | 2.70 | 0.19 | 0.848 |
| Gender \times Emotion (Tenderness) | 0.93 | 2.70 | 0.34 | 0.732 |
| Gender \times Emotion (Relief) | 2.33 | 2.70 | 0.86 | 0.389 |
| Gender \times Emotion (Sadness) | 4.69 | 2.70 | 1.73 | 0.084 |
| Gender \times Emotion (Irritation) | 5.02 | 2.70 | 1.86 | 0.065 |
| Speaker Language \times Emotion (Joy) | 2.84 | 2.70 | 1.05 | 0.294 |
| Speaker Language \times Emotion (Pride) | 14.39 | 2.70 | 5.32 | < .001 |
| Speaker Language \times Emotion (Anger) | 7.36 | 2.70 | 2.72 | < .01 |
| Speaker Language \times Emotion (Tenderness) | 18.37 | 2.70 | 6.80 | < .001 |
| Speaker Language \times Emotion (Relief) | 11.27 | 2.70 | 4.17 | < .001 |
| Speaker Language \times Emotion (Sadness) | 17.25 | 2.70 | 6.38 | < .001 |
| Speaker Language \times Emotion (Irritation) | 12.75 | 2.70 | 4.72 | < .001 |
| Gender \times Speaker Language \times Emotion (Joy) | -3.67 | 5.41 | -0.68 | 0.498 |
| Gender \times Speaker Language \times Emotion (Pride) | -5.63 | 5.41 | -1.04 | 0.298 |
| Gender \times Speaker Language \times Emotion (Anger) | -7.37 | 5.41 | -1.36 | 0.174 |
| Gender \times Speaker Language \times Emotion (Tenderness) | 1.39 | 5.41 | 0.26 | 0.798 |
| Gender \times Speaker Language \times Emotion (Relief) | 3.76 | 5.41 | 0.70 | 0.487 |
| Gender \times Speaker Language \times Emotion (Sadness) | 17.58 | 5.41 | 3.25 | < .010 |
| Gender \times Speaker Language \times Emotion (Irritation) | 1.13 | 5.41 | 0.21 | 0.834 |

| Int-SD | | | | |
|--|---------------------------|-----------------|----------|----------------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.17 | | | |
| Residual | 3.07 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 9.06 | 0.33 | 27.74 | < . 001 |
| Gender | -1.18 | 0.65 | -1.80 | 0.074 |
| Speaker Language | -3.36 | 0.65 | -5.14 | < . 001 |
| Emotion (Joy) | -0.54 | 0.44 | -1.24 | 0.216 |
| Emotion (Pride) | -0.28 | 0.44 | -0.63 | 0.530 |
| Emotion (Anger) | 1.06 | 0.44 | 2.43 | < . 050 |
| Emotion (Tenderness) | -1.72 | 0.44 | -3.92 | < . 001 |
| Emotion (Relief) | 0.33 | 0.44 | 0.75 | 0.453 |
| Emotion (Sadness) | -0.22 | 0.44 | -0.50 | 0.618 |
| Emotion (Irritation) | 0.10 | 0.44 | 0.24 | 0.813 |
| Gender \times Speaker Language | 0.93 | 0.60 | 1.55 | 0.147 |
| Gender \times Emotion (Joy) | -0.73 | 0.88 | -0.83 | 0.408 |
| Gender \times Emotion (Pride) | 0.92 | 0.88 | 1.05 | 0.297 |
| Gender \times Emotion (Anger) | 0.22 | 0.88 | 0.25 | 0.805 |
| Gender \times Emotion (Tenderness) | 0.05 | 0.88 | 0.06 | 0.951 |
| Gender \times Emotion (Relief) | 0.20 | 0.88 | 0.23 | 0.816 |
| Gender \times Emotion (Sadness) | -0.44 | 0.88 | -0.50 | 0.618 |
| Gender \times Emotion (Irritation) | 0.55 | 0.88 | 0.62 | 0.534 |
| Speaker Language \times Emotion (Joy) | 0.14 | 0.88 | 0.16 | 0.873 |
| Speaker Language \times Emotion (Pride) | 1.21 | 0.88 | 1.38 | 0.168 |
| Speaker Language \times Emotion (Anger) | 0.82 | 0.88 | 0.94 | 0.348 |
| Speaker Language \times Emotion (Tenderness) | 2.88 | 0.88 | 3.29 | < . 010 |
| Speaker Language \times Emotion (Relief) | 4.65 | 0.88 | 5.31 | < . 001 |
| Speaker Language \times Emotion (Sadness) | 3.78 | 0.88 | 4.31 | < . 001 |
| Speaker Language \times Emotion (Irritation) | 1.10 | 0.88 | 1.26 | 0.210 |

| Tilt | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 3.33 | | | |
| Residual | 6.71 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | -6.28 | 0.65 | -9.72 | < .001 |
| Gender | 2.49 | 1.29 | 1.93 | 0.062 |
| Speaker Language | 4.39 | 1.29 | 3.39 | < .010 |
| Emotion (Joy) | -1.80 | 0.65 | -2.77 | < .010 |
| Emotion (Pride) | -0.69 | 0.65 | -1.07 | 0.288 |
| Emotion (Anger) | -1.10 | 0.65 | -1.69 | 0.092 |
| Emotion (Tenderness) | 1.24 | 0.65 | 1.91 | 0.057 |
| Emotion (Relief) | 1.70 | 0.65 | 2.63 | < .010 |
| Emotion (Sadness) | 1.38 | 0.65 | 2.14 | < .050 |
| Emotion (Irritation) | 1.62 | 0.65 | 2.50 | < .050 |
| Gender × Speaker Language | -4.04 | 2.58 | -1.56 | 0.127 |
| Gender × Emotion (Joy) | -2.77 | 1.29 | -2.14 | < .050 |
| Gender × Emotion (Pride) | -2.02 | 1.29 | -1.56 | 0.120 |
| Gender × Emotion (Anger) | -0.80 | 1.29 | -0.62 | 0.537 |
| Gender × Emotion (Tenderness) | -4.58 | 1.29 | -3.53 | < .001 |
| Gender × Emotion (Relief) | -4.71 | 1.29 | -3.63 | < .001 |
| Gender × Emotion (Sadness) | -3.79 | 1.29 | -2.93 | < .010 |
| Gender × Emotion (Irritation) | -4.63 | 1.29 | -3.57 | < .001 |
| Speaker Language × Emotion (Joy) | -3.53 | 1.29 | -2.73 | < .010 |
| Speaker Language × Emotion (Pride) | -6.29 | 1.29 | -4.86 | < .001 |
| Speaker Language × Emotion (Anger) | -4.64 | 1.29 | -3.58 | < .001 |
| Speaker Language × Emotion (Tenderness) | -7.52 | 1.29 | -5.80 | < .001 |
| Speaker Language × Emotion (Relief) | -5.65 | 1.29 | -4.36 | < .001 |
| Speaker Language × Emotion (Sadness) | -3.89 | 1.29 | -3.01 | < .010 |
| Speaker Language × Emotion (Irritation) | -5.52 | 1.29 | -4.26 | < .001 |
| Gender × Speaker Language × Emotion (Joy) | 1.76 | 2.59 | 0.68 | 0.497 |
| Gender × Speaker Language × Emotion (Pride) | 6.81 | 2.59 | 2.63 | < .010 |
| Gender × Speaker Language × Emotion (Anger) | 2.13 | 2.59 | 0.82 | 0.411 |
| Gender × Speaker Language × Emotion (Tenderness) | 4.70 | 2.59 | 1.81 | 0.071 |
| Gender × Speaker Language × Emotion (Relief) | 4.63 | 2.59 | 1.79 | 0.075 |
| Gender × Speaker Language × Emotion (Sadness) | 2.44 | 2.59 | 0.94 | 0.348 |
| Gender × Speaker Language × Emotion (Irritation) | 7.32 | 2.59 | 2.83 | < .010 |

| F1 | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.06 | | | |
| Residual | 0.40 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 6.43 | 0.13 | 50.65 | < .001 |
| Gender | 0.38 | 0.25 | 1.51 | 0.135 |
| Speaker Language | -0.10 | 0.25 | -0.41 | 0.682 |
| Emotion (Joy) | -0.11 | 0.16 | -0.70 | 0.483 |
| Emotion (Pride) | -0.18 | 0.16 | -1.12 | 0.263 |
| Emotion (Anger) | 0.02 | 0.16 | 0.15 | 0.880 |
| Emotion (Tenderness) | -0.76 | 0.16 | -4.79 | < .001 |
| Emotion (Relief) | -0.19 | 0.16 | -1.21 | 0.226 |
| Emotion (Sadness) | -0.31 | 0.16 | -1.96 | 0.052 |
| Emotion (Irritation) | -0.27 | 0.16 | -1.70 | 0.091 |
| Gender \times Speaker Language | 2.25 | 0.51 | 4.43 | < .001 |
| Gender \times Emotion (Joy) | -0.01 | 0.32 | -0.03 | 0.975 |
| Gender \times Emotion (Pride) | 0.21 | 0.32 | 0.66 | 0.511 |
| Gender \times Emotion (Anger) | -0.64 | 0.32 | -2.01 | < .050 |
| Gender \times Emotion (Tenderness) | -0.19 | 0.32 | -0.59 | 0.557 |
| Gender \times Emotion (Relief) | -0.17 | 0.32 | -0.52 | 0.604 |
| Gender \times Emotion (Sadness) | 0.11 | 0.32 | 0.35 | 0.724 |
| Gender \times Emotion (Irritation) | -0.31 | 0.32 | -0.98 | 0.330 |
| Speaker Language \times Emotion (Joy) | -0.28 | 0.32 | -0.87 | 0.384 |
| Speaker Language \times Emotion (Pride) | 0.59 | 0.32 | 1.85 | 0.065 |
| Speaker Language \times Emotion (Anger) | -0.07 | 0.32 | -0.23 | 0.818 |
| Speaker Language \times Emotion (Tenderness) | -0.50 | 0.32 | -1.56 | 0.120 |
| Speaker Language \times Emotion (Relief) | -0.38 | 0.32 | -1.19 | 0.237 |
| Speaker Language \times Emotion (Sadness) | -0.26 | 0.32 | -0.83 | 0.410 |
| Speaker Language \times Emotion (Irritation) | -0.04 | 0.32 | -0.14 | 0.892 |
| Gender \times Speaker Language \times Emotion (Joy) | -0.33 | 0.64 | -0.52 | 0.602 |
| Gender \times Speaker Language \times Emotion (Pride) | -1.34 | 0.64 | -2.10 | < .050 |
| Gender \times Speaker Language \times Emotion (Anger) | -2.82 | 0.64 | -4.43 | < .001 |
| Gender \times Speaker Language \times Emotion (Tenderness) | -2.85 | 0.64 | -4.48 | < .001 |
| Gender \times Speaker Language \times Emotion (Relief) | -2.36 | 0.64 | -3.71 | < .001 |
| Gender \times Speaker Language \times Emotion (Sadness) | -2.88 | 0.64 | -4.53 | < .001 |
| Gender \times Speaker Language \times Emotion (Irritation) | -1.14 | 0.64 | -1.79 | 0.075 |

| F2 | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.04 | | | |
| Residual | 0.18 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 12.10 | 0.09 | 132.83 | < .001 |
| Gender | 0.50 | 0.18 | 2.72 | < .010 |
| Speaker Language | 0.66 | 0.18 | 3.63 | < .001 |
| Emotion (Joy) | -0.20 | 0.10 | -1.87 | 0.063 |
| Emotion (Pride) | -0.15 | 0.10 | -1.42 | 0.157 |
| Emotion (Anger) | -0.07 | 0.10 | -0.69 | 0.488 |
| Emotion (Tenderness) | -0.06 | 0.10 | -0.55 | 0.586 |
| Emotion (Relief) | 0.05 | 0.10 | 0.46 | 0.647 |
| Emotion (Sadness) | 0.00 | 0.10 | 0.04 | 0.964 |
| Emotion (Irritation) | -0.12 | 0.10 | -1.10 | 0.270 |
| Gender \times Speaker Language | 0.32 | 0.36 | 0.89 | 0.379 |
| Gender \times Emotion (Joy) | -0.16 | 0.21 | -0.75 | 0.452 |
| Gender \times Emotion (Pride) | 0.21 | 0.21 | 1.01 | 0.314 |
| Gender \times Emotion (Anger) | -0.08 | 0.21 | -0.37 | 0.714 |
| Gender \times Emotion (Tenderness) | -0.36 | 0.21 | -1.74 | 0.084 |
| Gender \times Emotion (Relief) | -0.03 | 0.21 | -0.14 | 0.891 |
| Gender \times Emotion (Sadness) | 0.13 | 0.21 | 0.64 | 0.523 |
| Gender \times Emotion (Irritation) | -0.27 | 0.21 | -1.30 | 0.197 |
| Speaker Language \times Emotion (Joy) | -0.27 | 0.21 | -1.31 | 0.192 |
| Speaker Language \times Emotion (Pride) | -0.55 | 0.21 | -2.62 | < .010 |
| Speaker Language \times Emotion (Anger) | -0.81 | 0.21 | -3.88 | < .001 |
| Speaker Language \times Emotion (Tenderness) | -1.17 | 0.21 | -5.57 | < .001 |
| Speaker Language \times Emotion (Relief) | -0.80 | 0.21 | -3.81 | < .001 |
| Speaker Language \times Emotion (Sadness) | -0.59 | 0.21 | -2.82 | < .010 |
| Speaker Language \times Emotion (Irritation) | -0.84 | 0.21 | -4.01 | < .001 |
| Gender \times Speaker Language \times Emotion (Joy) | 0.001 | 0.42 | 0.001 | 0.998 |
| Gender \times Speaker Language \times Emotion (Pride) | 0.20 | 0.42 | 0.48 | 0.632 |
| Gender \times Speaker Language \times Emotion (Anger) | -0.43 | 0.42 | -1.02 | 0.311 |
| Gender \times Speaker Language \times Emotion (Tenderness) | -0.76 | 0.42 | -1.81 | 0.071 |
| Gender \times Speaker Language \times Emotion (Relief) | -0.06 | 0.42 | -0.14 | 0.891 |
| Gender \times Speaker Language \times Emotion (Sadness) | -0.69 | 0.42 | -1.65 | 0.100 |
| Gender \times Speaker Language \times Emotion (Irritation) | 0.33 | 0.42 | 0.78 | 0.438 |

