Original Article

Language *and* Speech

Investigating Cross-Cultural Vocal Emotion Recognition With an Affectively and Linguistically Balanced Design

Language *and* Speech
1–27

DOI: 10.1177/00238309251318730
journals.sagepub.com/home/las
Sage

Yachan Liang
Centre for Language Studies, Radboud University Nijmegen, The Netherlands

Martijn Goudbeek
Department of Communication and Cognition, Tilburg University, The Netherlands

Agnieszka Konopka
University of Aberdeen, UK

Jiyoun Choi
Department of Social Psychology, Sookmyung Women's University, Korea

Mirjam Broersma
Centre for Language Studies, Radboud University Nijmegen, The Netherlands

Abstract

This study investigates cross-cultural vocal emotion recognition in a corpus with an affectively and linguistically balanced design. It has two main goals, one theoretical and the other methodological. 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, and communicated in a single pseudo 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 for each one which emotion they thought it expressed. Both groups of listeners recognized emotions

Corresponding author:
Jiyoun Choi, Department of Social Psychology, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 04310, Korea.
Email: jiyoun.choi@sookmyung.ac.kr

significantly above chance in both languages, but more accurately in their native language, in line with the Dialect Theory of emotion. Low-arousal emotions, negative emotions, and basic emotions were recognized more accurately than their counterparts. While some of these results replicate earlier findings, others—the effect of arousal and the within-cultural effect of valence and basicness—had not been previously investigated. This study provides new insights in cross-cultural vocal emotion recognition and contributes to the methodological toolkit of intercultural emotion recognition research.

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

Introduction

Emotions are an inseparable part of all human behavior. They guide all our actions, thoughts, and beliefs (Pessoa, 2015)—or, vice versa (with emotions being constructed from core affect, cognitive processes, and language; Feldman Barrett, 2017; Feldman Barrett & Russell, 2014; Russell, 2003). The ability to understand other people's emotions plays an important role in our daily communication and social interactions (Jensen, 2014). Revealing the dimensions along which emotions are perceived thus provides a glance of the core of human nature. Moreover, the same can be said about other species, even species as far removed from humans as domestic chickens (Marino, 2017). The study of human emotions has a long history: Charles Darwin (1872/1998) proposed that the production and perception of emotions are innate and universal, and that they have developed through evolution. Since then, emotions have been the topic of many studies. 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 conveyed emotions with their facial expressions and recognized facially expressed 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 the finding that emotions can be accurately recognized across cultures (for a meta-analysis, see Elfenbein & Ambady, 2002).

At the same time, there is overwhelming evidence that culture and language 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 Feldman Barrett & Russell, 2014). Elfenbein and Ambady (2002), in a meta-analysis of 97 studies, found that emotions were recognized with above-chance accuracies across cultures. They also found robust evidence that people who belonged to the same national, ethnic, or regional group displayed an in-group advantage, with more exposure to the other group reducing the differences between in-group and out-group performance (Elfenbein & Ambady, 2002). To account for these findings, they proposed the Dialect Theory of emotion (Elfenbein, 2013; Elfenbein & Ambady, 2002), which compares the expression of emotion to language processing: while different dialects of the same language are typically to some extent mutually intelligible, there is an in-group advantage for people sharing the same dialect. It is important to note, however, that very few studies had investigated emotion recognition in the auditory domain by 2002, such that the studies contained in Elfenbein and Ambady's (2002) meta-analysis almost exclusively addressed the visual domain.

There is now a general consensus that visual cross-cultural emotion recognition is influenced by both universal and cultural-linguistic factors (Elfenbein, 2013; Elfenbein & Ambady, 2002; 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.

1.1 Cross-cultural vocal emotion recognition

Humans express their emotions in many ways: through the face; through bodily signals such as gestures and postures; through nonlinguistic vocalizations like laughs, growls, and sighs; through the semantic content of spoken utterances; and through the paralinguistic characteristics of those utterances like prosody (Keltner et al., 2019; Mehrabian, 2017; Scherer, 2003, 2019). The vocal expression of emotion has now become a lively topic of research (e.g., Juslin & Laukka, 2003; Paulmann & Uskul, 2014; Pell et al., 2009; for reviews, see also Laukka et al., 2016, and Scherer et al., 2011). Vocal emotion expressions can be recognized cross-culturally at above-chance levels, both when they occur in nonlinguistic vocalizations (Cordaro et al., 2016; Laukka et al., 2013; Sauter et al., 2010; Sauter & Scott, 2007) and in linguistic vocalizations like phrases, words, or nonwords (Juslin & Laukka, 2003; Laukka et al., 2016; Paulmann & Uskul, 2014; Pell et al., 2009).

A meta-analysis of 37 studies of cross-cultural vocal emotion recognition (Laukka & Elfenbein, 2021) confirms that vocal emotions are recognized above chance across cultures. It also shows that there is an in-group advantage in vocal emotion recognition similar to the one found in visual emotion recognition, with listeners recognizing emotions expressed by members from the same cultural/linguistic group more accurately than those expressed by members from another group (Laukka & Elfenbein, 2021). Vocal emotion recognition is therefore, like facial emotion recognition, taken to be a product of both universal principles and language-specific factors (Juslin & Laukka, 2003; Keltner et al., 2019; Laukka & Elfenbein, 2021; Laukka et al., 2016; Mesquita & Frijda, 1992; Paulmann & Uskul, 2014; Pell et al., 2009).

Most of these findings are based on a categorical conceptualization of emotions (cf. Laukka, 2003). Emotions can also be understood, however, as entities in a multidimensional space formed by (at least) the affective dimensions arousal (or excitement) and valence (with the poles positive vs. negative, or pleasant vs. unpleasant; Laukka et al., 2005; Russell, 2003; Scherer, 2009). Arousal refers to the intensity with which an emotion is experienced. (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 positive phrases and facial expressions as being high in arousal too (Clark et al., 1984). Arousal also affects the vocal characteristics of speech. 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 with high-arousal emotions both within and across cultures.

Like arousal, valence plays an important role in emotion recognition (Russell, 1994). A number of positive and negative emotions can be identified in vocal signals (Cowen et al., 2019; Laukka & Elfenbein, 2021). Recognition accuracy is higher for negative than positive emotions

(Laukka et al., 2016; Sauter et al., 2010; Scherer et al., 2011). This trend was first observed by Sauter et al. (2010), who investigated recognition of emotion 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 across cultures, 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 that has been proposed for this difference is that negative emotions are directly associated with danger and survival, while positive emotions are linked to social bonds, and thus more likely to be shared by members from the same culture (Shiota et al., 2004). The impact of valence on emotion recognition *within* cultures, however, remains unclear.

Finally, according to Basic Emotion theory, there is a small set of emotions that are fixed physical and behavioral responses to fixed triggers that all humans share regardless of their cultural background (Ekman, 1972, 1992a; 1992b; Ekman et al., 1969; but see Gendron et al., 2018, for a different view on this matter). In line with the observations on facial expressions (Ekman, 1972; Elfenbein & Ambady, 2002), Sauter et al. (2010) found that European English and Himba listeners did reliably decode vocal expressions of basic emotions (anger, fear, joy, sadness, disgust, surprise) cross-culturally, but not those of nonbasic emotions. An open question remains as to whether basic emotions are also recognized better than nonbasic emotions *within* cultures.

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.

