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

Liang, Y.

Chapter Two

Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design[5]

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 × n matrix), where the sets of speaker and listener languages are the same, provide complete information regarding these interactions. However, rectangular designs (n × 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* /ˈkanɔn/ vs. *kaNON* /kaˈnɔn/, 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 [o] 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 × 8 actors × 2 tokens × 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 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 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”.

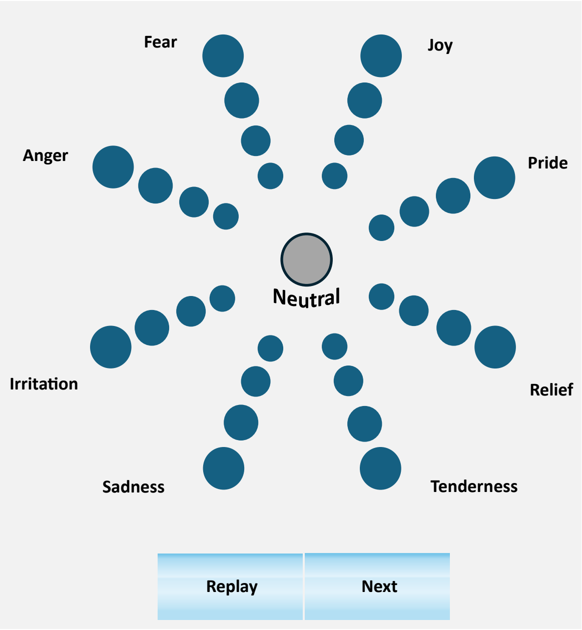

Figure 2.1. The emotion wheel in English (reproduced from Liang et al., 2023). Translation in Dutch and Korean, Joy: “Blijdschap”, “행복”; Pride: “Trots”, “자랑스러움”; Relief: “Opluchting”, “안도감”; Tenderness: “Vertedering”, “애정”; Sadness: “Verdriet”, “슬픔”; Irritation: “Irritatie”, “짜증”; Anger: “Woede”, “분노”;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

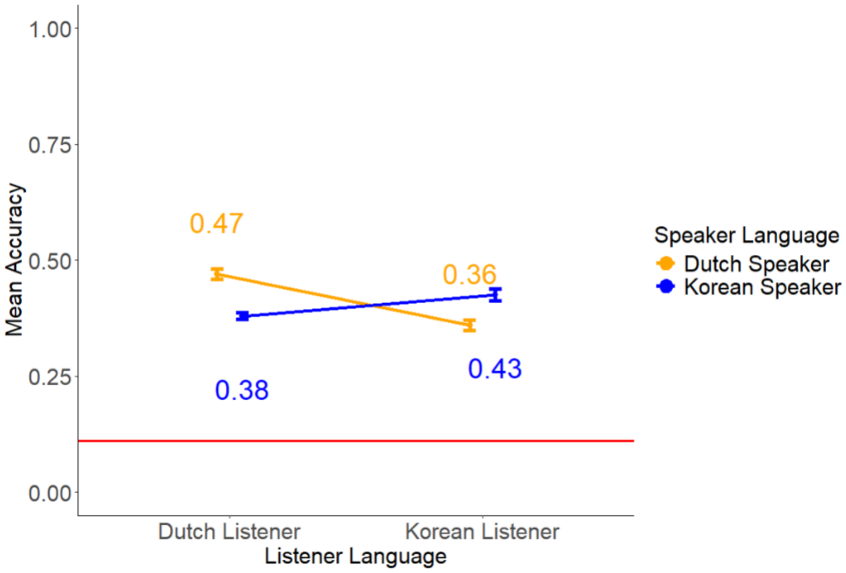

Figure 2.2. Accuracy (proportion of correct responses) for Dutch and Korean recordings by Dutch and Korean listeners. The red line indicates chance performance (.11). Error bars are ±1 SE in all figures.

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 & Elfenbein, 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 (*ß*) are transformed back to odds (*exp*(*ß*)) 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				
	ß	*Exp (ß)*	*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 × LL	0.92	2.51	0.15	6.08	**< .001**

Note. The Speaker Language × Listener Language interaction showed a reliable in-group effect (*ß* =.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.