| F3 | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.03 | | | |
| Residual | 0.07 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 15.61 | 0.06 | 256.84 | < .001 |
| Gender | 0.02 | 0.12 | 0.17 | 0.866 |
| Speaker Language | 0.46 | 0.12 | 3.82 | < .001 |
| Emotion (Joy) | -0.12 | 0.07 | -1.81 | 0.071 |
| Emotion (Pride) | -0.09 | 0.07 | -1.31 | 0.190 |
| Emotion (Anger) | 0.00 | 0.07 | -0.06 | 0.954 |
| Emotion (Tenderness) | -0.17 | 0.07 | -2.58 | < .050 |
| Emotion (Relief) | -0.06 | 0.07 | -0.97 | 0.331 |
| Emotion (Sadness) | -0.03 | 0.07 | -0.50 | 0.615 |
| Emotion (Irritation) | -0.01 | 0.07 | -0.18 | 0.856 |
| Gender \times Speaker Language | 0.48 | 0.24 | 1.97 | 0.055 |
| Gender \times Emotion (Joy) | -0.10 | 0.13 | -0.81 | 0.421 |
| Gender \times Emotion (Pride) | -0.01 | 0.13 | -0.08 | 0.939 |
| Gender \times Emotion (Anger) | -0.19 | 0.13 | -1.43 | 0.154 |
| Gender \times Emotion (Tenderness) | -0.33 | 0.13 | -2.57 | < .050 |
| Gender \times Emotion (Relief) | -0.09 | 0.13 | -0.67 | 0.506 |
| Gender \times Emotion (Sadness) | 0.02 | 0.13 | 0.16 | 0.874 |
| Gender \times Emotion (Irritation) | -0.17 | 0.13 | -1.32 | 0.187 |
| Speaker Language \times Emotion (Joy) | -0.20 | 0.13 | -1.51 | 0.134 |
| Speaker Language \times Emotion (Pride) | -0.37 | 0.13 | -2.81 | < .010 |
| Speaker Language \times Emotion (Anger) | -0.43 | 0.13 | -3.31 | < .010 |
| Speaker Language \times Emotion (Tenderness) | -0.59 | 0.13 | -4.52 | < .001 |
| Speaker Language \times Emotion (Relief) | -0.41 | 0.13 | -3.11 | < .010 |
| Speaker Language \times Emotion (Sadness) | -0.50 | 0.13 | -3.84 | < .001 |
| Speaker Language \times Emotion (Irritation) | -0.34 | 0.13 | -2.58 | < .050 |
| Gender \times Speaker Language \times Emotion (Joy) | -0.37 | 0.26 | -1.42 | 0.157 |
| Gender \times Speaker Language \times Emotion (Pride) | 0.07 | 0.26 | 0.29 | 0.774 |
| Gender \times Speaker Language \times Emotion (Anger) | -0.70 | 0.26 | -2.67 | < .010 |
| Gender \times Speaker Language \times Emotion (Tenderness) | -0.68 | 0.26 | -2.59 | < .050 |
| Gender \times Speaker Language \times Emotion (Relief) | -0.29 | 0.26 | -1.10 | 0.273 |
| Gender \times Speaker Language \times Emotion (Sadness) | -0.81 | 0.26 | -3.12 | < .010 |
| Gender \times Speaker Language \times Emotion (Irritation) | -0.05 | 0.26 | -0.17 | 0.863 |

| AR | | | | |
|---|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.28 | | | |
| Residual | 0.66 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 5.48 | 0.20 | 28.08 | < .001 |
| Gender | 0.99 | 0.39 | 2.55 | < .050 |
| Speaker Language | 3.33 | 0.39 | 8.54 | < .001 |
| Emotion (Joy) | -0.48 | 0.20 | -2.38 | < .050 |
| Emotion (Pride) | -0.75 | 0.20 | -3.68 | < .001 |
| Emotion (Anger) | -0.78 | 0.20 | -3.85 | < .001 |
| Emotion (Tenderness) | -0.99 | 0.20 | -4.89 | < .001 |
| Emotion (Relief) | -0.90 | 0.20 | -4.45 | < .001 |
| Emotion (Sadness) | -0.85 | 0.20 | -4.18 | < .001 |
| Emotion (Irritation) | -0.85 | 0.20 | -4.20 | < .001 |
| Gender × Speaker Language | 0.83 | 0.57 | 1.46 | 0.169 |
| Gender × Emotion (Joy) | 0.37 | 0.41 | 0.92 | 0.357 |
| Gender × Emotion (Pride) | -0.45 | 0.41 | -1.10 | 0.271 |
| Gender × Emotion (Anger) | -0.62 | 0.41 | -1.53 | 0.128 |
| Gender × Emotion (Tenderness) | -0.10 | 0.41 | -0.26 | 0.799 |
| Gender × Emotion (Relief) | 0.27 | 0.41 | 0.67 | 0.507 |
| Gender × Emotion (Sadness) | -0.51 | 0.41 | -1.26 | 0.209 |
| Gender × Emotion (Irritation) | -0.57 | 0.41 | -1.40 | 0.163 |
| Speaker Language × Emotion (Joy) | -0.45 | 0.41 | -1.10 | 0.271 |
| Speaker Language × Emotion (Pride) | -1.10 | 0.41 | -2.70 | < .010 |
| Speaker Language × Emotion (Anger) | -0.96 | 0.41 | -2.37 | < .050 |
| Speaker Language × Emotion (Tenderness) | -1.88 | 0.41 | -4.64 | < .001 |
| Speaker Language × Emotion (Relief) | -1.34 | 0.41 | -3.29 | < .010 |
| Speaker Language × Emotion (Sadness) | -1.99 | 0.41 | -4.91 | < .001 |
| Speaker Language × Emotion (Irritation) | -1.27 | 0.41 | -3.13 | < .010 |

| PPQ | | | | |
|--|----------|--------|-------|--------|
| Random effects | Variance | | | |
| Speaker (Intercept) | 1.91 | | | |
| Residual | 1.28 | | | |
| Fixed effects | β | SE | t | p |
| Intercept | < .01 | < .001 | 15.10 | < .001 |
| Gender | < .001 | < .001 | 2.82 | < .010 |
| Speaker Language | < .01 | < .001 | 3.86 | < .001 |
| Emotion (Joy) | < .001 | < .001 | -0.93 | 0.353 |
| Emotion (Pride) | < .001 | < .001 | -2.27 | < .050 |
| Emotion (Anger) | < .001 | < .001 | -1.63 | 0.105 |
| Emotion (Tenderness) | < .001 | < .001 | -0.42 | 0.675 |
| Emotion (Relief) | < .001 | < .001 | 0.75 | 0.455 |
| Emotion (Sadness) | < .001 | < .001 | 1.31 | 0.192 |
| Emotion (Irritation) | < .001 | < .001 | -0.53 | 0.598 |
| Gender \times Speaker Language | < .01 | < .001 | 2.38 | < .050 |
| Gender \times Emotion (Joy) | < .001 | < .001 | 0.55 | 0.585 |
| Gender \times Emotion (Pride) | < .001 | < .001 | -0.86 | 0.389 |
| Gender \times Emotion (Anger) | < .001 | < .001 | -1.95 | 0.050 |
| Gender \times Emotion (Tenderness) | < .001 | < .001 | 0.27 | 0.788 |
| Gender \times Emotion (Relief) | < .001 | < .001 | 0.63 | 0.530 |
| Gender \times Emotion (Sadness) | < .001 | < .001 | 0.65 | 0.515 |
| Gender \times Emotion (Irritation) | < .001 | < .001 | -0.82 | 0.411 |
| Speaker Language \times Emotion (Joy) | < .001 | < .001 | -0.37 | 0.712 |
| Speaker Language \times Emotion (Pride) | < -.01 | < .001 | -3.06 | < .010 |
| Speaker Language \times Emotion (Anger) | < -.01 | < .001 | -3.93 | < .001 |
| Speaker Language \times Emotion (Tenderness) | < -.01 | < .001 | -4.95 | < .001 |
| Speaker Language \times Emotion (Relief) | < -.01 | < .001 | -3.36 | < .001 |
| Speaker Language \times Emotion (Sadness) | < -.01 | < .001 | -3.32 | < .010 |
| Speaker Language \times Emotion (Irritation) | < -.01 | < .001 | -3.12 | < .010 |
| Gender \times Speaker Language \times Emotion (Joy) | < .001 | < .001 | -0.83 | 0.411 |
| Gender \times Speaker Language \times Emotion (Pride) | < .001 | < .001 | -1.20 | 0.233 |
| Gender \times Speaker Language \times Emotion (Anger) | < -.01 | < .001 | -1.41 | 0.161 |
| Gender \times Speaker Language \times Emotion (Tenderness) | < -.01 | < .001 | -3.11 | < .010 |
| Gender \times Speaker Language \times Emotion (Relief) | < -.01 | < .001 | -2.11 | < .050 |
| Gender \times Speaker Language \times Emotion (Sadness) | < -.01 | < .001 | -3.61 | < .001 |
| Gender \times Speaker Language \times Emotion (Irritation) | < -.01 | < .001 | -1.93 | 0.055 |

| APQ | | | | |
|--|---------------------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | < .001 | | | |
| Residual | < .001 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.05 | 0.003 | 14.69 | < .001 |
| Gender | 0.02 | 0.010 | 2.62 | < .050 |
| Speaker Language | 0.01 | 0.010 | 2.04 | < .050 |
| Emotion (Joy) | 0.01 | 0.004 | 1.18 | 0.238 |
| Emotion (Pride) | 0.01 | 0.004 | 1.27 | 0.207 |
| Emotion (Anger) | 0.01 | 0.004 | 1.95 | 0.052 |
| Emotion (Tenderness) | 0.01 | 0.004 | 1.90 | 0.059 |
| Emotion (Relief) | 0.004 | 0.004 | 1.00 | 0.320 |
| Emotion (Sadness) | 0.01 | 0.004 | 2.40 | < .050 |
| Emotion (Irritation) | 0.01 | 0.004 | 1.95 | 0.053 |
| Gender \times Speaker Language | 0.03 | 0.01 | 2.07 | < .050 |
| Gender \times Emotion (Joy) | 0.02 | 0.01 | 1.91 | 0.058 |
| Gender \times Emotion (Pride) | -0.001 | 0.01 | -0.10 | 0.924 |
| Gender \times Emotion (Anger) | -0.01 | 0.01 | -0.72 | 0.470 |
| Gender \times Emotion (Tenderness) | 0.02 | 0.01 | 2.68 | < .010 |
| Gender \times Emotion (Relief) | 0.02 | 0.01 | 2.55 | < .050 |
| Gender \times Emotion (Sadness) | 0.02 | 0.01 | 2.40 | < .050 |
| Gender \times Emotion (Irritation) | 0.003 | 0.01 | 0.37 | 0.713 |
| Speaker Language \times Emotion (Joy) | -0.002 | 0.01 | -0.24 | 0.810 |
| Speaker Language \times Emotion (Pride) | -0.02 | 0.01 | -2.16 | < .050 |
| Speaker Language \times Emotion (Anger) | -0.03 | 0.01 | -3.21 | < .010 |
| Speaker Language \times Emotion (Tenderness) | -0.04 | 0.01 | -5.51 | < .001 |
| Speaker Language \times Emotion (Relief) | -0.02 | 0.01 | -2.66 | < .010 |
| Speaker Language \times Emotion (Sadness) | -0.03 | 0.01 | -3.90 | < .001 |
| Speaker Language \times Emotion (Irritation) | -0.02 | 0.01 | -2.94 | < .010 |
| Gender \times Speaker Language \times Emotion (Joy) | -0.01 | 0.02 | -0.63 | 0.527 |
| Gender \times Speaker Language \times Emotion (Pride) | -0.04 | 0.02 | -2.25 | < .050 |
| Gender \times Speaker Language \times Emotion (Anger) | -0.02 | 0.02 | -1.28 | 0.201 |
| Gender \times Speaker Language \times Emotion (Tenderness) | -0.06 | 0.02 | -4.06 | < .010 |
| Gender \times Speaker Language \times Emotion (Relief) | -0.03 | 0.02 | -2.13 | < .050 |
| Gender \times Speaker Language \times Emotion (Sadness) | -0.07 | 0.02 | -4.44 | < .001 |
| Gender \times Speaker Language \times Emotion (Irritation) | -0.03 | 0.02 | -1.98 | < .050 |

| HNR | | | | |
|--|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.95 | | | |
| Residual | 4.42 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 14.42 | 0.44 | 32.44 | < .001 |
| Gender | -2.49 | 0.89 | -2.80 | < .010 |
| Speaker Language | -2.74 | 0.89 | -3.08 | < .010 |
| Emotion (Joy) | -0.52 | 0.53 | -0.99 | 0.324 |
| Emotion (Pride) | -0.70 | 0.53 | -1.33 | 0.184 |
| Emotion (Anger) | -1.87 | 0.53 | -3.57 | < .001 |
| Emotion (Tenderness) | 0.70 | 0.53 | 1.33 | 0.185 |
| Emotion (Relief) | 0.44 | 0.53 | 0.84 | 0.399 |
| Emotion (Sadness) | -0.36 | 0.53 | -0.68 | 0.494 |
| Emotion (Irritation) | -0.58 | 0.53 | -1.11 | 0.269 |
| Gender \times Speaker Language | -4.69 | 1.78 | -2.64 | < .050 |
| Gender \times Emotion (Joy) | -1.08 | 1.05 | -1.03 | 0.305 |
| Gender \times Emotion (Pride) | -0.08 | 1.05 | -0.07 | 0.942 |
| Gender \times Emotion (Anger) | 0.58 | 1.05 | 0.56 | 0.578 |
| Gender \times Emotion (Tenderness) | -1.28 | 1.05 | -1.21 | 0.226 |
| Gender \times Emotion (Relief) | -1.17 | 1.05 | -1.11 | 0.267 |
| Gender \times Emotion (Sadness) | -0.44 | 1.05 | -0.42 | 0.673 |
| Gender \times Emotion (Irritation) | 0.56 | 1.05 | 0.53 | 0.596 |
| Speaker Language \times Emotion (Joy) | 0.57 | 1.05 | 0.54 | 0.590 |
| Speaker Language \times Emotion (Pride) | 2.09 | 1.05 | 1.99 | < .050 |
| Speaker Language \times Emotion (Anger) | 3.25 | 1.05 | 3.09 | < .010 |
| Speaker Language \times Emotion (Tenderness) | 6.16 | 1.05 | 5.87 | < .001 |
| Speaker Language \times Emotion (Relief) | 4.39 | 1.05 | 4.18 | < .001 |
| Speaker Language \times Emotion (Sadness) | 3.94 | 1.05 | 3.75 | < .001 |
| Speaker Language \times Emotion (Irritation) | 3.32 | 1.05 | 3.16 | < .010 |
| Gender \times Speaker Language \times Emotion (Joy) | 4.05 | 2.10 | 1.93 | 0.055 |
| Gender \times Speaker Language \times Emotion (Pride) | 6.26 | 2.10 | 2.98 | < .010 |
| Gender \times Speaker Language \times Emotion (Anger) | 4.53 | 2.10 | 2.16 | < .050 |
| Gender \times Speaker Language \times Emotion (Tenderness) | 6.92 | 2.10 | 3.30 | < .010 |
| Gender \times Speaker Language \times Emotion (Relief) | 7.35 | 2.10 | 3.50 | < .001 |
| Gender \times Speaker Language \times Emotion (Sadness) | 9.27 | 2.10 | 4.41 | < .001 |
| Gender \times Speaker Language \times Emotion (Irritation) | 4.45 | 2.10 | 2.12 | < .050 |