1.2.1 Balance in the emotions' characteristics of interest. 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 (e.g., Bailey et al., 1998; Bryant & Barrett, 2008; Chronaki et al., 2018; Chung, 1999; Huang et al., 2008; Mandal, 2008; Pell et al., 2009; Scherer et al., 2001; Thompson & Balkwill, 2006; notable exceptions being e.g., Cowen & Keltner, 2017, and Scherer et al., 2011). Other studies have used several basic emotions and only a few nonbasic emotions (e.g., Cordaro et al., 2016; Kramer, 1964; Laukka et al., 2016; Shochi et al., 2009). Most studies have included more high-arousal than low-arousal emotions (e.g., Bailey et al., 1998; Paulmann & Uskul, 2014; Pell et al., 2009; Thompson & Balkwill, 2006; see Laukka & Elfenbein, 2021), and more negative than positive emotions (see Laukka & Elfenbein, 2021, for an overview), 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 (sadness), and only one positive emotion (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 these variables by the choice of emotions included in a study.

1.2.2 Balance in the 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. While for some research questions speakers from one language and listeners from multiple languages might be desirable, other research questions require using speakers from multiple languages and listeners from a single language, or speakers and listeners from the same two language backgrounds. Some studies, using what we will call a "one-to-many" approach[1] (e.g., Beier & Zautra, 1972; Scherer et al., 2001; van Bezooijen, 1984), have presented stimuli recorded by a single group of speakers to several groups of listeners. 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 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 (e.g., many-to-many) designs provide more information than other designs. For example, Paulmann and Uskul (2014) crucially needed a two-to-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 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 use of two typologically unrelated languages, and Laukka et al.'s use of different varieties of the same language.

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, cross-cultural emotion studies that have used linguistic materials produced in two or more languages have allowed those materials to differ across languages—as is unavoidable if the materials consist of existing words or phrases. Such differences, however, introduce 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), that might affect the processing of emotional information. This therefore creates a confound between effects of culture and linguistic compatibility. Second, it is conceivable that some sounds carry more affective meaning than others (e.g., vowels vs. consonants; Majid, 2012), such that using different materials across languages entails the risk of further confounds.

Such confounds can be avoided by using pseudo-words or 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 identical in all the speakers' languages involved and to be phonologically compatible not only with the speakers' languages but also with the listeners' languages.

1.2.4 Acted versus spontaneous speech. Speech materials consist of either acted or spontaneous speech. The most important advantage of spontaneous speech is its greater ecological validity,

while the most important advantage of acted speech is the opportunity to control relevant aspects of the stimuli. 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 background noise can be produced in the laboratory, unlike in spontaneous speech. Third, acted speech can (at least aim to) express a single emotion per utterance, whereas there might be more than one dominant emotion per utterance in spontaneous speech. 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 et al., 2009; Thompson & Balkwill, 2006; van Bezooijen, 1984).

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 (Scherer et al., 2001; van Bezooijen, 1984). 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, for instance Jiang et al. (2015), have already started employing these methods.

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 (previously presented by Goudbeek & Broersma, 2010a, 2010b), adopting a two-to-two approach. Second, it aims to explore the recognition of emotions in Dutch and Korean (a language that is relatively underrepresented in affective science) with affectively and linguistically balanced materials within and across cultures.

Our first theoretical research aim concerns the recognition of emotions within and across cultures. Based on the Dialect Theory of emotion (Elfenbein & Ambady, 2002; Elfenbein et al., 2007), and the evidence reviewed above (Elfenbein, 2013; Juslin & Laukka, 2003; Laukka & Elfenbein, 2021; Pell 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 levels (Hypothesis 1), but that there will be an in-group advantage 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* and *valence*, and the *basicness* of the categorical emotions. While we have no prior expectations about the influence of arousal on emotion recognition, we test the novel hypothesis that high-arousal and low-arousal emotions will be recognized differently (Hypothesis 3). Furthermore, 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) (Hypothesis 4), and, finally, that basic emotions will be recognized more accurately than nonbasic emotions (Ekman, 1992a, 1992b, 1999; Elfenbein & Ambady, 2002) (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 in cross-cultural emotion recognition, it is crucial to have emotions balanced, as far as possible, on all these properties. In the current study, there were eight emotions (see Table 1), which are balanced in

Table 1. The eight emotions used in the current study in a valence by arousal grid.

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

Reproduced from Goudbeek and Broersma (2010a, p. 2212).
*Basic emotions.

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). 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). We adopt Ekman's classification of basic and nonbasic emotions (Ekman, 1992b, 1999; Ekman et al., 1969). 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 nonbasic emotions (pride, tenderness, relief, irritation) (see also Table 1).

Second, the study includes speakers as well as listeners from two languages: Dutch and Korean. Dutch and Korean differ strongly in their use of prosodic cues like pitch. Dutch employs pitch as a cue to signal word stress, which differentiates the meaning of segmentally identical word forms (Cutler & Van Donselaar, 2001; Gussenhoven, 1993) Pitch also contributes to the marking of two prosodic units in Dutch, namely Intonational Phrases (IP) and Phonological Phrases (PP) (Gussenhoven, 2005). In Korean, pitch contributes to the marking of two different prosodic units, namely Intonational Phrases (IP) and Accentual Phrases (AP) (Jun, 2005), the final boundaries of which are signaled by a rising pitch movement and lengthening (Jun, 2006; Kim et al., 2008).

Third, to ensure the similarity of the stimuli across the languages, we use a single pseudo phrase [nuto hɔm sɛpikaŋ], which is phonologically compatible with Dutch and Korean.

Fourth, this study uses acted speech to obtain well-controlled stimuli. The emotion portrayals from the Demo/Koremo corpus (Goudbeek & Broersma, 2010a) have been recorded following the methods developed by Scherer and colleagues (Banse & Scherer, 1996; Bänziger et al., 2012; 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 be able to statistically account for the effect of all variables of interest, including by-participant and by-item variability, we use logistic linear mixed effects models in our analyses.

2 Method

2.1 Participants

There were two groups of participants: 31 native listeners of Dutch (age: $M = 20.87$, $SD = 2.17$), all of whom were students at Radboud University Nijmegen in the Netherlands, and 24 native listeners of Korean (age: $M = 23.46$, $SD = 2.59$), all of whom were students at Korea University, in 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 judgment study that was used for the selection of the portrayals (described below).

2.2 Auditory materials

We used all the emotion portrayals from the Demo/Koremo (Dutch emotion/Korean emotion) corpus by Goudbeek and Broersma (2010a). 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. nonbasic emotions (Table 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 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 × 8 actors × 2 tokens × 2 languages). All portrayals use a single pseudo phrase [nuto hɔm sɛpikaŋ], which is phonologically legal in both Dutch and Korean. Elicitation and recording procedures were the same in Dutch and Korean.

2.2.1 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 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 college-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 Konstantin Stanislavski (1936/1988), 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 earlier work (Banse & Scherer, 1996; Bänziger et al., 2012; Bänziger & Scherer, 2007), three scenarios per emotion were provided to illustrate the emotions prior to reenactment.[2]

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 min (with 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 phrase 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 minimally five good portrayals. From those selected sequences, the final four portrayals per emotion per actor were used for the judgment 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 remaining earlier portrayals that the director had approved of.

2.2.2 Judgment study. To determine the quality and naturalness of each emotion portrayal, we conducted a judgment study (see also Goudbeek & Broersma, 2010a) 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 actors × 8 emotions × 4 tokens × 2 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[3] throughout the experiment. 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 experiment object (Boersma, 2001).

2.2.3 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 (i.e., with the highest unbiased hit rates) per actor per emotion 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 (2010a).

2.3 Visual materials

The main experiment made use of 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 (Figure 1) and a Korean version. 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 toward 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."