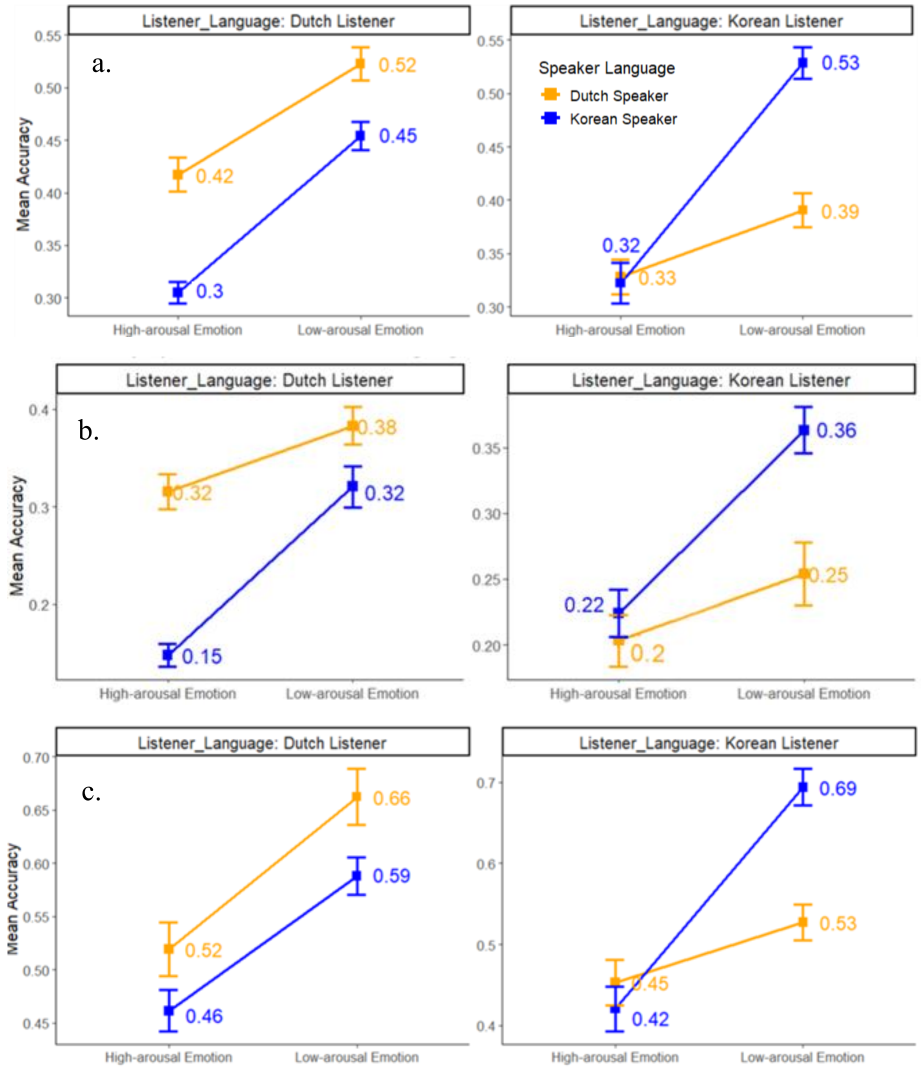

Figure 2.3. Recognition accuracy (Proportion correct) for high-arousal and low-arousal emotions in Dutch and Korean recordings by Dutch and Korean listeners (a) in the entire dataset (positive and negative emotions), (b) for positive emotions, and (c) for negative emotions.

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				
	ß	*Exp (ß)*	*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 × LL	0.92	2.51	0.15	6.09	< **.001**
SL × A	0.70	2.01	0.40	1.77	.076
LL × A	0.07	1.07	0.19	0.37	.711
SL × LL × 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 × LL	1.22	3.39	0.23	5.30	< **.001**
SL × A	0.97	2.64	0.47	2.05	< **.050**
LL × A	−0.22	0.80	0.25	−0.87	.387
SL × LL × 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 × LL	0.77	2.16	0.21	3.64	< **.001**
SL × A	0.45	1.57	0.49	0.92	.359
LL × A	0.30	1.35	0.28	1.10	.273
SL × LL × 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, $\chi^2(1) = .38$, $p = .54$).