| Dutch data | | F0-M | | | |
|-----------------------|--|---------------------------|-----------|----------|----------|
| Random effects | | <i>Variance</i> | | | |
| Speaker (Intercept) | | 1.67 | | | |
| Residual | | 8.58 | | | |
| Fixed effects | | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | | 32.88 | 0.86 | 38.10 | < .001 |
| Gender | | -7.18 | 1.05 | -6.84 | < .001 |
| Emotion (Joy) | | -1.48 | 1.04 | -1.43 | 0.155 |
| Emotion (Pride) | | -7.74 | 1.04 | -7.48 | < .001 |
| Emotion (Anger) | | -2.85 | 1.04 | -2.75 | < .010 |
| Emotion (Tenderness) | | -13.09 | 1.04 | -12.64 | < .001 |
| Emotion (Relief) | | -10.43 | 1.04 | -10.07 | < .001 |
| Emotion (Sadness) | | -10.61 | 1.04 | -10.25 | < .001 |
| Emotion (Irritation) | | -8.92 | 1.04 | -8.62 | < .001 |
| | | F0-SD | | | |
| Random effects | | <i>Variance</i> | | | |
| Speaker (Intercept) | | 0.36 | | | |
| Residual | | 1.38 | | | |
| Fixed effects | | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | | 14.56 | 0.36 | 40.33 | < .001 |
| Emotion (Joy) | | 0.98 | 0.41 | 2.37 | < .050 |
| Emotion (Pride) | | 1.68 | 0.41 | 4.06 | < .001 |
| Emotion (Anger) | | 1.31 | 0.41 | 3.15 | < .010 |
| Emotion (Tenderness) | | 0.30 | 0.41 | 0.71 | 0.476 |
| Emotion (Relief) | | 0.27 | 0.41 | 0.66 | 0.512 |
| Emotion (Sadness) | | -0.05 | 0.41 | -0.13 | 0.896 |
| Emotion (Irritation) | | 1.52 | 0.41 | 3.67 | < .001 |
| | | F0-min | | | |
| Random effects | | <i>Variance</i> | | | |
| Speaker (Intercept) | | 1.77 | | | |
| Residual | | 11.08 | | | |
| Fixed effects | | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | | 26.36 | 0.96 | 27.59 | < .001 |
| Gender | | -6.53 | 1.11 | -5.89 | < .010 |
| Emotion (Joy) | | -3.51 | 1.18 | -2.99 | < .010 |
| Emotion (Pride) | | -10.01 | 1.18 | -8.50 | < .001 |
| Emotion (Anger) | | -6.93 | 1.18 | -5.89 | < .001 |
| Emotion (Tenderness) | | -10.22 | 1.18 | -8.69 | < .001 |
| Emotion (Relief) | | -8.94 | 1.18 | -7.60 | < .001 |
| Emotion (Sadness) | | -7.86 | 1.18 | -6.68 | < .001 |
| Emotion (Irritation) | | -9.90 | 1.18 | -8.41 | < .001 |

| F0-max | | | | |
|-----------------------|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 2.58 | | | |
| Residual | 20.77 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 38.05 | 1.27 | 29.89 | < .001 |
| Gender | -6.66 | 1.39 | -4.79 | < .010 |
| Emotion (Joy) | -0.96 | 1.61 | -0.60 | 0.552 |
| Emotion (Pride) | -6.02 | 1.61 | -3.74 | < .001 |
| Emotion (Anger) | -2.00 | 1.61 | -1.24 | 0.216 |
| Emotion (Tenderness) | -9.75 | 1.61 | -6.05 | < .001 |
| Emotion (Relief) | -10.11 | 1.61 | -6.27 | < .001 |
| Emotion (Sadness) | -10.21 | 1.61 | -6.34 | < .001 |
| Emotion (Irritation) | -7.51 | 1.61 | -4.66 | < .001 |
| Sync | | | | |
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.0002 | | | |
| Residual | 0.02 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.54 | 0.04 | 14.52 | < .001 |
| Gender | -0.05 | 0.03 | -1.67 | 0.146 |
| Emotion (Joy) | 0.03 | 0.05 | 0.64 | 0.521 |
| Emotion (Pride) | 0.01 | 0.05 | 0.20 | 0.840 |
| Emotion (Anger) | 0.11 | 0.05 | 2.15 | < .050 |
| Emotion (Tenderness) | 0.04 | 0.05 | 0.79 | 0.433 |
| Emotion (Relief) | -0.06 | 0.05 | -1.20 | 0.231 |
| Emotion (Sadness) | -0.01 | 0.05 | -0.14 | 0.886 |
| Emotion (Irritation) | 0.06 | 0.05 | 1.13 | 0.260 |
| Int-M | | | | |
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 31.08 | | | |
| Residual | 23.22 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 69.06 | 2.31 | 29.90 | < .001 |
| Emotion (Joy) | 1.50 | 1.70 | 0.88 | 0.381 |
| Emotion (Pride) | -10.27 | 1.70 | -6.03 | < .001 |
| Emotion (Anger) | 0.82 | 1.70 | 0.48 | 0.633 |
| Emotion (Tenderness) | -19.67 | 1.70 | -11.55 | < .001 |
| Emotion (Relief) | -13.95 | 1.70 | -8.19 | < .001 |
| Emotion (Sadness) | -19.64 | 1.70 | -11.53 | < .001 |
| Emotion (Irritation) | -11.66 | 1.70 | -6.85 | < .001 |

| Int-SD | | | | |
|-------------------------------|---------------------------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.38 | | | |
| Residual | 2.54 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 10.74 | 0.45 | 23.66 | < .001 |
| Gender | -1.55 | 0.52 | -2.98 | < .050 |
| Emotion (Joy) | -0.61 | 0.56 | -1.09 | 0.278 |
| Emotion (Pride) | -0.88 | 0.56 | -1.57 | 0.120 |
| Emotion (Anger) | 0.65 | 0.56 | 1.16 | 0.249 |
| Emotion (Tenderness) | -3.16 | 0.56 | -5.60 | < .001 |
| Emotion (Relief) | -2.00 | 0.56 | -3.54 | < .001 |
| Emotion (Sadness) | -2.11 | 0.56 | -3.74 | < .001 |
| Emotion (Irritation) | -0.45 | 0.56 | -0.79 | 0.429 |
| Tilt | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 3.73 | | | |
| Residual | 5.25 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | -8.47 | 0.89 | -9.51 | < .001 |
| Gender | 4.51 | 1.78 | 2.53 | < .050 |
| Emotion (Joy) | -0.03 | 0.81 | -0.04 | 0.970 |
| Emotion (Pride) | 2.46 | 0.81 | 3.03 | < .010 |
| Emotion (Anger) | 1.22 | 0.81 | 1.51 | 0.135 |
| Emotion (Tenderness) | 5.00 | 0.81 | 6.17 | < .001 |
| Emotion (Relief) | 4.53 | 0.81 | 5.59 | < .001 |
| Emotion (Sadness) | 3.33 | 0.81 | 4.11 | < .001 |
| Emotion (Irritation) | 4.38 | 0.81 | 5.41 | < .001 |
| Gender × Emotion (Joy) | -3.65 | 1.62 | -2.25 | < .050 |
| Gender × Emotion (Pride) | -5.42 | 1.62 | -3.35 | < .010 |
| Gender × Emotion (Anger) | -1.87 | 1.62 | -1.15 | 0.252 |
| Gender × Emotion (Tenderness) | -6.92 | 1.62 | -4.27 | < .001 |
| Gender × Emotion (Relief) | -7.02 | 1.62 | -4.33 | < .001 |
| Gender × Emotion (Sadness) | -5.01 | 1.62 | -3.09 | < .010 |
| Gender × Emotion (Irritation) | -8.29 | 1.62 | -5.11 | < .001 |

| F1 | | | | |
|-------------------------------|----------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.02 | | | |
| Residual | 0.26 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 6.48 | 0.14 | 46.97 | < .001 |
| Gender | -0.74 | 0.28 | -2.68 | < .010 |
| Emotion (Joy) | 0.03 | 0.18 | 0.15 | 0.881 |
| Emotion (Pride) | -0.47 | 0.18 | -2.64 | < .010 |
| Emotion (Anger) | 0.06 | 0.18 | 0.34 | 0.736 |
| Emotion (Tenderness) | -0.51 | 0.18 | -2.86 | < .010 |
| Emotion (Relief) | 0.00 | 0.18 | -0.02 | 0.981 |
| Emotion (Sadness) | -0.18 | 0.18 | -1.00 | 0.318 |
| Emotion (Irritation) | -0.25 | 0.18 | -1.39 | 0.168 |
| Gender × Emotion (Joy) | 0.16 | 0.36 | 0.44 | 0.664 |
| Gender × Emotion (Pride) | 0.88 | 0.36 | 2.45 | < .050 |
| Gender × Emotion (Anger) | 0.77 | 0.36 | 2.14 | < .050 |
| Gender × Emotion (Tenderness) | 1.24 | 0.36 | 3.45 | < .001 |
| Gender × Emotion (Relief) | 1.02 | 0.36 | 2.83 | < .010 |
| Gender × Emotion (Sadness) | 1.55 | 0.36 | 4.33 | < .001 |
| Gender × Emotion (Irritation) | 0.26 | 0.36 | 0.72 | 0.475 |
| F2 | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.03 | | | |
| Residual | 0.15 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 11.77 | 0.12 | 102.00 | < .001 |
| Gender | 0.35 | 0.14 | 2.51 | < .050 |
| Emotion (Joy) | -0.06 | 0.14 | -0.43 | 0.671 |
| Emotion (Pride) | 0.13 | 0.14 | 0.91 | 0.365 |
| Emotion (Anger) | 0.33 | 0.14 | 2.42 | < .050 |
| Emotion (Tenderness) | 0.53 | 0.14 | 3.81 | < .001 |
| Emotion (Relief) | 0.45 | 0.14 | 3.24 | < .010 |
| Emotion (Sadness) | 0.30 | 0.14 | 2.18 | < .050 |
| Emotion (Irritation) | 0.30 | 0.14 | 2.21 | < .050 |

| F3 | | | | |
|-------------------------------|----------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.03 | | | |
| Residual | 0.05 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 15.38 | 0.08 | 186.32 | < .001 |
| Gender | -0.22 | 0.17 | -1.33 | 0.204 |
| Emotion (Joy) | -0.02 | 0.08 | -0.26 | 0.799 |
| Emotion (Pride) | 0.10 | 0.08 | 1.25 | 0.215 |
| Emotion (Anger) | 0.21 | 0.08 | 2.71 | < .010 |
| Emotion (Tenderness) | 0.13 | 0.08 | 1.62 | 0.107 |
| Emotion (Relief) | 0.14 | 0.08 | 1.78 | 0.077 |
| Emotion (Sadness) | 0.22 | 0.08 | 2.78 | < .010 |
| Emotion (Irritation) | 0.16 | 0.08 | 2.00 | < .050 |
| Gender × Emotion (Joy) | 0.08 | 0.16 | 0.51 | 0.610 |
| Gender × Emotion (Pride) | -0.05 | 0.16 | -0.30 | 0.762 |
| Gender × Emotion (Anger) | 0.16 | 0.16 | 1.03 | 0.305 |
| Gender × Emotion (Tenderness) | 0.00 | 0.16 | 0.02 | 0.981 |
| Gender × Emotion (Relief) | 0.06 | 0.16 | 0.36 | 0.720 |
| Gender × Emotion (Sadness) | 0.43 | 0.16 | 2.74 | < .010 |
| Gender × Emotion (Irritation) | -0.15 | 0.16 | -0.96 | 0.339 |
| AR | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.10 | | | |
| Residual | 0.40 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 3.81 | 0.19 | 19.66 | < .001 |
| Gender | 0.68 | 0.39 | 1.76 | 0.088 |
| Emotion (Joy) | -0.26 | 0.22 | -1.16 | 0.250 |
| Emotion (Pride) | -0.20 | 0.22 | -0.89 | 0.375 |
| Emotion (Anger) | -0.30 | 0.22 | -1.34 | 0.182 |
| Emotion (Tenderness) | -0.05 | 0.22 | -0.23 | 0.820 |
| Emotion (Relief) | -0.24 | 0.22 | -1.05 | 0.297 |
| Emotion (Sadness) | 0.15 | 0.22 | 0.66 | 0.512 |
| Emotion (Irritation) | -0.22 | 0.22 | -0.97 | 0.333 |
| Gender × Emotion (Joy) | 0.23 | 0.45 | 0.50 | 0.615 |
| Gender × Emotion (Pride) | -0.53 | 0.45 | -1.18 | 0.240 |
| Gender × Emotion (Anger) | -0.91 | 0.45 | -2.02 | < .050 |
| Gender × Emotion (Tenderness) | -0.56 | 0.45 | -1.24 | 0.217 |
| Gender × Emotion (Relief) | 0.75 | 0.45 | 1.67 | 0.098 |
| Gender × Emotion (Sadness) | -0.56 | 0.45 | -1.26 | 0.211 |
| Gender × Emotion (Irritation) | -0.84 | 0.45 | -1.88 | 0.064 |

| PPQ | | | | |
|-------------------------------|---------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 3.50 | | | |
| Residual | 1.26 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.008 | 0.001 | 7.32 | < .001 |
| Gender | 0.001 | 0.002 | 0.28 | 0.781 |
| Emotion (Joy) | -0.001 | 0.001 | -0.40 | 0.690 |
| Emotion (Pride) | 0.001 | 0.001 | 0.56 | 0.575 |
| Emotion (Anger) | 0.002 | 0.001 | 1.64 | 0.103 |
| Emotion (Tenderness) | 0.004 | 0.001 | 3.23 | < .010 |
| Emotion (Relief) | 0.004 | 0.001 | 2.93 | < .010 |
| Emotion (Sadness) | 0.004 | 0.001 | 3.30 | < .010 |
| Emotion (Irritation) | 0.002 | 0.001 | 1.85 | 0.068 |
| Gender × Emotion (Joy) | 0.002 | 0.003 | 0.98 | 0.330 |
| Gender × Emotion (Pride) | 0.001 | 0.003 | 0.24 | 0.813 |
| Gender × Emotion (Anger) | -0.001 | 0.003 | -0.39 | 0.698 |
| Gender × Emotion (Tenderness) | 0.006 | 0.003 | 2.41 | < .050 |
| Gender × Emotion (Relief) | 0.005 | 0.003 | 1.95 | 0.054 |
| Gender × Emotion (Sadness) | 0.008 | 0.003 | 3.04 | < .010 |
| Gender × Emotion (Irritation) | 0.002 | 0.003 | 0.79 | 0.432 |
| APQ | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 0.0001 | | | |
| Residual | 0.0003 | | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.04 | 0.01 | 7.55 | < .001 |
| Gender | 0.00 | 0.01 | 0.33 | 0.745 |
| Emotion (Joy) | 0.01 | 0.01 | 0.92 | 0.359 |
| Emotion (Pride) | 0.01 | 0.01 | 2.22 | < .050 |
| Emotion (Anger) | 0.02 | 0.01 | 3.34 | < .010 |
| Emotion (Tenderness) | 0.03 | 0.01 | 4.79 | < .001 |
| Emotion (Relief) | 0.01 | 0.01 | 2.37 | < .050 |
| Emotion (Sadness) | 0.02 | 0.01 | 4.08 | < .001 |
| Emotion (Irritation) | 0.02 | 0.01 | 3.16 | < .010 |
| Gender × Emotion (Joy) | 0.02 | 0.01 | 1.65 | 0.103 |
| Gender × Emotion (Pride) | 0.02 | 0.01 | 1.39 | 0.167 |
| Gender × Emotion (Anger) | 0.00 | 0.01 | 0.36 | 0.719 |
| Gender × Emotion (Tenderness) | 0.05 | 0.01 | 4.36 | < .001 |
| Gender × Emotion (Relief) | 0.04 | 0.01 | 3.03 | < .010 |
| Gender × Emotion (Sadness) | 0.05 | 0.01 | 4.42 | < .001 |
| Gender × Emotion (Irritation) | 0.02 | 0.01 | 1.52 | 0.131 |