2.4 Procedure

Participants were tested individually in a sound-attenuated booth at Radboud University and at Korea University. Participants were seated in front of a computer screen which showed the emotion wheel in the participant's native language. Recordings were played 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 by choosing from the eight emotions on the screen, as well as the intensity with which they thought the speaker had experienced the emotion

Figure 1. The emotion wheel in Dutch: *Blijdschap* (Joy), *Trots* (Pride), *Opluchting* (Relief), *Vertedering* (Tenderness), *Verdriet* (Sadness), *Irritatie* (Irritation), *Woede* (Anger), *Angst* (Fear), *Neutraal* (Neutral); *Opnieuw* (Again), *Volgende* (Next).

or, alternatively, to choose the "Neutral" option (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). 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 min.

3 Results

The data were analyzed in R (R Core Team, 2018). We ran one-sample *t*-tests and binomial 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; glmer provides *p*-values from Wald z-tests) to address all other hypotheses. We also report pairwise comparisons (all obtained with the *emmeans* package) to further elaborate on some of the findings.

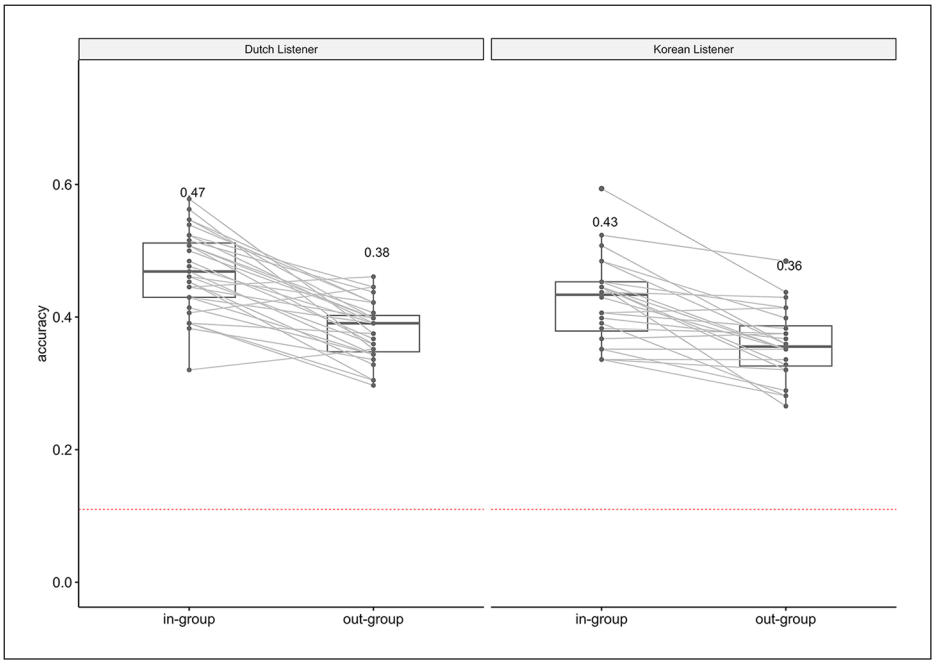

Figure 2. Recognition accuracy in the in-group condition (i.e., responding to recordings produced by speakers of the same language) and in the out-group condition (i.e., responding to recordings produced by speakers of the different language) for Dutch and Korean listeners. In all figures, the red line indicates chance performance (.11), and printed values are by-participant condition means.

The models used a combination of five predictors (fixed factors) as outlined in each analysis below: In-/out-group (listeners and speakers from the same [in-group] vs. different language groups [out-group]), Listener Language (Dutch vs. Korean listeners), Arousal (high-arousal vs. low-arousal emotions), Valence (positive vs. negative emotions), and Basicness (basic vs. nonbasic 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). Significance was determined with Wald *z*-tests (provided in the output of glmer models).

The models included the maximal random structure justified by the design 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 nonconvergence, models were simplified by iteratively removing the random slopes accounting for the smallest amount of variance (Barr et al., 2013) until convergence was reached.[4]

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. We expected above-chance performance (with chance level in a 9-alternative forced choice task, i.e., 1 out of 9, being .11) in both listener groups and both in in-group and in out-group condition. Figure 2 shows that all participants performed much higher than chance in both In-/out-group conditions (i.e., with recordings produced by speakers of the same and different language groups), in line with Hypothesis 1. This hypothesis was tested with four one-sample *t*-tests, which

compared the average recognition accuracy of (1) Dutch listeners in Dutch recordings and (2) in Korean recordings, as well as recognition accuracy of (3) Korean listeners in Dutch recordings and (4) in Korean recordings, to the chance level. Indeed, the four *t*-tests showed that performance was significantly above chance in all conditions and across all participants, all $ts > 22$, all $ps < .001$ (see Appendix A). Binomial tests further confirmed this finding, with the probabilities of observing responses above chance at $p < .001$ for each individual participant. Thus, both groups of listeners were able to recognize vocal emotion expressions above chance in their own language and also in the unknown language.

3.2 The effect of In-/out-group, Arousal and Valence in emotion recognition (Hypothesis 2, 3, and 4)

We hypothesized that listeners would recognize emotions from their own language more accurately than emotions from the other language (Hypothesis 2), that Arousal would influence recognition accuracy (Hypothesis 3), and that negative emotions would be recognized more accurately than positive emotions (Hypothesis 4).

We tested these hypotheses by assessing the effects of In-/out-group, Arousal, and Valence on emotion recognition in a joint analysis that also included Listener Language. The model included these four variables as fixed effects (testing for a four-way interaction), as well as random by-participant slopes for Arousal and Valence, and random by-item interacting slopes for In-/out-group and Listener Language. All random slopes, except slopes for In-/out-group, improved model fit significantly. Table 2 reports the model output. We first discuss the results addressing Hypothesis 2, 3, and 4 in turn, and then discuss the joint effects of all four variables.

3.2.1 Hypothesis 2 (in-group effect). Figure 2 suggests that there was an in-group effect, with listeners recognizing emotions correctly more often when produced by speakers of the same language (in-group condition), than by speakers of the other language (out-group condition). Table 2 shows that there was indeed a significant main effect of In-/out-group, supporting Hypothesis 2.

There was also a significant main effect of Listener Language, as Dutch listeners had generally higher accuracy than Korean listeners (recognition accuracy was .03 higher in Dutch listeners than Korean listeners; also see Figure 2).[5] The interaction between In-/out-group and Listener Language was not reliable; thus, the in-group effect did not reliably differ between the Dutch and Korean listener groups. However, to confirm the presence of an in-group advantage in both groups, we report on results from Dutch and Korean listeners separately (with two one-tailed Bonferroni-corrected pairwise comparisons). As Figure 2 shows, there was an in-group recognition benefit of .09 in Dutch listeners responding to Dutch over Korean recordings, and an in-group recognition benefit of .07 in Korean listeners responding to Korean over Dutch recordings (see Table 3 for further details and pairwise comparisons). Thus, both groups of listeners contributed to the in-group advantage.

Finally, there was a significant three-way interaction among In-/out-group, Arousal and Valence (see Table 2), which will be discussed below.

3.2.2 Hypothesis 3 (Arousal). Figure 3 suggests that recognition was more accurate for low-arousal emotions than for high-arousal emotions in line with Hypothesis 3 (which did not specify the direction of the expected effect). The model (Table 2) indeed showed a main effect of Arousal on emotion recognition accuracy, which was .13 higher for low-arousal than high-arousal emotions, thus confirming Hypothesis 3.