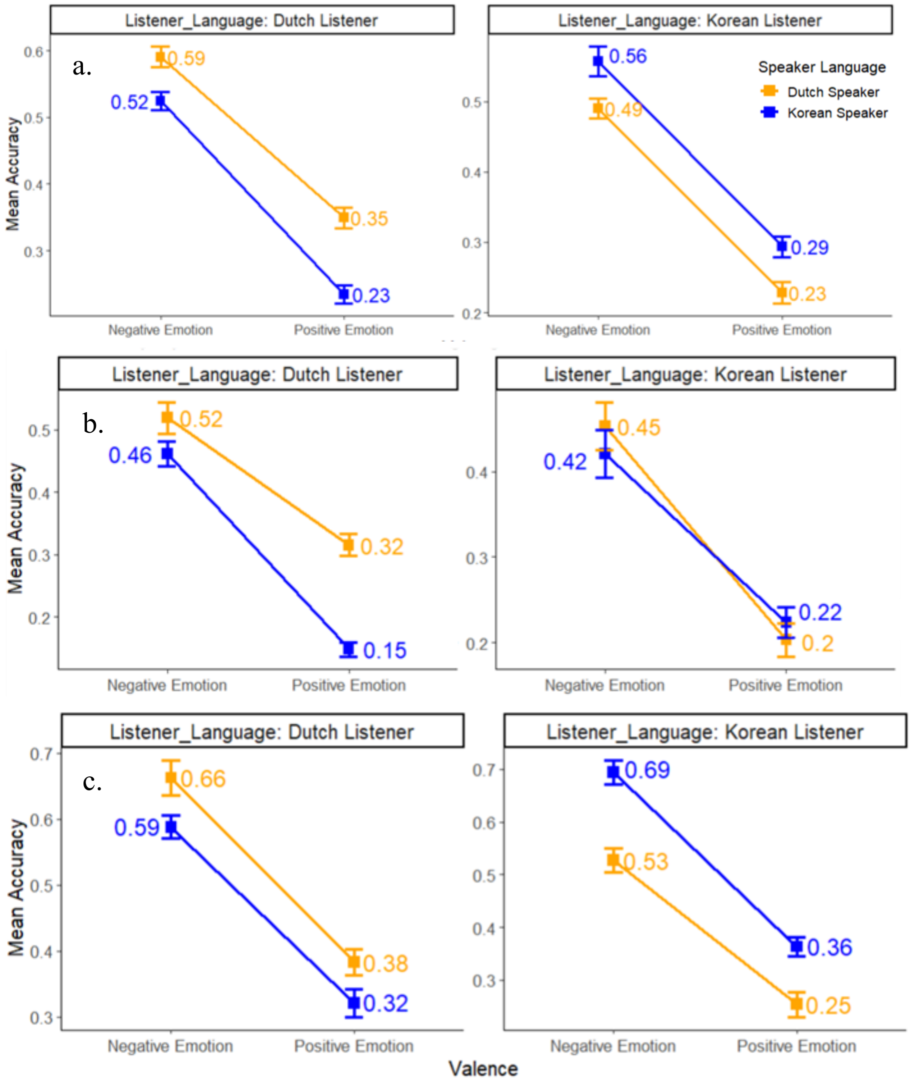

Figure 2.4. Recognition accuracy (Proportion correct) for positive and negative emotions in Dutch and Korean recordings by Dutch and Korean listeners in the entire dataset (high-arousal and low-arousal emotions), (b) for high-arousal emotions, and (c) for low-arousal emotions.

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				
	ß	*Exp (ß)*	*SE*	*z*	*p*
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 × LL	0.95	2.59	0.15	6.20	< **.001**
SL × V	−0.25	0.78	0.35	−0.71	.477
LL × V	−0.11	0.90	0.20	−0.54	.591
SL × LL × 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 × LL	0.76	2.14	0.22	3.42	< **.001**
SL × V	−0.52	0.59	0.46	−1.14	.256
LL × V	0.19	1.21	0.26	0.73	.466
SL × LL × 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 × LL	1.25	3.50	0.22	5.74	< **.001**
SL × V	0.02	1.02	0.51	0.05	.962
LL× V	−0.38	0.68	0.29	−1.32	.188
SL × LL × 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				
	ß	*Exp (β)*	*SE*	*z*	*p*
Intercept	−0.55	0.58	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 × LL	0.92	2.51	0.15	6.20	< **.001**
SL × B	0.44	1.55	0.38	1.15	.251
LL × B	0.01	1.01	0.19	0.07	.947
SL × LL × 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.

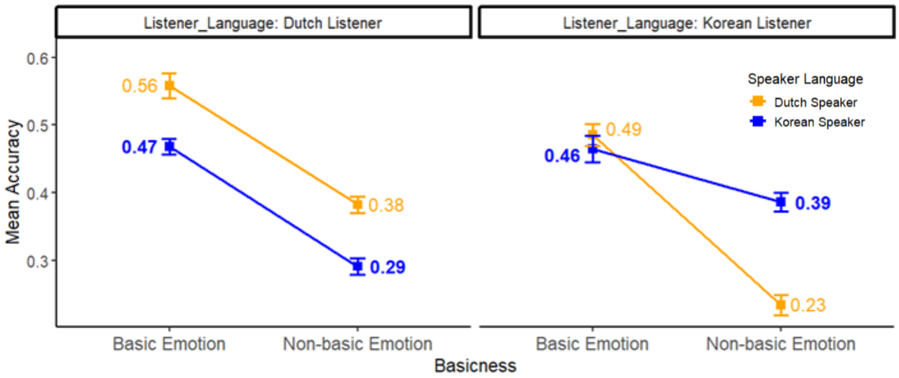

Figure 2.5. Recognition accuracy (Proportion correct) for basic and non-basic emotions in Dutch and Korean recordings by Dutch and Korean listeners.

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.