| HNR | | | | |
|--------------------------------------|-----------------|-----------|----------|----------|
| Random effects | | | | |
| Speaker (Intercept) | <i>Variance</i> | | | |
| | 0.78 | | | |
| Residual | 3.72 | | | |
| Fixed effects | | | | |
| | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 15.79 | 0.57 | 27.49 | < .001 |
| Gender | -0.15 | 1.15 | -0.13 | 0.898 |
| Emotion (Joy) | -0.80 | 0.68 | -1.18 | 0.242 |
| Emotion (Pride) | -1.75 | 0.68 | -2.56 | < .050 |
| Emotion (Anger) | -3.50 | 0.68 | -5.13 | < .001 |
| Emotion (Tenderness) | -2.38 | 0.68 | -3.50 | < .001 |
| Emotion (Relief) | -1.75 | 0.68 | -2.57 | < .050 |
| Emotion (Sadness) | -2.33 | 0.68 | -3.42 | < .001 |
| Emotion (Irritation) | -2.24 | 0.68 | -3.29 | < .010 |
| Gender \times Emotion (Joy) | -3.11 | 1.36 | -2.28 | < .050 |
| Gender \times Emotion (Pride) | -3.21 | 1.36 | -2.35 | < .050 |
| Gender \times Emotion (Anger) | -1.68 | 1.36 | -1.23 | 0.221 |
| Gender \times Emotion (Tenderness) | -4.74 | 1.36 | -3.47 | < .001 |
| Gender \times Emotion (Relief) | -4.84 | 1.36 | -3.55 | < .001 |
| Gender \times Emotion (Sadness) | -5.08 | 1.36 | -3.73 | < .001 |
| Gender \times Emotion (Irritation) | -1.67 | 1.36 | -1.22 | 0.224 |

| Korean data | | F0-M | | | |
|-------------------------------|--|---------------------------|-----------|----------|----------|
| Random effects | | <i>Variance</i> | | | |
| Speaker (Intercept) | | 3.93 | | | |
| Residual | | 10.92 | | | |
| Fixed effects | | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | | 27.21 | 1.08 | 25.11 | < .001 |
| Gender | | -8.68 | 1.52 | -5.72 | < .010 |
| Emotion (Joy) | | -1.20 | 1.17 | -1.03 | 0.305 |
| Emotion (Pride) | | -1.25 | 1.17 | -1.07 | 0.286 |
| Emotion (Anger) | | 0.96 | 1.17 | 0.82 | 0.412 |
| Emotion (Tenderness) | | -6.58 | 1.17 | -5.63 | < .001 |
| Emotion (Relief) | | -4.76 | 1.17 | -4.08 | < .001 |
| Emotion (Sadness) | | -4.24 | 1.17 | -3.62 | < .001 |
| Emotion (Irritation) | | -3.16 | 1.17 | -2.70 | < .010 |
| | | F0-SD | | | |
| Random effects | | <i>Variance</i> | | | |
| Speaker (Intercept) | | 0.20 | | | |
| Residual | | 1.21 | | | |
| Fixed effects | | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | | 14.04 | 0.32 | 44.37 | < .001 |
| Gender | | 0.76 | 0.63 | 1.21 | 0.233 |
| Emotion (Joy) | | 0.59 | 0.39 | 1.52 | 0.133 |
| Emotion (Pride) | | 0.60 | 0.39 | 1.54 | 0.126 |
| Emotion (Anger) | | 1.42 | 0.39 | 3.65 | < .001 |
| Emotion (Tenderness) | | 0.55 | 0.39 | 1.41 | 0.161 |
| Emotion (Relief) | | -0.10 | 0.39 | -0.25 | 0.800 |
| Emotion (Sadness) | | 0.36 | 0.39 | 0.93 | 0.356 |
| Emotion (Irritation) | | 1.24 | 0.39 | 3.20 | < .010 |
| Gender × Emotion (Joy) | | 0.19 | 0.78 | 0.24 | 0.810 |
| Gender × Emotion (Pride) | | -1.51 | 0.78 | -1.95 | 0.054 |
| Gender × Emotion (Anger) | | -0.96 | 0.78 | -1.24 | 0.218 |
| Gender × Emotion (Tenderness) | | -2.38 | 0.78 | -3.07 | < .010 |
| Gender × Emotion (Relief) | | 0.08 | 0.78 | 0.10 | 0.918 |
| Gender × Emotion (Sadness) | | -0.95 | 0.78 | -1.23 | 0.222 |
| Gender × Emotion (Irritation) | | 0.01 | 0.78 | 0.01 | 0.992 |

| F0-min | | | | |
|-------------------------------|-----------------|-----------|----------|----------|
| Random effects | | | | |
| | <i>Variance</i> | | | |
| Speaker (Intercept) | 2.09 | | | |
| Residual | 8.47 | | | |
| Fixed effects | | | | |
| | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 23.11 | 0.89 | 25.99 | < .001 |
| Gender | -12.75 | 1.78 | -7.17 | < .001 |
| Emotion (Joy) | -2.66 | 1.03 | -2.58 | < .050 |
| Emotion (Pride) | -2.38 | 1.03 | -2.31 | < .050 |
| Emotion (Anger) | -3.41 | 1.03 | -3.32 | < .010 |
| Emotion (Tenderness) | -6.92 | 1.03 | -6.73 | < .001 |
| Emotion (Relief) | -3.91 | 1.03 | -3.80 | < .001 |
| Emotion (Sadness) | -3.53 | 1.03 | -3.43 | < .001 |
| Emotion (Irritation) | -4.92 | 1.03 | -4.78 | < .001 |
| Gender × Emotion (Joy) | 3.19 | 2.06 | 1.55 | 0.125 |
| Gender × Emotion (Pride) | 6.37 | 2.06 | 3.09 | < .010 |
| Gender × Emotion (Anger) | 4.39 | 2.06 | 2.13 | < .050 |
| Gender × Emotion (Tenderness) | 6.56 | 2.06 | 3.19 | < .010 |
| Gender × Emotion (Relief) | 3.44 | 2.06 | 1.67 | 0.098 |
| Gender × Emotion (Sadness) | 7.41 | 2.06 | 3.60 | < .001 |
| Gender × Emotion (Irritation) | 4.40 | 2.06 | 2.14 | < .050 |
| F0-max | | | | |
| Random effects | | | | |
| | <i>Variance</i> | | | |
| Speaker (Intercept) | 4.43 | | | |
| Residual | 16.09 | | | |
| Fixed effects | | | | |
| | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 31.28 | 1.25 | 25.05 | < .001 |
| Gender | -8.35 | 1.65 | -5.07 | < .010 |
| Emotion (Joy) | -0.25 | 1.42 | -0.17 | 0.862 |
| Emotion (Pride) | -0.76 | 1.42 | -0.54 | 0.591 |
| Emotion (Anger) | 1.89 | 1.42 | 1.33 | 0.185 |
| Emotion (Tenderness) | -6.42 | 1.42 | -4.53 | < .001 |
| Emotion (Relief) | -5.83 | 1.42 | -4.11 | < .001 |
| Emotion (Sadness) | -2.93 | 1.42 | -2.06 | < .050 |
| Emotion (Irritation) | -2.21 | 1.42 | -1.56 | 0.123 |

| Sync | | | | |
|-------------------------------|-----------------|-----------|----------|----------|
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.0002 | | | |
| Residual | 0.03 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.61 | 0.04 | 13.92 | < .001 |
| Gender | -0.06 | 0.03 | -1.73 | 0.133 |
| Emotion (Joy) | 0.01 | 0.06 | 0.09 | 0.927 |
| Emotion (Pride) | -0.05 | 0.06 | -0.84 | 0.405 |
| Emotion (Anger) | -0.04 | 0.06 | -0.65 | 0.516 |
| Emotion (Tenderness) | -0.04 | 0.06 | -0.60 | 0.549 |
| Emotion (Relief) | -0.10 | 0.06 | -1.67 | 0.097 |
| Emotion (Sadness) | -0.14 | 0.06 | -2.25 | < .050 |
| Emotion (Irritation) | 0.00 | 0.06 | 0.06 | 0.951 |
| Int-M | | | | |
| Random effects | <i>Variance</i> | | | |
| Speaker (Intercept) | 1.54 | | | |
| Residual | 35.98 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 69.16 | 1.56 | 44.26 | < .001 |
| Gender | 2.29 | 3.12 | 0.73 | 0.466 |
| Emotion (Joy) | 4.34 | 2.12 | 2.05 | < .050 |
| Emotion (Pride) | 4.11 | 2.12 | 1.94 | 0.055 |
| Emotion (Anger) | 8.17 | 2.12 | 3.85 | < .001 |
| Emotion (Tenderness) | -1.30 | 2.12 | -0.62 | 0.540 |
| Emotion (Relief) | -2.67 | 2.12 | -1.26 | 0.210 |
| Emotion (Sadness) | -2.39 | 2.12 | -1.13 | 0.263 |
| Emotion (Irritation) | 1.09 | 2.12 | 0.51 | 0.610 |
| Gender × Emotion (Joy) | -0.40 | 4.24 | -0.09 | 0.925 |
| Gender × Emotion (Pride) | -1.93 | 4.24 | -0.46 | 0.650 |
| Gender × Emotion (Anger) | -3.17 | 4.24 | -0.75 | 0.457 |
| Gender × Emotion (Tenderness) | 1.62 | 4.24 | 0.38 | 0.703 |
| Gender × Emotion (Relief) | 4.21 | 4.24 | 0.99 | 0.323 |
| Gender × Emotion (Sadness) | 13.48 | 4.24 | 3.18 | < .010 |
| Gender × Emotion (Irritation) | 5.59 | 4.24 | 1.32 | 0.191 |

| Int-SD | | | | |
|-----------------------|-----------------|-----------|----------|----------|
| Random effects | | | | |
| | <i>Variance</i> | | | |
| Speaker (Intercept) | 0.05 | | | |
| Residual | 3.55 | | | |
| Fixed effects | | | | |
| | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 7.38 | 0.48 | 15.47 | < .001 |
| Emotion (Joy) | -0.47 | 0.67 | -0.71 | 0.478 |
| Emotion (Pride) | 0.33 | 0.67 | 0.50 | 0.621 |
| Emotion (Anger) | 1.48 | 0.67 | 2.22 | < .050 |
| Emotion (Tenderness) | -0.28 | 0.67 | -0.41 | 0.679 |
| Emotion (Relief) | 2.66 | 0.67 | 3.99 | < .001 |
| Emotion (Sadness) | 1.67 | 0.67 | 2.51 | < .050 |
| Emotion (Irritation) | 0.65 | 0.67 | 0.98 | 0.328 |
| Tilt | | | | |
| Random effects | | | | |
| | <i>Variance</i> | | | |
| Speaker (Intercept) | 2.52 | | | |
| Residual | 8.19 | | | |
| Fixed effects | | | | |
| | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | -4.09 | 0.91 | -4.49 | < .001 |
| Emotion (Joy) | -3.56 | 1.01 | -3.52 | < .001 |
| Emotion (Pride) | -3.84 | 1.01 | -3.79 | < .001 |
| Emotion (Anger) | -3.42 | 1.01 | -3.38 | < .010 |
| Emotion (Tenderness) | -2.52 | 1.01 | -2.49 | < .050 |
| Emotion (Relief) | -1.12 | 1.01 | -1.11 | 0.270 |
| Emotion (Sadness) | -0.56 | 1.01 | -0.56 | 0.579 |
| Emotion (Irritation) | -1.14 | 1.01 | -1.13 | 0.263 |

| F1 | | | | |
|--------------------------------------|---------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | | 0.09 | | |
| Residual | | 0.55 | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 6.38 | 0.21 | 29.93 | < .001 |
| Gender | 1.51 | 0.43 | 3.54 | < .001 |
| Emotion (Joy) | -0.25 | 0.26 | -0.95 | 0.342 |
| Emotion (Pride) | 0.12 | 0.26 | 0.44 | 0.659 |
| Emotion (Anger) | -0.01 | 0.26 | -0.05 | 0.962 |
| Emotion (Tenderness) | -1.01 | 0.26 | -3.85 | < .001 |
| Emotion (Relief) | -0.38 | 0.26 | -1.45 | 0.149 |
| Emotion (Sadness) | -0.44 | 0.26 | -1.68 | 0.095 |
| Emotion (Irritation) | -0.29 | 0.26 | -1.11 | 0.269 |
| Gender \times Emotion (Joy) | -0.18 | 0.53 | -0.34 | 0.738 |
| Gender \times Emotion (Pride) | -0.46 | 0.53 | -0.88 | 0.383 |
| Gender \times Emotion (Anger) | -2.05 | 0.53 | -3.90 | < .001 |
| Gender \times Emotion (Tenderness) | -1.61 | 0.53 | -3.07 | < .010 |
| Gender \times Emotion (Relief) | -1.35 | 0.53 | -2.56 | < .050 |
| Gender \times Emotion (Sadness) | -1.33 | 0.53 | -2.53 | < .050 |
| Gender \times Emotion (Irritation) | -0.88 | 0.53 | -1.67 | 0.097 |
| F2 | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | | 0.06 | | |
| Residual | | 0.22 | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 12.43 | 0.14 | 86.17 | < .001 |
| Gender | 0.50 | 0.19 | 2.64 | < .050 |
| Emotion (Joy) | -0.33 | 0.16 | -2.02 | < .050 |
| Emotion (Pride) | -0.42 | 0.16 | -2.58 | < .050 |
| Emotion (Anger) | -0.48 | 0.16 | -2.92 | < .010 |
| Emotion (Tenderness) | -0.64 | 0.16 | -3.90 | < .001 |
| Emotion (Relief) | -0.35 | 0.16 | -2.13 | < .050 |
| Emotion (Sadness) | -0.29 | 0.16 | -1.77 | 0.079 |
| Emotion (Irritation) | -0.54 | 0.16 | -3.26 | < .010 |