Table 2. Summary of results of the logistic mixed effects model analysis for Hypothesis 2, 3, and 4. In all tables, coefficients (β) are transformed back to odds ratios (*exp*(β)) for ease of interpretation. (We provide an interpretation for the In-/out-group effect separately for Dutch and Korean listeners in terms of differences in the odds of correct responses between the in-group and out-group conditions below the table. All other interpretations are reported in the main text).

Model 1 (Hypothesis 2, 3, 4)	Estimates					
	β	*Exp (β)*	*SE*	z value	*p* value	95% CI
Intercept	−0.57	0.56	0.09	−6.38	<.001	[−.75, −.40]
In-/out-group	0.50	0.61	0.07	6.85	<.001	[−.64, −.36]
Listener Language	−0.20	0.82	0.10	−2.02	0.043	[−.40, −.01]
Arousal	0.83	2.30	0.18	4.65	<.001	[−.48, 1.18]
Valence	−1.64	0.19	0.18	−9.09	<.001	[−1.99, −1.29]
In-/out-group × Listener Language	0.20	1.22	0.33	0.59	0.56	[−.46, .85]
In-/out-group × Arousal	−0.21	0.81	0.15	−1.41	0.16	[−.49, .08]
In-/out-group × Valence	−0.28	0.76	0.15	−1.89	0.059	[−.57, .01]
Listener Language × Arousal	0.03	1.03	0.19	0.16	0.873	[−.35, .41]
Listener Language × Valence	−0.08	0.92	0.20	−0.40	0.689	[−.47, .31]
Arousal × Valence	−0.12	0.88	0.33	−0.37	0.710	[−.78, .53]
In-/out-group × Listener Language × Arousal	−1.45	0.24	0.67	−2.17	0.030	[−2.76, −.14]
In-/out-group × Listener Language × Valence	0.58	1.79	0.67	0.87	0.384	[−.73, 1.89]
In-/out-group × Arousal × Valence	0.72	2.05	0.29	2.48	0.013	[.15, 1.29]
Listener Language × Arousal × Valence	−0.50	0.61	0.29	−1.73	0.084	[−1.07, .07]
In-/out-group × Listener Language × Arousal × Valence	1.04	2.83	1.31	.79	0.433	[−3.65, 1.56]

Notes. Model 1 showed a significant main effect of In-/out-group. In Dutch listeners, the odds of a correct response were 1.45 times higher when listening to Dutch (in-group) than Korean (out-group) recordings (i.e., .37 log odds higher when listening to Dutch than Korean recordings). In Korean listeners, the odds of a correct response were 1.32 times higher when listening to Korean (in-group) than Dutch (out-group) recordings (i.e., .28 log odds higher when listening to Korean than Dutch recordings).

Figure 3 further suggests that this benefit occurred both in the in-group condition (a benefit of .10 for Dutch listeners and .21 for Korean listeners) and in the out-group condition (a benefit of .15 for Dutch listeners and .06 for Korean listeners). Indeed, there was no significant interaction between In-/out-group and Arousal.

Finally, Figure 3 suggests that the magnitude of the effect of Arousal was relatively large for recordings produced by Korean speakers (i.e., in the in-group condition for Korean listeners and in the out-group condition for Dutch listeners). There was a reliable interaction between In-/out-group, Listener Language and Arousal (Table 2).

3.2.3 Hypothesis 4 (Valence). Figure 4 suggests that recognition was more accurate for negative emotions than for positive emotions as predicted by Hypothesis 4. Table 2 indeed shows the presence of a main effect of Valence on emotion recognition, confirming Hypothesis 4; accuracy was .26 higher for negative than positive emotions.

Table 3. Summary of the pairwise comparisons per analysis. All tests were one-tailed, reflecting the directionality of the predicted effects. Bonferroni corrections were applied for the number of tests listed in each section below (two, four, or eight tests). The column "*n* participants showing effect" indicates the number of participants showing a difference in the expected direction as a fraction of the total number of participants in that subset of the data.

	Means	Difference	*z*	*p*	*n* participants showing effect
Hypothesis 2 (in-group effect). NB There was no reliable interaction In-/out-group × Listener Language.					
Comparison of in-group vs. out-group conditions					
Dutch listeners	.47 vs. .38	.09	3.60	<.001	28/31
Korean listeners	.43 vs. .36	.07	2.04	.041	22/24
Joint effects of In-/out-group, Arousal and Valence. Exploring the significant interaction In-/out-group × Arousal × Valence.					
Comparison of in-group vs. out-group conditions					
High arousal, positive emotions	.28 vs. .17	.11	4.56	<.0001	42/55
High arousal, negative emotions	.48 vs. .46	.02	.57	*ns*	31/55
Low arousal, positive emotions	.37 vs. .29	.08	3.88	<.0001	40/55
Low arousal, negative emotions	.68 vs. 56	.12	4.48	<.0001	39/55
Comparison of high vs. low arousal emotions					
In-group, positive emotions	.28 vs. .37	.09	–2.70	.014	47/55
In-group, negative emotions	.48 vs. .68	.20	–4.64	<.0001	47/55
Out-group, positive emotions	.17 vs. .29	.12	–3.03	.005	48/55
Out-group, negative emotions	.46 vs. .56	.10	–2.26	.048	40/55
Comparison of negative vs. positive emotions					
In-group, high arousal emotions	.48 vs. .28	.20	4.90	<.0001	49/55
In-group, low arousal emotions	.68 vs. .37	.31	6.82	<.0001	53/55
Out-group, high arousal emotions	.46 vs. .17	.29	6.82	<.0001	53/55
Out-group, low arousal emotions	.56 vs. .29	.27	6.03	<.0001	52/55
Comparison of in-group vs. out-group conditions					
Dutch listeners, positive, high-arousal emotions	.32 vs. .15	.17	4.22	.0001	29/31
Dutch listeners, positive, low-arousal emotions	.38 vs. .32	.06	.96	*ns*	21/31
Korean listeners, positive, high-arousal emotions	.22 vs. .20	.02	.05	*ns*	13/24
Korean listeners, positive, low-arousal emotions	.36 vs. .25	.11	2.06	.16	19/24
Dutch listeners, negative, high-arousal emotions	.52 vs. .46	.06	.81	*ns*	19/31
Dutch listeners, negative, low-arousal emotions	.66 vs. .59	.07	1.11	*ns*	18/31
Korean listeners, negative, high-arousal emotions	.42 vs. .45	–.03	–.27	*ns*	12/24
Korean listeners, negative, low-arousal emotions	.68 vs. .53	.15	2.35	.07	21/24
Hypothesis 5 (Basicness). Exploring the significant interaction In-/out-group × Basicness.					
Comparison of basic vs. nonbasic emotions					
In-group	.52 vs. .38	.14	3.26	<.01	46/55
Out-group	.48 vs. .27	.21	5.70	<.0001	53/55

We thank Associate Editor Susannah Levi for requesting all the pairwise comparisons in this table.

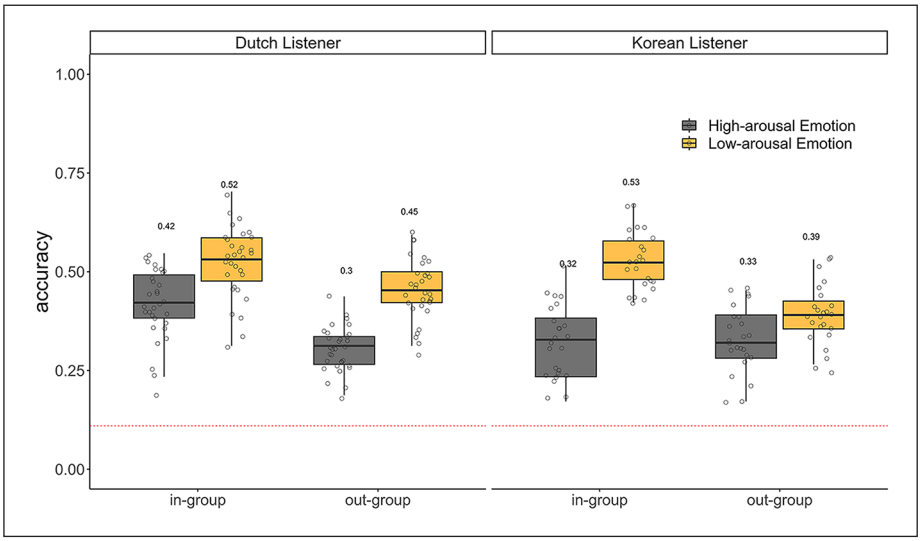

Figure 3. Recognition accuracy in the in-group and out-group conditions for Dutch and Korean listeners responding to recordings of high and low-arousal emotions.

Figure 4 suggests that the effect of Valence was grounded both in the in-group condition (a benefit of .24 for Dutch listeners and .27 for Korean listeners) and in the out-group condition (a benefit of .29 for Dutch listeners and .26 for Korean listeners). The two-way interaction between In-/out-group and Valence was marginally significant ($p = .059$), as the magnitude of the effect of Valence was relatively large in the out-group compared with the in-group condition, and the magnitude of the in-group effect was relatively large for positive compared with negative emotions.

3.2.4 Joint effects of In-/out-group, Arousal and Valence. The above effects were further qualified by a significant three-way interaction between In-/out-group, Arousal and Valence. We further assessed the effects across the relevant cells in this interaction with 12 one-tailed, Bonferroni-corrected pairwise comparisons.

Figure 5 shows that the In-/out-group effect was modulated by Arousal and Valence. The two-level variables Arousal and Valence can be combined in four ways. In each of the four combinations of Arousal and Valence, there was an in-group advantage, either statistically significant or as a nonsignificant trend (see Table 3). The magnitude of the effect of In-/out-group varied across conditions (the effect was reliable in three out of four comparisons). Figure 5 shows that the in-group advantage was relatively small for high-arousal, negative emotions.

Furthermore, the effect of Arousal was modulated by In-/out-group and Valence. In all four combinations of In-/out-group and Valence, there was a trend (or effect) of Arousal in the same direction, with more accurate recognition of low-arousal emotions than of high-arousal emotions (see Table 3). Figure 5 shows that the magnitude of this effect was relatively large in the in-group condition for negative emotions.

Finally, the effect of Valence was modulated by In-/out-group and Arousal. In all four combinations of In-/out-group and Arousal there was a significant effect of Valence, with more accurate recognition of negative than of positive emotions (see Table 3). Figure 5 shows that the magnitude of this effect was relatively large in the in-group condition for low-arousal emotions.

The four-way interaction was not reliable; thus, the joint effects of In-/out-group, Arousal and Valence (a reliable three-way interaction) were not further modulated by Listener Language.[6]

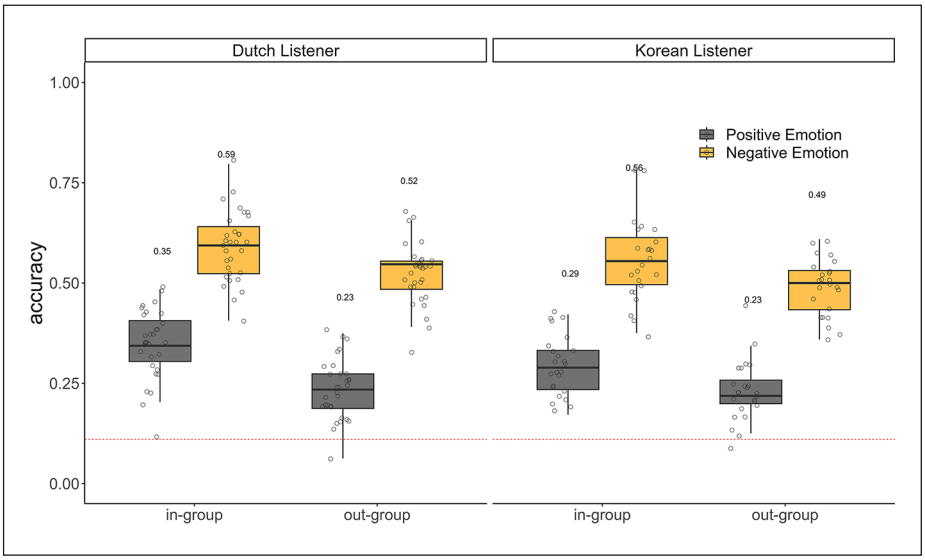

Figure 4. Recognition accuracy in the in-group and out-group conditions for Dutch and Korean listeners responding to recordings of positive and negative emotions.

The In-/out-group effect was thus observed relatively consistently in Dutch and Korean participants and contributed to across all levels of Arousal and Valence, but more consistently for positive emotions than for negative emotions.

3.3 The effect of basicness on emotion recognition (Hypothesis 5)

Next, we tested whether listeners could recognize basic and nonbasic emotions above chance, both within and across cultures. We predicted that performance would be above chance in both listener groups and both in in-group and in out-group condition. Figure 6 suggests that that was indeed the case. To test this prediction, we compared recognition accuracy of each listener group for both In-/out-group conditions, separately in basic and nonbasic emotions, to the chance level (.11) with eight one-sample *t*-tests (see Appendix B). Performance was above chance in all condition, all $ts > 8.45$, all $ps < .001$ (confirmed with binomial tests, $p < .001$ for all participants on basic emotions trials, and $p < .05$, for all but one Dutch participant and four Korean participants on nonbasic emotions trials). These results show that listeners identified basic as well as nonbasic emotions above chance within and across cultures.

Furthermore, we tested whether basic emotions were recognized more accurately than nonbasic emotions as predicted by Hypothesis 5. This hypothesis was addressed in Model 2 (see Table 4). The analysis tested for all interactions between In-/out-group, Listener Language and Basicness, and included random by-participant slopes for In-/out-group and Basicness and random interacting by-item slopes for In-/out-group and Listener Language. All random slopes, except by-item slopes for In-/out-group, improved model fit significantly.

Figure 6 suggests that there was a recognition accuracy benefit for basic over nonbasic emotions. Indeed, there was a significant main effect of Basicness: recognition accuracy was .17 higher for basic than nonbasic emotions, supporting Hypothesis 5.