| F3 | | | | |
|-------------------------------|---------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | | 0.02 | | |
| Residual | | 0.09 | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 15.84 | 0.09 | 177.53 | < .001 |
| Gender | 0.26 | 0.18 | 1.46 | 0.155 |
| Emotion (Joy) | -0.22 | 0.10 | -2.07 | < .050 |
| Emotion (Pride) | -0.27 | 0.10 | -2.58 | < .050 |
| Emotion (Anger) | -0.22 | 0.10 | -2.10 | < .050 |
| Emotion (Tenderness) | -0.46 | 0.10 | -4.44 | < .001 |
| Emotion (Relief) | -0.27 | 0.10 | -2.55 | < .050 |
| Emotion (Sadness) | -0.28 | 0.10 | -2.72 | < .010 |
| Emotion (Irritation) | -0.18 | 0.10 | -1.73 | 0.087 |
| Gender × Emotion (Joy) | -0.29 | 0.21 | -1.39 | 0.167 |
| Gender × Emotion (Pride) | 0.03 | 0.21 | 0.13 | 0.895 |
| Gender × Emotion (Anger) | -0.53 | 0.21 | -2.56 | < .050 |
| Gender × Emotion (Tenderness) | -0.67 | 0.21 | -3.22 | < .010 |
| Gender × Emotion (Relief) | -0.23 | 0.21 | -1.10 | 0.273 |
| Gender × Emotion (Sadness) | -0.39 | 0.21 | -1.85 | 0.067 |
| Gender × Emotion (Irritation) | -0.20 | 0.21 | -0.94 | 0.352 |
| AR | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | | 0.46 | | |
| Residual | | 0.89 | | |
| Fixed effects | β | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 7.14 | 0.34 | 21.21 | < .001 |
| Gender | 1.21 | 0.51 | 2.38 | 0.055 |
| Emotion (Joy) | -0.71 | 0.33 | -2.12 | < .050 |
| Emotion (Pride) | -1.30 | 0.33 | -3.88 | < .001 |
| Emotion (Anger) | -1.26 | 0.33 | -3.78 | < .001 |
| Emotion (Tenderness) | -1.93 | 0.33 | -5.79 | < .001 |
| Emotion (Relief) | -1.57 | 0.33 | -4.70 | < .001 |
| Emotion (Sadness) | -1.85 | 0.33 | -5.52 | < .001 |
| Emotion (Irritation) | -1.49 | 0.33 | -4.45 | < .001 |

| PPQ | | | | |
|-------------------------------|----------|-----------------|----------|----------|
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | < .002 | | | |
| Residual | < .001 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.01 | 0.001 | 14.577 | < .001 |
| Gender | 0.004 | 0.001 | 5.266 | < .010 |
| Emotion (Joy) | -0.001 | 0.001 | -0.897 | 0.371 |
| Emotion (Pride) | -0.01 | 0.001 | -3.673 | < .001 |
| Emotion (Anger) | -0.01 | 0.001 | -3.835 | < .001 |
| Emotion (Tenderness) | -0.01 | 0.001 | -3.705 | < .001 |
| Emotion (Relief) | -0.002 | 0.001 | -1.802 | 0.074 |
| Emotion (Sadness) | -0.002 | 0.001 | -1.386 | 0.168 |
| Emotion (Irritation) | -0.003 | 0.001 | -2.514 | < .050 |
| APQ | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | < .001 | | | |
| Residual | < 0.002 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 0.06 | 0.004 | 15.33 | < .001 |
| Gender | 0.03 | 0.01 | 4.30 | < .001 |
| Emotion (Joy) | 0.004 | 0.005 | 0.74 | 0.460 |
| Emotion (Pride) | -0.003 | 0.005 | -0.71 | 0.481 |
| Emotion (Anger) | -0.005 | 0.005 | -0.99 | 0.326 |
| Emotion (Tenderness) | -0.01 | 0.005 | -2.84 | < .010 |
| Emotion (Relief) | -0.01 | 0.005 | -1.31 | 0.193 |
| Emotion (Sadness) | -0.01 | 0.005 | -1.18 | 0.240 |
| Emotion (Irritation) | -0.004 | 0.005 | -0.78 | 0.438 |
| Gender × Emotion (Joy) | 0.01 | 0.01 | 1.00 | 0.318 |
| Gender × Emotion (Pride) | -0.02 | 0.01 | -1.84 | 0.068 |
| Gender × Emotion (Anger) | -0.02 | 0.01 | -1.58 | 0.117 |
| Gender × Emotion (Tenderness) | -0.01 | 0.01 | -1.09 | 0.278 |
| Gender × Emotion (Relief) | 0.003 | 0.01 | 0.33 | 0.739 |
| Gender × Emotion (Sadness) | -0.02 | 0.01 | -1.61 | 0.111 |
| Gender × Emotion (Irritation) | -0.01 | 0.01 | -1.27 | 0.207 |
| HNR | | | | |
| Random effects | | <i>Variance</i> | | |
| Speaker (Intercept) | 1.12 | | | |
| Residual | 5.22 | | | |
| Fixed effects | <i>β</i> | <i>SE</i> | <i>t</i> | <i>p</i> |
| Intercept | 13.05 | 0.68 | 19.12 | < .001 |
| Gender | -2.52 | 0.85 | -2.96 | < .050 |
| Emotion (Joy) | -0.24 | 0.81 | -0.29 | 0.771 |
| Emotion (Pride) | 0.35 | 0.81 | 0.43 | 0.669 |
| Emotion (Anger) | -0.25 | 0.81 | -0.31 | 0.758 |
| Emotion (Tenderness) | 3.78 | 0.81 | 4.68 | < .001 |
| Emotion (Relief) | 2.64 | 0.81 | 3.27 | < .010 |
| Emotion (Sadness) | 1.61 | 0.81 | 2.00 | < .050 |
| Emotion (Irritation) | 1.08 | 0.81 | 1.34 | 0.185 |

Appendix H. Plots of LME models 1 through 15 for Hypotheses 1—3 in all data.

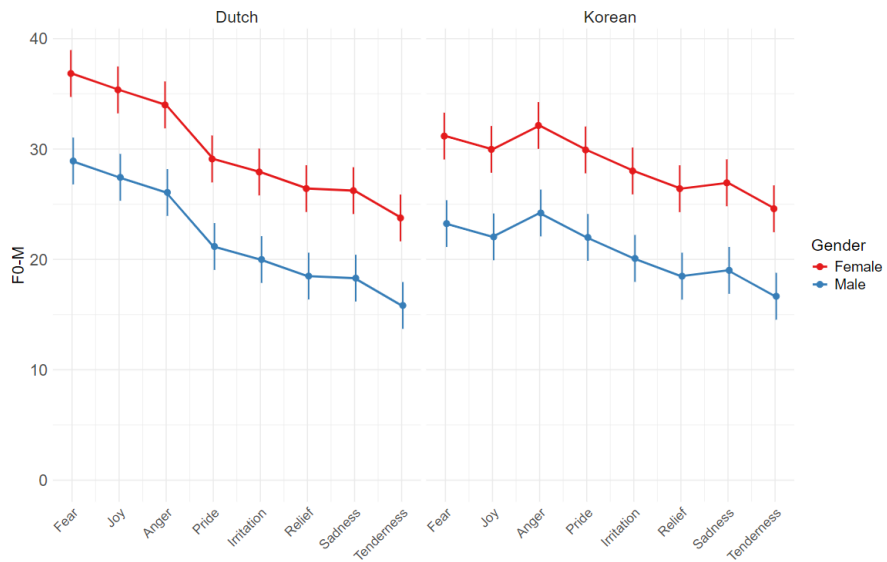


Figure H1. F0-M (semitones re. 50 Hz) between Dutch and Korean across females and males (the order of emotions from left to right is listed in a descending order of the respective acoustic parameter value in the following figures).

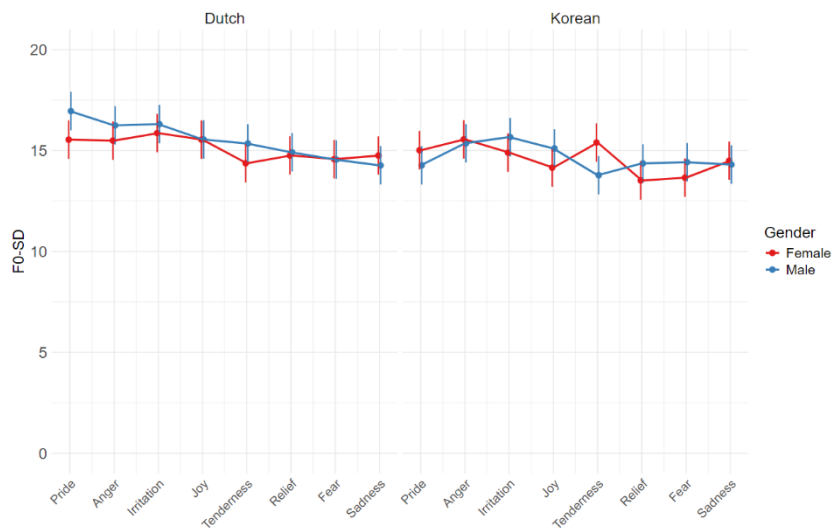


Figure H2. F0-SD (semitones) between Dutch and Korean across females and males.

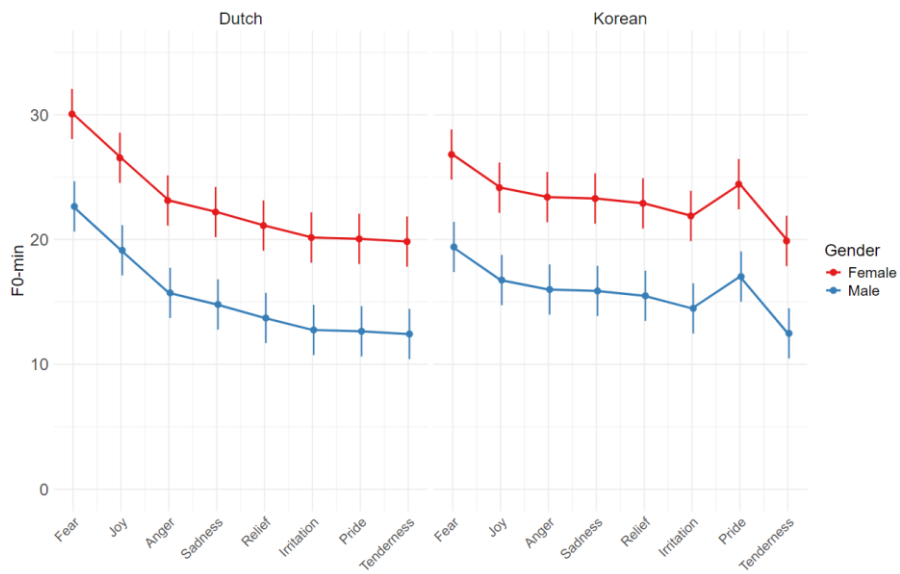


Figure H3. F0-min (semitones re. 50 Hz) between Dutch and Korean across females and males.

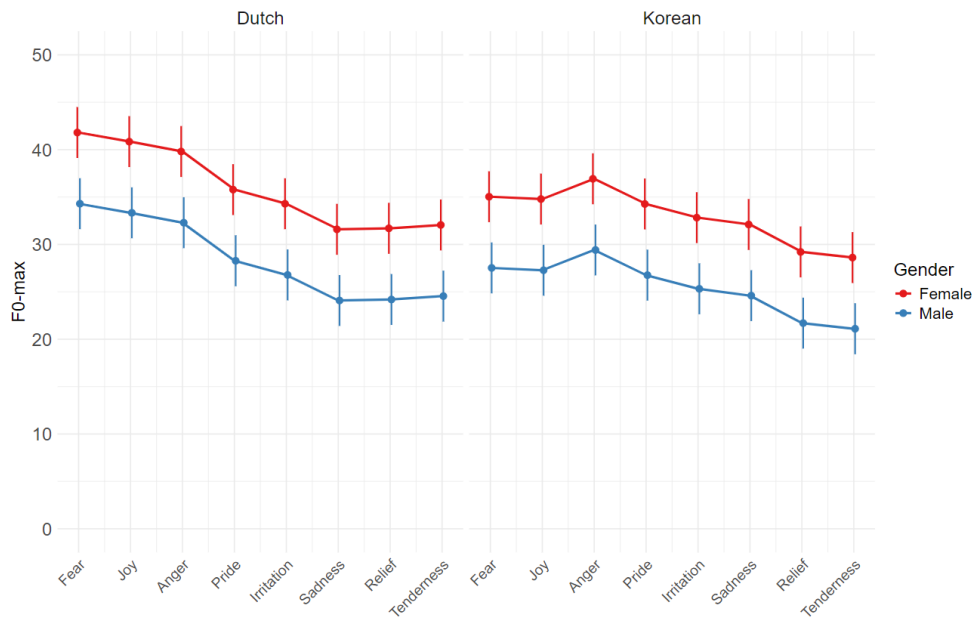


Figure H4. F0-max (semitones re. 50 Hz) between Dutch and Korean across females and males.

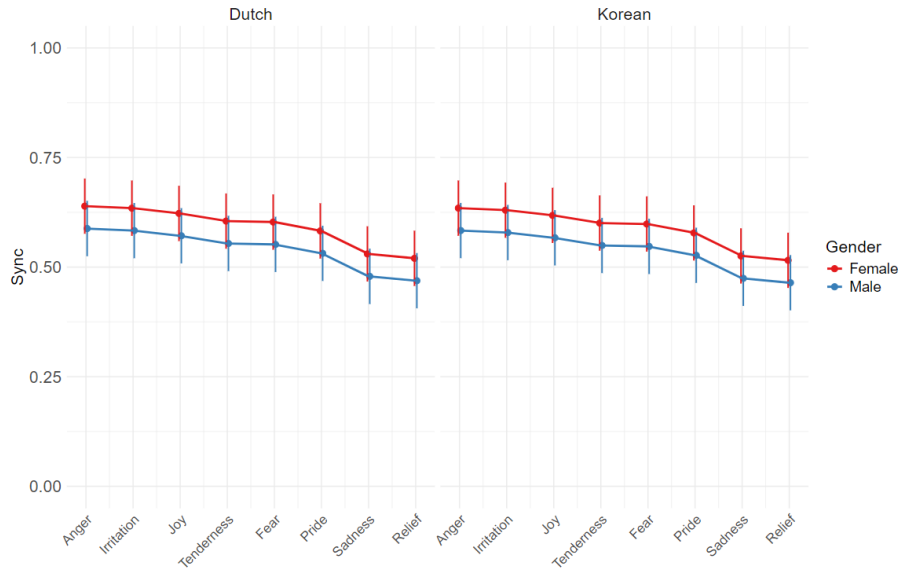


Figure H5. Synchronization (proportion of utterance duration) of F0-change between Dutch and Korean across females and males.