As in previous models, the model also showed the expected main effect of In-/out-group. Figure 6 suggests that the effect of Basicness was found both in the in-group condition (a benefit of .18

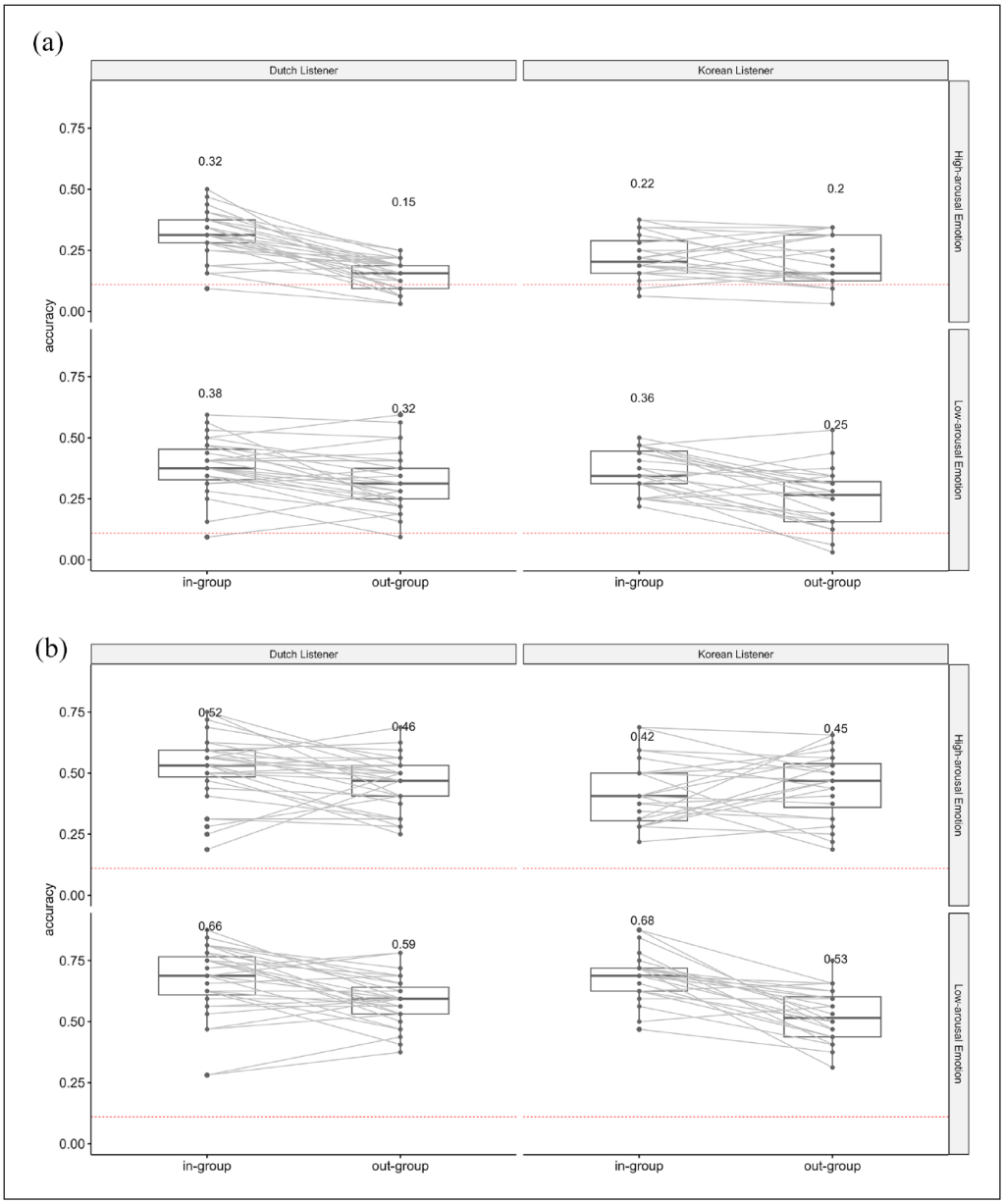

Figure 5. Recognition accuracy in the in-group and out-group conditions for Dutch and Korean listeners responding to recordings of high and low-arousal emotions, shown separately for (a) positive and (b) negative emotions.

for Dutch listeners and .07 for Korean listeners) and in the out-group condition (a benefit of .18 for Dutch listeners and .26 for Korean listeners). There was a significant interaction between In-/out-group and Basicness, as the magnitude of the in-group effect was relatively large for nonbasic compared with basic emotions (see Figure 6). The effect of Basicness was significant both in the in-group and in the out-group condition (Table 3). The three-way interaction was not reliable; thus, the joint effects of In-/out-group and Basicness was not further modulated by Listener Language.

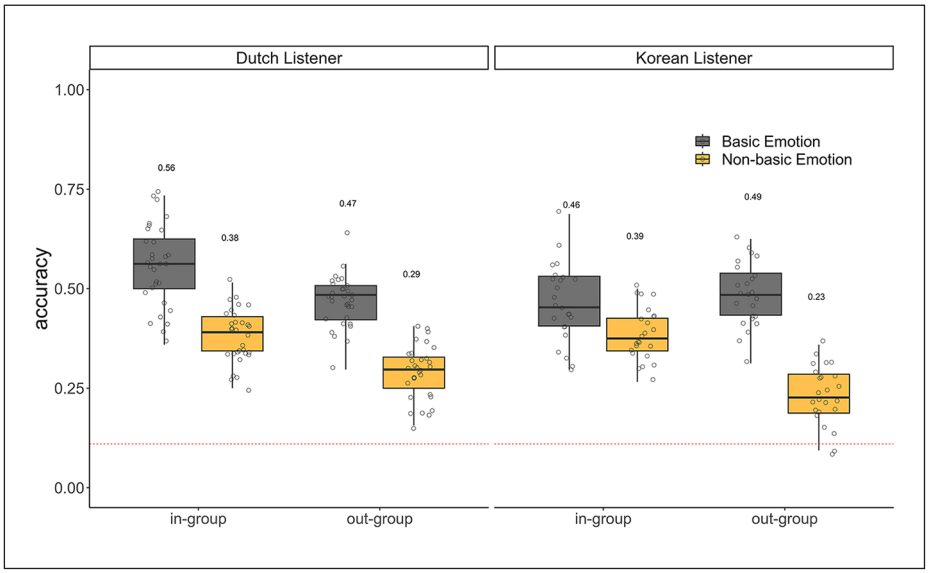

Figure 6. Recognition accuracy in the in-group and out-group conditions for Dutch and Korean listeners responding to recordings of basic and nonbasic emotions.

Table 4. Summary of results of the logistic mixed effect model analyses for Hypothesis 5.

Model 2 (Hypothesis 5)	Estimates					
	β	*Exp (β)*	*SE*	*z* value	*p* value	95% CI
Intercept	−0.56	.57	0.10	−5.50	< .001	[−.76, -.36]
In-/out-group	−0.46	.63	0.08	−6.16	< .001	[−.61, -.32]
Listener Language	−0.21	.81	0.10	−2.21	0.027	[−.40, -.02]
Basicness	−0.96	.38	0.20	−4.76	< .001	[−1.35, -.56]
In-/out-group × Listener Language	0.21	1.23	0.39	.54	0.591	[−.55, .97]
In-/out-group × Basicness	−0.53	.59	0.14	−3.67	< .001	[−.81, -.25]
Listener Language × Basicness	0.02	1.02	0.19	0.09	0.927	[−.35, .38]
In-/out-group × Listener Language × Basicness	−0.84	.43	0.77	−1.10	0.272	[−2.35, .66]

4 Discussion

This study investigated cross-cultural emotion recognition with a carefully balanced design. We replicated and extended earlier findings and provided several novel insights in vocal emotion recognition. The study had two main goals, one theoretical and the other methodological.

Our first theoretical aim (expressed in Hypotheses 1 and 2) was to test the predictions of the Dialect Theory of emotion (Elfenbein, 2013; Elfenbein & Ambady, 2002). 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. Our study has thus replicated 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

vocal emotion recognition (Laukka & Elfenbein, 2021; Laukka et al., 2016; Scherer et al., 2001). Second, as predicted in Hypothesis 2, we found an in-group advantage in cross-linguistic emotion recognition (Elfenbein & Ambady, 2002). Both groups of listeners recognized emotions produced by same-language speakers correctly more often than emotions produced by different-language speakers. 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 (Laukka & Elfenbein, 2021; Pell et al., 2009), as a result of cultural norms and language-specific prosodic cues influencing intercultural emotion recognition (Elfenbein & Ambady, 2002; Pell et al., 2009; Scherer et al., 2001). Taking the results for Hypotheses 1 and 2 together, the present study provides support for the Dialect Theory of emotion 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, 2002; Elfenbein et al., 2007).

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

First, we found that low-arousal emotions were recognized more accurately than high-arousal emotions. 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) and that listeners are able to distinguish between emotions that are high and low in arousal (Laukka et al., 2005), this is the first study, to the best of our knowledge, that has directly compared the recognition of low-arousal and high-arousal emotions, both within and across cultures. 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 adds new insights into the role of arousal in the communication of emotion.

Second, we found that negative emotions were recognized more accurately than positive emotions, as predicted in Hypothesis 4. As far as we are aware, our 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 nonlinguistic vocalizations. Furthermore, 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). Furthermore, our findings show that valence also affects the in-group advantage, as the magnitude of the in-group effect was relatively large for positive compared with negative emotions.

Third, we found that basic emotions were recognized more accurately than nonbasic emotions, as predicted in Hypothesis 5. Our cross-cultural findings are consistent with earlier findings that basic emotions can be decoded more accurately than nonbasic emotions across cultures in nonlinguistic vocalizations (Sauter et al., 2010) as well as in facial expressions (Ekman, 1972; Elfenbein & Ambady, 2002). 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, 1992b; Ekman et al., 1969). However, the finding that listeners recognized not only our four basic emotions but also our four nonbasic emotions above chance across (as well as within) cultures, provides a challenge for the strong version of Basic Emotion theory (Gendron et al., 2018). As far as we are aware, our study has been the first to compare the recognition of basic and nonbasic emotions within cultures. We found that basic emotions were recognized more accurately not only

across but also within cultures. The recognition advantage of basic over nonbasic emotions was relatively large when listening to out-group compared with in-group speakers, which can be seen to extend Basic Emotion theory. Furthermore, the magnitude of the in-group advantage was relatively large for nonbasic compared with basic emotions, which underlines the importance of the choice of emotions when studying the in-group effect.

We further observe that there is a close relationship between valence and basicness in our results—perhaps not surprisingly, because the two characteristics are, by definition, dependent. 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). Indeed, both positive and nonbasic emotions have been proposed to be closely connected to the formation and maintenance of social bonds (Shiota et al., 2004, 2017). Our findings showed that negative emotions were recognized more accurately than positive emotions, and that basic emotions were recognized more accurately than nonbasic emotions. The high recognition accuracy of negative and basic emotions reflects that valence and basicness often overlap, in our stimuli as well as by definition.

We do not observe a similar relationship between arousal and basicness, although they are dependent too. Among the four basic emotions in our experiment, only one was low in arousal (sadness), while the other three were high in arousal (joy, anger, fear). Yet, low-arousal emotions were, as a group, recognized *more* accurately than high-arousal emotions. This underscores the importance of the choice of the affective dimensions, and specific emotions, in emotion recognition research, as they will necessarily impact the outcomes of any study.

In addition to the above-mentioned theoretical aims, this study also had a methodological aim. We have presented and demonstrated the Demo/Koremo corpus for Dutch and Korean emotional speech (previously presented by Goudbeek & Broersma, 2010a, 2010b) with the aim of contributing to the methodological toolkit of intercultural emotion recognition research in general, and the methodological development of the study of cross-cultural vocal emotion recognition in particular. What has it yielded?

First, in the realm of reproducibility, this study has replicated previous findings of above-chance cross-cultural vocal emotion recognition, and of the in-group advantage in cross-cultural vocal emotion recognition. The results support the current consensus that the expression and recognition of emotions are affected by both universal and cultural/linguistic factors.

Second, this study's affectively and linguistically balanced design has allowed new insights into the influence of arousal, valence, and basicness on emotion recognition. The present findings underline the importance of the dimensions of arousal and valence, and of the concept of basicness, and the urgent need for studies of emotion recognition to take them into account. While some of the results have been shown before, others—the effect of arousal, and the within-cultural effect of valence and basicness—have not been previously investigated, and *could* not be investigated without an affectively and linguistically balanced design.

In spoken language, the effects of arousal, valence, and basicness on emotion recognition are thus evident both within and across cultures. Their effects outside the domain of spoken language might not be the same. While arousal, valence, and basicness are known to play a role in emotion recognition in the visual (Laukka et al., 2005; Russell, 2003; Scherer, 2009) and tactile modality (Raheel et al., 2019) as well, their role can be expected to differ there (Bänziger et al., 2012). For example, while valence is easily recognizable in the face, it is harder to recognize in the voice; arousal, on the other hand, is hard to recognize in the face but particularly easy to recognize in the

voice (Goudbeek & Scherer, 2010). Future studies could thus explore how arousal, valence, and basicness affect emotion recognition in all the channels that humans use to express their emotions, including linguistic and nonlinguistic vocalizations, facial expressions, gestures, body posture, and proximity to the interlocutor (Keltner et al., 2019; Mehrabian, 2017; Scherer, 2003, 2019).

Investigating the dimensions along which emotions are perceived in all those domains will provide a more thorough understanding of an essential part of human nature, that is at the core, or flows from the core, of all cognitive processes and behavior (Feldman Barret, 2017; Feldman Barrett & Russell 2014; Pessoa, 2015; Russell, 2003).

Acknowledgements

We thank Kichun Nam of the School of Psychology, Korea University, Seoul, for giving us the opportunity to test participants in his lab. We thank Taehong Cho of the Hanyang Institute for Phonetics and Cognitive Sciences of Language, Hanyang University, Seoul, for the opportunity to record the Korean emotion portrayals for the Demo/Koremo corpus in his lab.

Author contributions

This statement uses the roles defined in the Contributor Role Taxonomy (Brand et al., 2015). Y.L. contributed to the first stage of formal analysis and to writing the original draft as a student under supervision of M.G., A.K., J.C. and M.B.; M.G.: conceptualization, methodology, software, writing—review and editing, supervision; A.K.: formal analysis, writing—original draft, writing—review and editing, visualization, supervision; J.C.: methodology, investigation, writing—review and editing, supervision; M.B.: conceptualization, methodology, resources, data curation, writing—original draft, writing—review and editing, supervision, project administration, funding acquisition.

Data availability

The following data are available in the Radboud Data Repository via https://doi.org/10.34973/5kg3-9852: data, R-scripts, the Demo/Koremo corpus for Dutch and Korean emotional speech, scenarios used in creation of and supplementary materials for the Demo/Koremo corpus.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a Veni grant and a Vidi grant from the Dutch Research Council NWO to M.B.

Ethical considerations

This study was conducted in accordance with the Declaration of Helsinki.

Consent to participate

Participants provided verbal informed consent to conduct the study.

Consent for publication

Participants provided verbal informed consent to publish the study. All information has been anonymized.