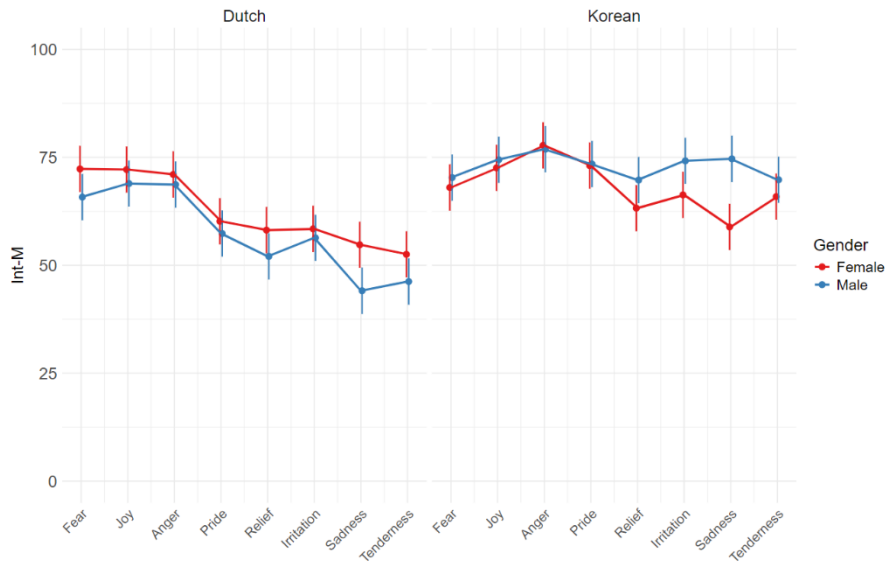


Figure H6. Int-M (decibels) between Dutch and Korean across females and males.

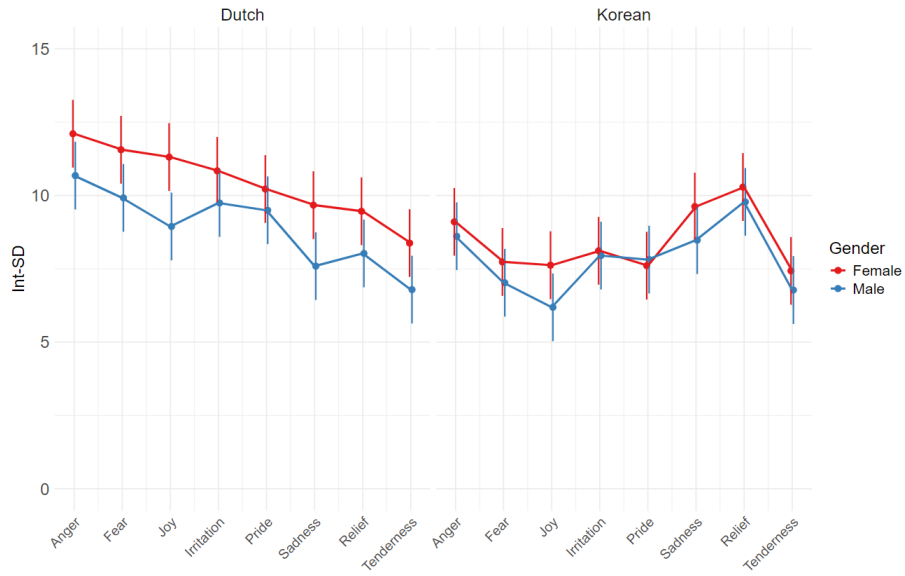


Figure H7. Int-SD (decibels) between Dutch and Korean across females and males.

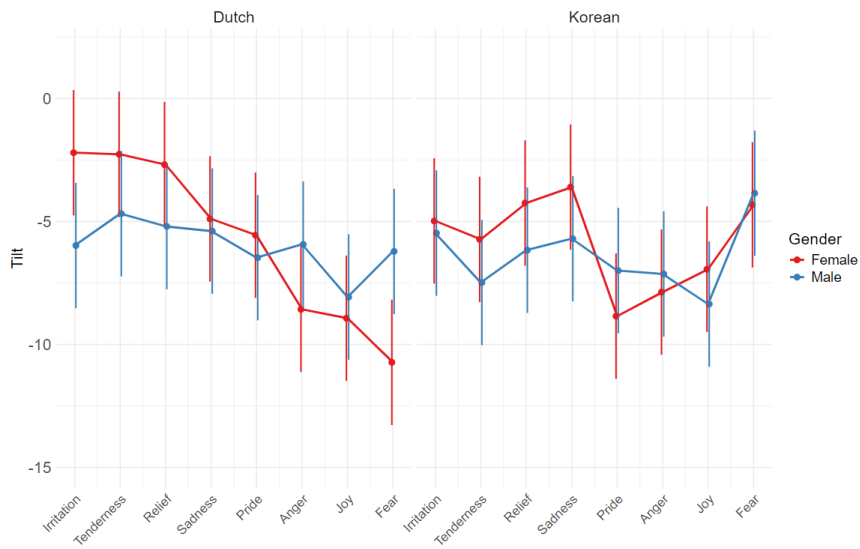


Figure H8. Spectral Tilt (decibels per octave) between Dutch and Korean across females and males.

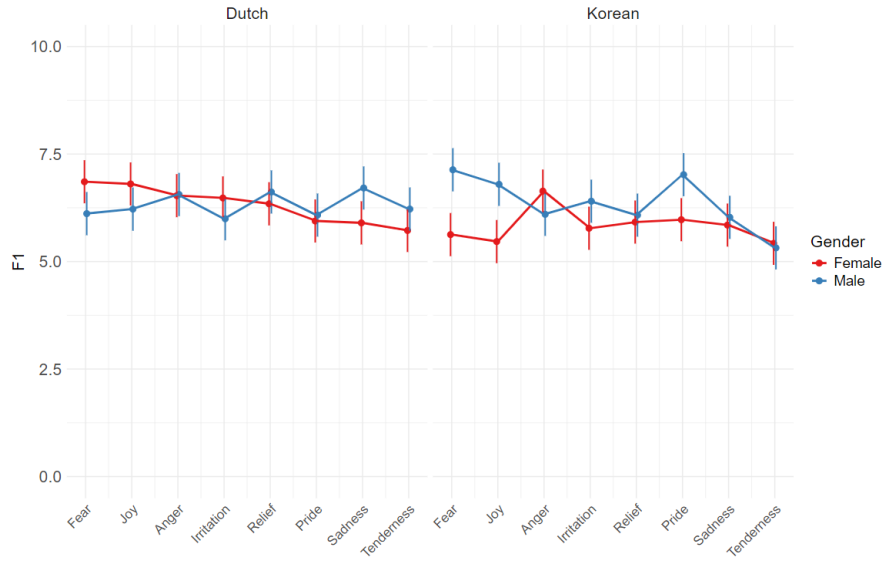


Figure H9. F1 (Equivalent Rectangular Bandwidths, ERB) between Dutch and Korean across females and males.

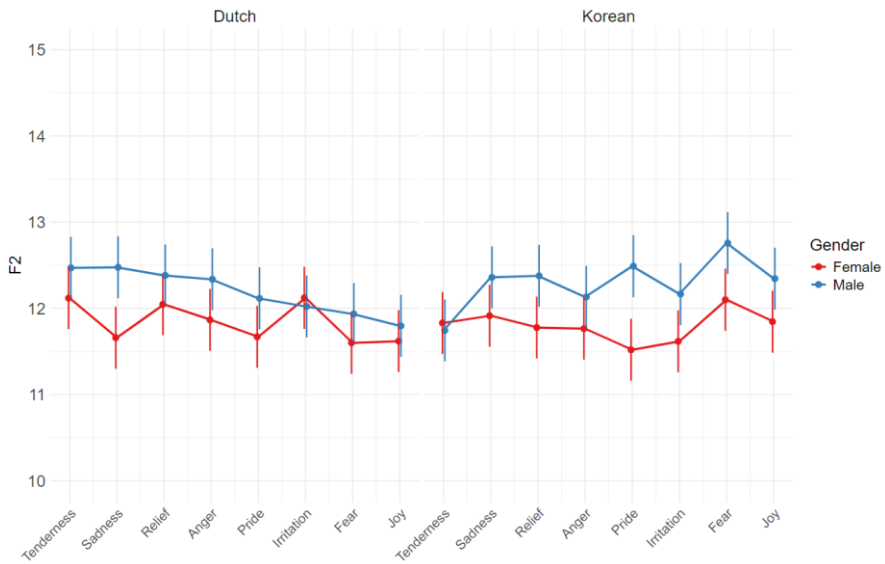


Figure H10. F2 (Equivalent Rectangular Bandwidths, ERB) between Dutch and Korean across females and males.

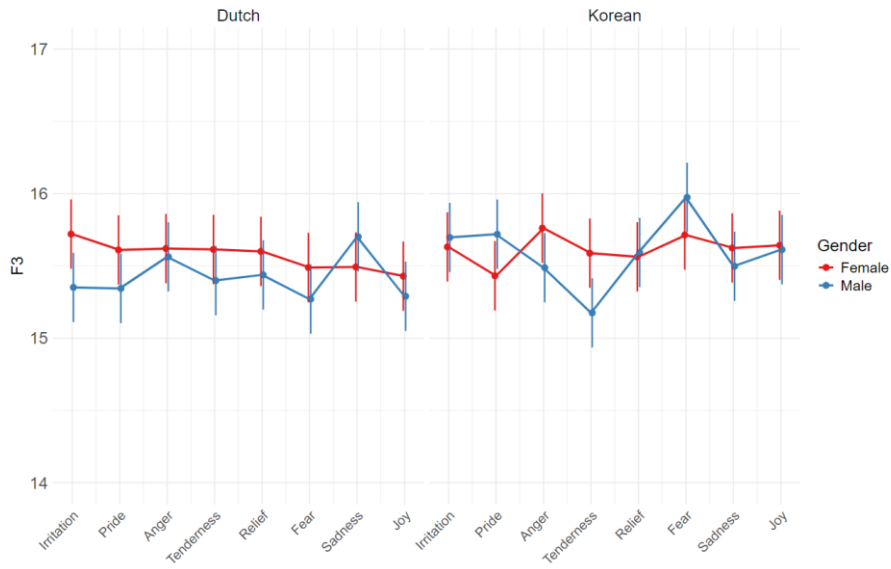


Figure H11. F3 (Equivalent Rectangular Bandwidths, ERB) between Dutch and Korean across females and males.

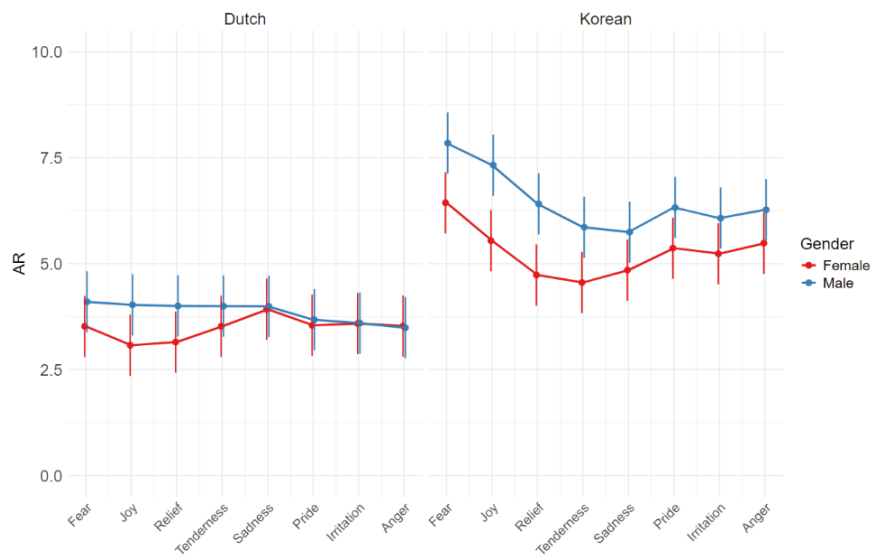


Figure H12. AR (syllables per second) between Dutch and Korean across females and males.

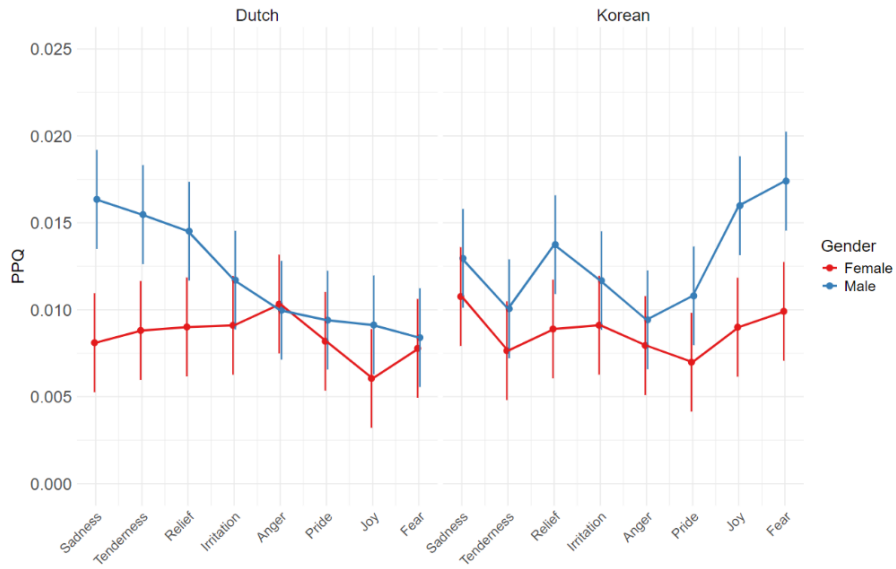


Figure H13. PPQ between Dutch and Korean across females and males.

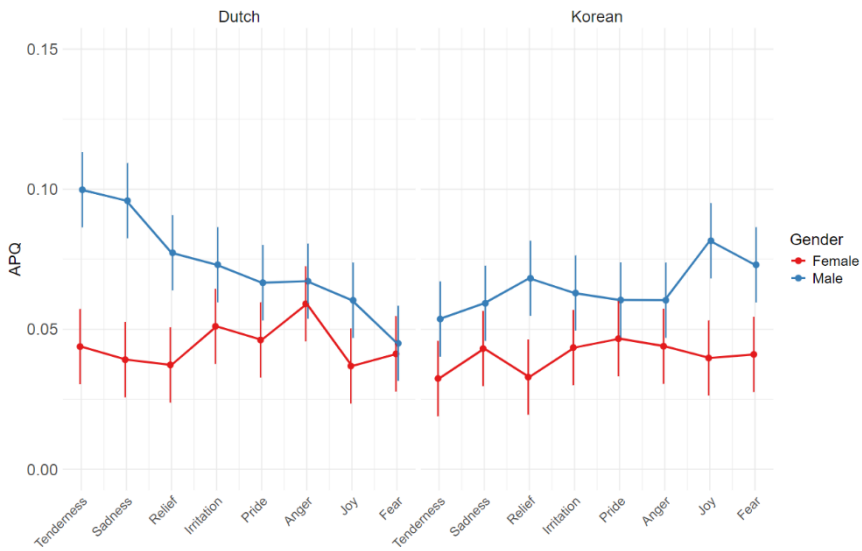


Figure H14. APQ between Dutch and Korean across females and males.