ORCID iDs

Yachan Liang https://orcid.org/0000-0001-5077-7529
Martijn Goudbeek https://orcid.org/0000-0002-7787-4123
Agnieszka Konopka https://orcid.org/0000-0001-5557-8668
Jiyoun Choi https://orcid.org/0000-0002-1560-2384
Mirjam Broersma https://orcid.org/0000-0001-8511-2877

Notes

1. Note that we follow the terminology used by Goudbeek and Broersma (2010a), whereas Laukka and 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.
2. The scenarios and corpus are available online: https://doi.org/10.34973/5kg3-9852.
3. 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."
4. The maximal random structure for all models included random by-participant and by-item intercepts, random by-participant and by-item slopes for In-/out-group as this variable was manipulated within participants and within items, random by-participant slopes for 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. In the best-fitting models reported in the paper, 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. 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).
5. This might be due to characteristics of the participant groups or differences in the test conditions beyond our control. Alternatively, it might result from a cultural bias in the construction of the corpus. Even though care was taken to make the corpus as balanced as possible, with actors, directors, and the researchers involved in corpus construction all including Korean and Dutch individuals, the choice of the emotions and of the scenarios that were provided to the actors to illustrate the emotions prior to reenactment were based on the prior literature, which is predominantly Indo-European-based.
6. For the sake of completeness, we further assessed the In-/out-group effect for both Dutch and Korean listeners by presenting numerical differences across conditions. The two-level variables Listener Language, Arousal and Valence can be combined in eight ways. In seven out of eight combinations, there was a trend (or significant effect) of In-/out-group in the same direction (see Table 3). For Korean listeners and high-arousal negative emotions, there was a trend in the opposite direction, i.e., a numerical *out-group* advantage. This trend may be due to a difference in the salience of high-arousal negative emotions in Korean versus Dutch. Alternatively, it might result from a cultural bias in the construction of the corpus, as suggested in footnote 5.

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

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., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, *12*(5), 1161.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In A. C. R. Paiva, R. Prada, & R. W. Picard (Eds.), *International Conference on Affective Computing and Intelligent Interaction* (pp. 476–487). Springer.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (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

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beier, E., & Zautra, A. J. (1972). The identification of vocal expressions of emotion across cultures. *Journal of Consulting and Clinical Psychology*, *40*(4), 560.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

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

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

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. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Conference of Phonetic Sciences (ICPhS)*, San Francisco, CA, United States (pp. 969–972).

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

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

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), E7900–E7909. 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

Cutler, A., & Van Donselaar, W. (2001). Voornaam is not (really) a homophone: Lexical prosody and lexical access in Dutch. *Language and Speech*, *44*(2), 171–195. https://doi.org/10.1177/00238309010440020301

Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed.). John Murray. (Original work published 1872)

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. K. Cole (Ed.), *Nebraska Symposium on Motivation* (Vol. 19, pp. 207–283). University of Nebraska Press.

Ekman, P. (1992a). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

Ekman, P. (1992b). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

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., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88.

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. (2002). 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., 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

Feldman Barrett, L. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Feldman Barrett, L., & Russell, J. A. (2014). *The psychological construction of emotion*. Guilford Publications.

Gendron, M., Crivelli, C., & Feldman Barrett, L. (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

Goudbeek, M., & Broersma, M. (2010a). 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.), *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta (pp. 2211–2215).

Goudbeek, M., & Broersma, M. (2010b). Language specific effects of emotion on phoneme duration. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Chiba, Japan (pp. 2026–2029).

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. https://doi.org/10.1121/1.3466853

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 P. In Sun-Ah Jun (Ed.), *The phonology of intonation and phrasing* (pp. 118–145). Oxford University Press. https://doi.org/10.1093/acprof

Harre, R. (1986). *The social construction of emotions*. Blackwell.

Huang, C. F., Erickson, D., & Akagi, M. (2008). Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners. In *Proceedings European Conference on Noise Control* (pp. 2317–2322). https://doi.org/10.1121/1.2933803

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

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

Jun, S. A. (2005). Prosody in sentence processing: Korean vs. English. *UCLA Working Papers in Phonetics*, 104, 26–45. http://phonetics.linguistics.ucla.edu/workpapph/104/3-Jun-WPP104.pdf

Jun, S. A. (2006). Intonational phonology of Seoul Korean revisited. *Japanese/Korean Linguistics*, *14*, 15–26.

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., & 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

Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of Nonverbal Behavior*, *43*, 133–160.

Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and Speech*, *51*(4), 343–359.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. https://doi.org/10.1121/1.398894

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

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

Laukka, P. (2003). Categorical perception of emotion in vocal expression. *Annals of the New York Academy of Sciences*, *1000*, 283–287. https://doi.org/10.1196/annals.1280.026

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., 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*, 1–8. 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., 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

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.

Marino, L. (2017). Thinking chickens: A review of cognition, emotion, and behavior in the domestic chicken. *Animal Cognition*, *20*(2), 127–147.

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

Mehrabian, A. (2017). *Nonverbal communication*. Routledge.

Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin*, *112*(2), 179–204.

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

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

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

Pessoa, L. (2015). Précis on the cognitive-emotional brain. *Behavioral and Brain Sciences*, *38*, e71.

Raheel, A., Anwar, S. M., & Majid, M. (2019). Emotion recognition in response to traditional and tactile enhanced multimedia using electroencephalography. *Multimedia Tools and Applications*, *78*(10), 13971–13985.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141.

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

Sacharin, V., Schlegel, K., & Scherer, K. R. (2012). *Geneva emotion wheel rating study*. Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences.

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., & 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

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

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., 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, S., Siegert, I., Bigalke, L., & Meudt, S. (2010). Developing an expressive speech labeling tool incorporating the temporal characteristics of emotion. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta (pp. 1172–1175).

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). Lawrence Erlbaum.

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

Shochi, T., Rilliard, A., & Aubergé, V. (2009). Intercultural perception of English, French. *The Role of Prosody in Affective Speech*, *97*, 31.

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

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

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

van Bezooijen, R. (1984). *Characteristics and recognizability of vocal expressions of emotion*. Foris. https://doi.org/10.1515/9783110850390

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

Appendix A

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

	Mean	*t*	*df*	*p*	Range	95% confidence interval for the mean	
						Lower	Upper
Dutch listeners responding to Dutch recordings	0.47	32.69	30	$<.001$	.32–.58	0.45	0.49
Dutch listeners responding to Korean recordings	0.38	34.77	30	$<.001$	.30–.46	0.36	0.40
Korean listeners responding to Dutch recordings	0.36	22.80	23	$<.001$	.27–.48	0.34	0.38
Korean listeners responding to Korean recordings	0.43	23.70	23	$<.001$	.34–.59	0.40	0.46

Notes. 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	*t*	*df*	*p*	Range	95% confidence interval for the mean	
						Lower	Upper
Basic emotions: Dutch listeners responding to Dutch recordings	0.56	24.55	30	$<.001$	.36–.73	0.52	0.60
Basic emotions: Dutch listeners responding to Korean recordings	0.47	30.81	30	$<.001$	.30–.64	0.45	0.49
Non-basic emotions: Dutch listeners responding to Dutch recordings	0.38	23.29	30	$<.001$	.25–.52	0.36	0.40
Non-basic emotions: Dutch listeners responding to Korean recordings	0.29	15.40	30	$<.001$	.16–.41	0.27	0.31
Basic emotions: Korean listeners responding to Dutch recordings	0.49	23.04	23	$<.001$	.31–.63	0.46	0.52
Basic emotions: Korean listeners responding to Korean recordings	0.46	18.08	23	$<.001$	.30–.69	0.42	0.50
Non-basic emotions: Korean listeners responding to Dutch recordings	0.23	8.45	23	$<.001$	.09–.36	0.20	0.26
Non-basic emotions: Korean listeners responding to Korean recording	0.39	20.63	23	$<.001$	.27–.50	0.36	0.42

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