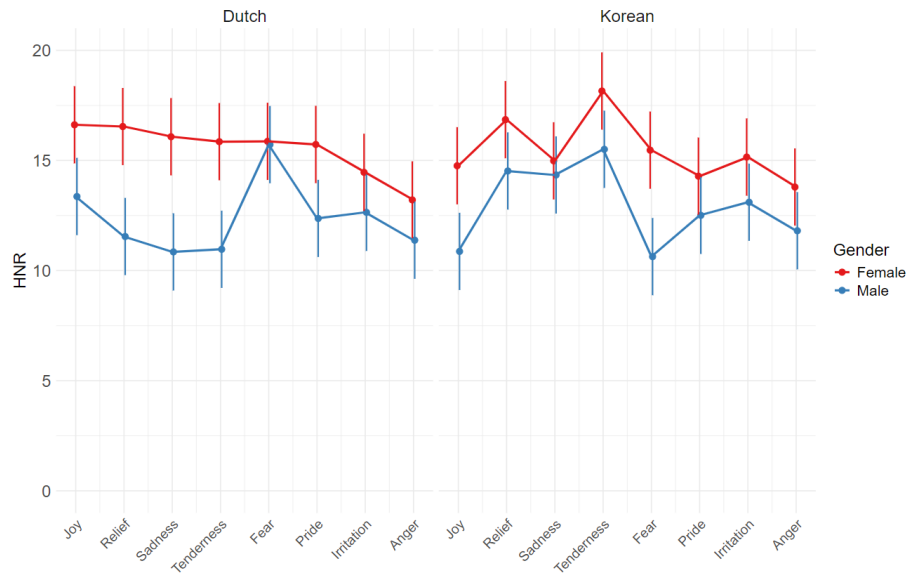


Figure H15. HNR (decibels) between Dutch and Korean across females and males.

Appendix I. Summary of confusion matrices for SVM models and classifiers.**Table II.** Confusion matrices for SVM models tested in-group (A, C) with Leave-One-Out cross-validation) and out-group (B, D). Panels E-F-G-H show the corresponding responses by human listeners. Correct classifications (%) are shown in bold in main diagonal. Empty cells have zero confusions.

| Responses (down) | In-group | | | | | | | | Out-group | | | | | | | |
|---------------------|----------------------------------|------------|------------|-----------|-----------|-----------|-----------|------------|---------------------------------|-----------|------------|-----------|-----------|-----------|-----------|------------|
| | A. Dutch model | | | | | | | | B. Dutch model, Korean data | | | | | | | |
| | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness |
| Anger | 100 | | 13 | 25 | | | | | 69 | 25 | 38 | 50 | 13 | 6 | 6 | |
| Fear | | 100 | | 56 | | | | | 13 | 31 | | 13 | 38 | 6 | 19 | |
| Irritation | | | 19 | | 13 | | | 6 | 6 | | 6 | | 6 | 13 | 13 | 13 |
| Joy | | | | 19 | | | | | | 6 | | 0 | | | | 6 |
| Pride | | | 50 | | 75 | 6 | 6 | | 13 | | 19 | 25 | 25 | 6 | 6 | 19 |
| Relief | | | | | | 13 | | | | | | | | 0 | | |
| Sadness | | | | | 13 | 81 | 88 | 25 | | 19 | 19 | | 13 | 56 | 38 | 50 |
| Tenderness | | | 19 | | | | 6 | 69 | | 19 | 19 | 13 | 6 | 13 | 13 | 19 |
| | C. Korean model | | | | | | | | D. Korean model, Dutch data | | | | | | | |
| Anger | 94 | 19 | 44 | 44 | 63 | 6 | 13 | | 81 | 63 | 19 | 88 | 25 | 6 | | |
| Fear | 6 | 69 | | 31 | 6 | 6 | | | 19 | 25 | 19 | | | 6 | 19 | 13 |
| Irritation | | | 0 | | | | | | | | 0 | | | | | |
| Joy | | 6 | 13 | 13 | 13 | 6 | 19 | 6 | | | 19 | 0 | 19 | 38 | 25 | 31 |
| Pride | | | | 6 | 6 | | 13 | 13 | | 6 | 13 | 13 | 13 | | 6 | |
| Relief | | | 19 | | 6 | 6 | 13 | 13 | | | 31 | | 44 | 19 | 13 | |
| Sadness | | | 6 | | 6 | 50 | 44 | | | 6 | | | | 19 | 19 | 6 |
| Tenderness | | 6 | 19 | 6 | | 25 | | 69 | | | | | | 13 | 19 | 50 |
| | E. Dutch listeners, Dutch data | | | | | | | | F. Dutch listeners, Korean data | | | | | | | |
| Anger | 63 | 8 | 19 | 11 | 7 | 1 | | | 39 | 4 | 10 | 16 | 8 | 1 | | 2 |
| Fear | | 41 | 2 | 11 | | 4 | 7 | 6 | 5 | 56 | 4 | 11 | 13 | 15 | 9 | 5 |
| Irritation | 34 | 6 | 61 | 9 | 16 | 7 | | 1 | 41 | 4 | 53 | 12 | 20 | 5 | 1 | 4 |
| Joy | | 2 | 2 | 41 | 25 | 5 | | 14 | 3 | | 3 | 22 | 19 | 4 | 1 | 19 |
| Pride | 2 | 1 | 7 | 4 | 25 | 4 | | 8 | 3 | | 4 | 4 | 8 | 3 | 1 | 17 |
| Relief | | 3 | 3 | 15 | 18 | 54 | 2 | 11 | 4 | 6 | 16 | 14 | 19 | 48 | 5 | 14 |
| Sadness | | 38 | 4 | 9 | 1 | 16 | 85 | 24 | 3 | 27 | 5 | 16 | 9 | 14 | 76 | 9 |
| Tenderness | | 1 | 2 | 1 | 7 | 9 | 5 | 34 | 1 | 2 | 5 | 4 | 3 | 11 | 7 | 31 |
| | G. Korean listeners, Korean data | | | | | | | | H. Korean listeners, Dutch data | | | | | | | |
| Anger | 67 | 13 | 39 | 17 | 15 | 7 | | 2 | 35 | 10 | 12 | 18 | 6 | 4 | 1 | |
| Fear | | 26 | 5 | 7 | 2 | 2 | 2 | 1 | 1 | 52 | 1 | 8 | 4 | 7 | 3 | |
| Irritation | 29 | 13 | 27 | 18 | 18 | 10 | 1 | 2 | 61 | 4 | 63 | 12 | 20 | 10 | 2 | 2 |
| Joy | 1 | 4 | 5 | 21 | 16 | 4 | | 10 | | | 1 | 22 | 18 | 7 | 1 | 18 |
| Pride | 3 | 3 | 8 | 15 | 23 | 3 | | 3 | 1 | 1 | 3 | 9 | 26 | 1 | | 4 |
| Relief | | 2 | 12 | 4 | 15 | 39 | 10 | 22 | 1 | 3 | 11 | 12 | 16 | 44 | 8 | 28 |
| Sadness | 1 | 38 | 2 | 14 | 3 | 30 | 86 | 43 | 1 | 31 | 8 | 17 | 8 | 24 | 83 | 8 |
| Tenderness | | 2 | 2 | 4 | 8 | 6 | 1 | 16 | | | 1 | 3 | 3 | 2 | 2 | 40 |

Table I2. Distance matrix (lower triangle only) for all pairs of combinations of Dutch (Du) and Korean (Ko) in-group and out-group identifications of emotions by SVM models and by human listeners.

| Training/Listeners | SVM models | | | | Human listeners | | | |
|-----------------------|------------|------|------|------|-----------------|------|------|----|
| | Du | Du | Ko | Ko | Du | Du | Ko | Ko |
| Test data/Speakers | Du | Ko | Ko | Du | Du | Ko | Ko | Du |
| Train: Du, Test: Du | 0 | | | | | | | |
| Train: Du, Test: Ko | 1.69 | 0 | | | | | | |
| Train: Ko, Test: Ko | 1.80 | 1.57 | 0 | | | | | |
| Train: Ko, Test: Du | 2.22 | 1.98 | 1.69 | 0 | | | | |
| Listen: Du, Speak: Du | 1.40 | 1.57 | 1.68 | 1.81 | 0 | | | |
| Listen: Du Speak: Ko | 1.50 | 1.47 | 1.61 | 1.78 | 0.54 | 0 | | |
| Listen: Ko, Speak: Ko | 1.46 | 1.57 | 1.66 | 1.81 | 0.48 | 0.50 | 0 | |
| Listen: Ko, Speak: Du | 1.46 | 1.45 | 1.61 | 1.77 | 0.57 | 0.75 | 0.63 | 0 |

Appendix J

Table J1. Hofstede's Six Cultural Dimension Scores for America, France, The Netherlands, and South Korea²⁷

| Country | PDI | IDV | MAS | UAI | LTO | IVR |
|-----------------|-----|-----|-----|-----|-----|-----|
| America | 40 | 91 | 62 | 46 | 26 | 68 |
| France | 68 | 71 | 43 | 86 | 63 | 48 |
| The Netherlands | 38 | 80 | 14 | 53 | 67 | 68 |
| South Korea | 60 | 18 | 39 | 85 | 100 | 29 |

²⁷ PDI: power distance; IDV: individualism vs. collectivism; MAS: masculinity vs. femininity; UAI: uncertainty avoidance; LTO: long-term vs. short-term orientation; and IVR: indulgence vs. restraint. <https://geerthofstede.com/country-comparison-graphs/>

Table J2. Cultural distance matrix between America, France, the Netherlands, and South Korea²⁸

| | America | France | Netherlands | South Korea |
|-------------|---------|--------|-------------|-------------|
| America | | 70.10 | 64.49 | 121.56 |
| France | | | 57.68 | 67.97 |
| Netherlands | | | | 92.67 |
| South Korea | | | | |

Appendix K. Recognition accuracy in Dutch and Korean recordings by American English and French listeners

| Emotion | American English listeners | | French listeners | |
|------------|----------------------------|-----------------|------------------|-----------------|
| | Dutch speakers | Korean speakers | Dutch speakers | Korean speakers |
| Anger | 0.54 | 0.30 | 0.72 | 0.46 |
| Fear | 0.42 | 0.51 | 0.52 | 0.57 |
| Irritation | 0.46 | 0.48 | 0.40 | 0.39 |
| Joy | 0.45 | 0.17 | 0.52 | 0.28 |
| Pride | 0.22 | 0.08 | 0.30 | 0.12 |
| Relief | 0.47 | 0.47 | 0.39 | 0.29 |
| Sadness | 0.76 | 0.74 | 0.75 | 0.68 |
| Tenderness | 0.26 | 0.24 | 0.25 | 0.37 |

²⁸ The cultural distance was calculated according to the Euclidean distance formula in six dimensions.

Appendix L

Table L1. Confusion matrices for American English listeners responding to Dutch and Korean recordings.

| Responses (down) | Dutch recordings | | | | | | | | Korean recordings | | | | | | | |
|---------------------|------------------|-----------|------------|-----------|-----------|-----------|-----------|------------|-------------------|-----------|------------|-----------|----------|-----------|-----------|------------|
| | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness |
| Anger | 54 | 6 | 15 | 9 | 5 | 2 | 0 | 2 | 30 | 2 | 5 | 10 | 5 | 1 | 0 | 0 |
| Fear | 1 | 42 | 1 | 10 | 2 | 2 | 5 | 3 | 4 | 51 | 1 | 14 | 14 | 8 | 4 | 1 |
| Irritation | 36 | 6 | 46 | 12 | 16 | 5 | 0 | 3 | 45 | 4 | 49 | 13 | 17 | 7 | 1 | 4 |
| Joy | 1 | 5 | 4 | 45 | 22 | 5 | 0 | 10 | 3 | 2 | 2 | 17 | 18 | 2 | 1 | 12 |
| Pride | 6 | 2 | 12 | 5 | 22 | 3 | 0 | 4 | 9 | 0 | 6 | 4 | 8 | 1 | 1 | 7 |
| Relief | 0 | 2 | 4 | 8 | 8 | 47 | 3 | 12 | 3 | 5 | 8 | 13 | 16 | 47 | 6 | 10 |
| Sadness | 0 | 33 | 2 | 9 | 3 | 11 | 77 | 28 | 3 | 25 | 7 | 20 | 8 | 13 | 74 | 13 |
| Tenderness | 0 | 3 | 1 | 0 | 5 | 10 | 8 | 26 | 0 | 7 | 5 | 3 | 2 | 11 | 10 | 24 |
| Neutrality | 2 | 3 | 14 | 3 | 19 | 16 | 9 | 15 | 6 | 5 | 18 | 7 | 11 | 11 | 4 | 28 |

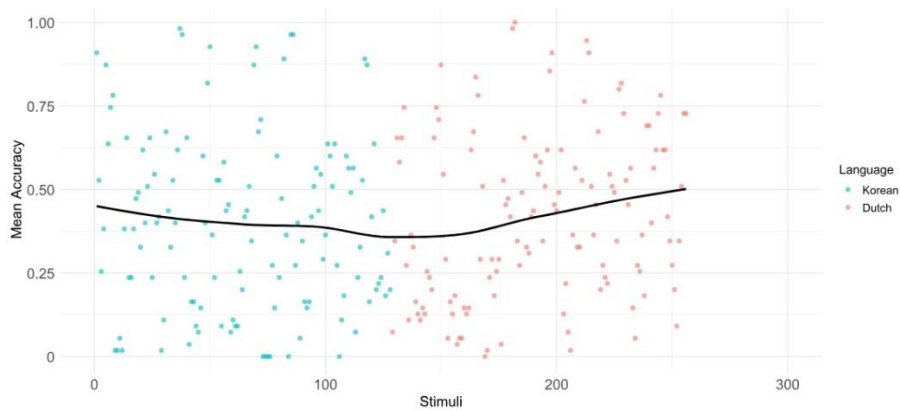
Table L2. Confusion matrices for French listeners responding to Dutch and Korean recordings.

| Responses (down) | Dutch recordings | | | | | | | | Korean recordings | | | | | | | |
|---------------------|------------------|-----------|------------|-----------|-----------|-----------|-----------|------------|-------------------|-----------|------------|-----------|-----------|-----------|-----------|------------|
| | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness |
| Anger | 73 | 8 | 25 | 10 | 5 | 3 | 1 | 1 | 46 | 5 | 14 | 13 | 6 | 2 | 0 | 0 |
| Fear | 0 | 52 | 1 | 9 | 0 | 2 | 8 | 2 | 4 | 57 | 5 | 15 | 11 | 12 | 8 | 0 |
| Irritation | 17 | 12 | 40 | 11 | 8 | 9 | 3 | 1 | 29 | 5 | 39 | 4 | 13 | 7 | 2 | 3 |
| Joy | 2 | 3 | 4 | 52 | 31 | 11 | 0 | 21 | 4 | 2 | 4 | 28 | 24 | 7 | 2 | 16 |
| Pride | 6 | 1 | 10 | 1 | 30 | 6 | 0 | 9 | 8 | 1 | 6 | 10 | 12 | 5 | 0 | 9 |
| Relief | 0 | 3 | 2 | 3 | 7 | 39 | 3 | 9 | 2 | 5 | 8 | 8 | 15 | 29 | 6 | 9 |
| Sadness | 0 | 21 | 1 | 4 | 0 | 9 | 75 | 18 | 0 | 19 | 5 | 15 | 6 | 19 | 68 | 4 |
| Tenderness | 0 | 0 | 2 | 0 | 6 | 10 | 3 | 25 | 1 | 2 | 3 | 2 | 3 | 7 | 9 | 37 |
| Neutrality | 1 | 1 | 14 | 2 | 12 | 11 | 7 | 14 | 5 | 4 | 16 | 4 | 11 | 13 | 5 | 21 |

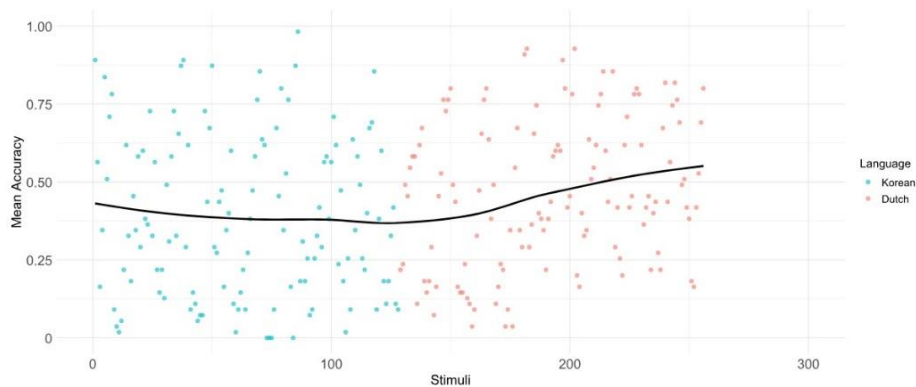
Appendix M. Distance matrix for vocal emotions of Dutch and Korean listeners (in-group only).

| | | A. Dutch listeners | | | | | | |
|------------|-------|---------------------|------------|------|-------|--------|---------|------------|
| | Anger | Fear | Irritation | Joy | Pride | Relief | Sadness | Tenderness |
| Anger | 0 | | | | | | | |
| Fear | 2.26 | 0 | | | | | | |
| Irritation | 1.35 | 2.04 | 0 | | | | | |
| Joy | 2.03 | 1.72 | 1.81 | 0 | | | | |
| Pride | 2.03 | 2.04 | 1.59 | 1.07 | 0 | | | |
| Relief | 2.37 | 1.98 | 2.03 | 1.61 | 1.50 | 0 | | |
| Sadness | 2.61 | 1.54 | 2.36 | 2.11 | 2.27 | 2.10 | 0 | |
| Tenderness | 2.40 | 1.87 | 2.09 | 1.60 | 1.45 | 1.60 | 1.82 | 0 |
| | | B. Korean listeners | | | | | | |
| Anger | 0 | | | | | | | |
| Fear | 2.25 | 0 | | | | | | |
| Irritation | 0.76 | 2.14 | 0 | | | | | |
| Joy | 1.53 | 1.71 | 1.40 | 0 | | | | |
| Pride | 1.79 | 2.07 | 1.48 | 0.85 | 0 | | | |
| Relief | 1.91 | 1.83 | 1.52 | 1.14 | 1.37 | 0 | | |
| Sadness | 2.38 | 1.93 | 2.08 | 1.70 | 2.04 | 1.58 | 0 | |
| Tenderness | 2.51 | 2.56 | 2.24 | 1.72 | 1.84 | 1.71 | 2.37 | 0 |

Appendix N. Mean Accuracy for Dutch and Korean listeners, with a smooth LOESS line to show the pattern over time (from stimulus 1 to stimulus 256).



Appendix O. Mean Accuracy for American English and French listeners, with a smooth LOESS line to show the pattern over time (from stimulus 1 to stimulus 256).



Appendix P. Instruction

P1. Dutch version

Instructie (a)

Het experiment van vandaag bestaat uit twee delen. Nu zullen we het eerste deel van het experiment uitleggen.

Je luistert naar de geluiden en bepaalt welke emotie de geluiden volgens jou uitdrukken.

De geluiden die je uit een headset hoort, zijn betekenisloos, maar verschillende Koreaanse sprekers hebben verschillende emoties geuit. Het is jouw taak om naar één geluid tegelijk te luisteren en uit voorbeelden op een scherm te kiezen welke emotie bij elk geluid hoort. Het voorbeeld op een scherm heeft 8 emoties. ① bang/angstig ② boos ③ geïrriteerd/geërgerd ④ verdrietig ⑤ tederheid ⑥ opgelucht ⑦ trots ⑧ gelukkig/blij. Als je denkt dat de zin zonder enige emotie wordt uitgedrukt, kies je voor de [neutrale] knop.

Terwijl je een soort van emoties bepaalt, moet je de intensiteit van de emotie bepalen. Met andere woorden, je moet beslissen hoe sterk de Koreaanse spreker de emotie voelt. Omdat sterke emotie op een rustige en introverte manier kan worden uitgedrukt, moet je je niet concentreren op de manier van expressie, maar op de intensiteit van de emotie zelf. Elk van de 8 emoties uit het voorbeeld op een scherm vormen vier cirkels van verschillende grootte.

De grootte van de cirkels staat voor de intensiteit van de emotie, wat betekent dat de grote cirkel staat voor sterke intensiteit en de kleine cirkel voor zwakke intensiteit. Klik op een cirkel op basis van het soort en de intensiteit van de emotie.

Als je het nodig vindt, kun je maximaal twee verschillende emoties voor één geluid kiezen. Houd ook rekening met de intensiteit van elke soort emotie. Als u op dezelfde cirkel klikt waarop u zojuist hebt geklikt, wordt deze keuze geannuleerd. Als u nog een keer wilt horen, drukt u nogmaals op de knop [Nogmaals]. Let er echter op dat u meer aandacht besteedt aan de eerste indruk die u krijgt. Druk na uw keuze op de [volgende]-knop om de volgende proefperiode te starten.

Nu gaan we oefenproeven starten. Stel al uw vragen tijdens de oefentrials. Na de oefentrials begint het eigenlijke experiment.

Instructie (b)

Laten we nu het tweede deel van het experiment uitleggen.

Zoals je in het eerste deel hebt gedaan, luister je naar de geluiden en beslis je welke emotie de geluiden uitdrukken.

Je hoort weer dezelfde zinloze zin met de zin die je in het eerste deel hebt gehoord. In deze tijd hebben verschillende Nederlandstaligen echter uiteenlopende emoties geuit. Jouw taak is om naar één geluid tegelijk te luisteren en uit voorbeelden op een scherm te kiezen welke emotie bij elk geluid hoort. Het voorbeeld op een scherm heeft 8 emoties. ①bang/angstig ②boos ③geïrriteerd/geërgerd ④verdrietig ⑤tederheid ⑥opgelucht ⑦trots ⑧gelukkig/blij. Als je denkt dat de zin zonder enige emotie wordt uitgedrukt, kies je voor de [neutrale] knop.

Terwijl je een soort van emoties bepaalt, moet je de intensiteit van de emotie bepalen. Met andere woorden, u moet beslissen hoe sterk de Nederlandstalige de emotie voelt. Omdat sterke emotie op een rustige en introverte manier kan worden uitgedrukt, moet je je niet concentreren op de manier van expressie, maar op de intensiteit van de emotie zelf. Elk van de 8 emoties uit het voorbeeld op een scherm vormen vier cirkels van verschillende grootte. De grootte van de cirkels staat voor de intensiteit van de emotie, wat betekent dat de grote cirkel staat voor sterke intensiteit en de kleine cirkel voor zwakke intensiteit. Klik op een cirkel op basis van het soort en de intensiteit van de emotie.

Als je het nodig vindt, kun je maximaal twee verschillende emoties voor één geluid kiezen. Houd ook rekening met de intensiteit van elke soort emotie. Als u op dezelfde cirkel klikt waarop u zojuist hebt geklikt, wordt deze keuze geannuleerd. Als u nog een keer wilt horen, drukt u nogmaals op de knop [Nogmaals]. Let er echter op dat u meer aandacht besteedt aan de eerste indruk die u krijgt. Druk na uw keuze op de [volgende]-knop om de volgende proefperiode te starten.

Nu gaan we oefenproeven starten. Stel al uw vragen tijdens de oefentrials. Na de oefentrials begint het eigenlijke experiment en ben je klaar.

P2. Korean version

지시 (a)

오늘의 실험은 두 부분으로 구성됩니다. 이제 실험의 첫 부분에 대해 설명하겠습니다.

소리를 듣고 소리가 표현하는 감정을 결정합니다.

헤드셋에서 들리는 소리는 의미가 없지만 여러 한국어 사용자가 다양한 감정을 표현했습니다. 당신의 임무는 한 번에 하나의 소리를 듣고 화면의 예에서 각 소리가 포함하는 감정을 선택하는 것입니다. 화면의 예에는 8 가지 감정이 있습니다. ①두렵다/두려워하다 ②화가 나다 ③화나다/성가시다 ④슬픈 ⑤부드러움 ⑥안심하다 ⑦자랑스럽다 ⑧기쁨/기쁨. 문장이 감정 없이 표현되었다고 생각되면 [중립] 버튼을 선택합니다.

감정의 종류를 결정하는 동안 감정의 강도를 결정해야 합니다. 즉, 한국어 화자가 감정을 얼마나 강하게 느끼는지를 결정해야 합니다. 강한 감정은 조용하고 내향적인 방식으로 표현할 수 있기 때문에 표현 방식이 아닌 감정의 강도 자체에 집중해야 합니다. 화면의 예에서 8 개의 감정 각각은 4 개의 다른 크기의 원을 구성합니다. 원의 크기는 감정의 강도를 나타내며, 큰 원은 강한 강도, 작은 원은 약한 강도를 나타냅니다. 감정의 종류와 정도에 따라 동그라미를 눌러주세요.

필요하다고 느끼시면 하나의 소리에 대해 최대 두 가지 감정을 선택할 수 있습니다. 각 감정의 강도를 고려하십시오. 방금 클릭한 동일한 원을 클릭하면 이 선택이 취소됩니다. 다시 듣고 싶으시면 [다시] 버튼을

눌러주세요. 하지만 첫인상을 더 중요하게 생각하시기 바랍니다. 다음 시도를 시작하려면 선택 후 [다음] 버튼을 눌러주세요.

이제 실습을 시작하겠습니다. 실습 중 궁금한 사항은 무엇이든 물어보세요. 실습이 끝나면 실제 실습이 시작됩니다.

지시 (b)

이제 실험의 두 번째 부분에 대해 설명하겠습니다.

첫 번째 부분에서 했던 것처럼 소리를 듣고 소리가 표현하는 감정을 결정합니다.

처음 부분에서 들은 문장과 똑같은 의미 없는 문장을 다시 듣게 됩니다. 그러나 이때 여러 네덜란드 화자들이 다양한 감정을 표현했습니다. 당신의 임무는 한 번에 하나의 소리를 듣고 화면의 예에서 각 소리가 포함하는 감정을 선택하는 것입니다. 화면의 예에는 8 가지 감정이 있습니다. ①두렵다/두려워하다 ②화가 나다 ③화나다/성가시다 ④슬픈 ⑤부드러움 ⑥안심하다 ⑦자랑스럽다 ⑧기쁨/기쁨. 문장이 감정 없이 표현되었다고 생각되면 [중립] 버튼을 선택합니다.

감정의 종류를 결정하는 동안 감정의 강도를 결정해야 합니다. 즉, 네덜란드 화자가 감정을 얼마나 강하게 느끼는지 결정해야 합니다. 강한 감정은 조용하고 내성적인 방식으로 표현할 수 있기 때문에 표현 방식이 아니라 감정 자체의 강도에 집중해야 합니다. 화면의 예에서 8 개의 감정 각각은 4 개의 다른 크기의 원을 구성합니다. 원의 크기는 감정의 강도를 나타내며, 큰 원은 강한 강도, 작은 원은 약한 강도를 나타냅니다. 감정의 종류와 정도에 따라 동그라미를 눌러주세요.

필요하다고 느끼시면 하나의 소리에 대해 최대 두 가지 감정을 선택할 수 있습니다. 각 감정의 강도를 고려하십시오. 방금 클릭한 동일한 원을 클릭하면 이 선택이 취소됩니다. 다시 듣고 싶으시면 [다시] 버튼을 눌러주세요. 하지만 첫인상을 더욱 중요하게 생각하시기 바랍니다. 다음 시도를 시작하려면 선택 후 [다음] 버튼을 눌러주세요.

이제 실습을 시작하겠습니다. 실습 중 궁금한 사항은 무엇이든 물어보세요. 연습 시도 후 실제 실험이 시작되고 완료됩니다.

Curriculum Vitae

Yachan Liang completed her Bachelor's degree in Chinese Language and Literature at Guangdong University of Finance in 2010 (PR China), where she developed a strong interest in language, communication, and culture. She then pursued her Master's studies in Linguistics at Utrecht University, focusing on cross-linguistic and cross-cultural aspects of language and communication. After obtaining her Master's degree in 2018, she began her doctoral research at Radboud University and later completed her PhD at Leiden University, focusing on cross-cultural emotion recognition and communication.

List of publications

Liang, Y., van Heuven, V., van Hout, R. (Submitted). Recognizing vocal emotions in unfamiliar languages: universal patterns versus cultural and prosodic proximity.

Liang, Y., van Heuven, V., van Hout, R. (Submitted). Classifying emotions cross-linguistically from acoustic parameters.

Liang, Y., van Hout, R., van Heuven, V. (Submitted). Interpreting the intensity of vocal emotions across cultures.

Liang, Y., Goudbeek, M., Konopka, A., & Choi, J., Broersma, M. (2025). Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design. *Language and Speech*. <https://doi.org/10.1177/00238309251318730>

Liang, Y., Broersma, M., Goudbeek, M., Konopka, A., & Choi, J. (2023). Rhythmic similarity hypothesis for cross-cultural vocal emotion recognition. *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, 1315–1319. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/253.pdf