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

Liang, Y.

Chapter Six

Conclusion and discussion

6.1 Introduction

This dissertation investigated cross-language (Dutch and Korean) vocal emotion recognition from multiple perspectives, examining 1) cross-cultural and/or cross-linguistic vocal emotion recognition by Dutch and Korean listeners; 2) intensity ratings of vocal emotions by Dutch and Korean listeners; 3) patterns of acoustic parameters across emotion, speaker language, and gender; 4) the role of cultural and prosodic similarity in vocal emotion recognition by American English and French listeners. Each perspective was addressed in a separate chapter, with four main research questions:

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

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

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

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

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

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

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

6.2 Main findings

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

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

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

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

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

6.2.2 Interpreting the intensity of vocal emotions across cultures

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

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

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

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

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

6.2.3 Classifying emotions from acoustic parameters

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

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

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

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

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

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

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

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

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

6.3 Discussion

6.3.1 Adapting the dimensional approach

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

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

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

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

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

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

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

Cluster analysis is a statistical method used to group items together that share similarities, where closer/clustered emotions are more similar than those farther apart (Albornoz et al., 2011; Kurematsu et al., 2010). In Chapter 2, the Korean and Dutch listeners had to recognize the emotions in their own language. The analysis focused on the recognition accuracies, not further analyzing the confusion matrices. We present all confusion matrices in Table I1 in Appendix I. These confusion matrices may be used to investigate which emotions are more similar and which are more dissimilar. This is especially relevant in comparing the Dutch and Korean data when listeners evaluate the emotional expressions in their own language. Is the similarity structure of the emotions in the two languages the same?

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

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

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

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

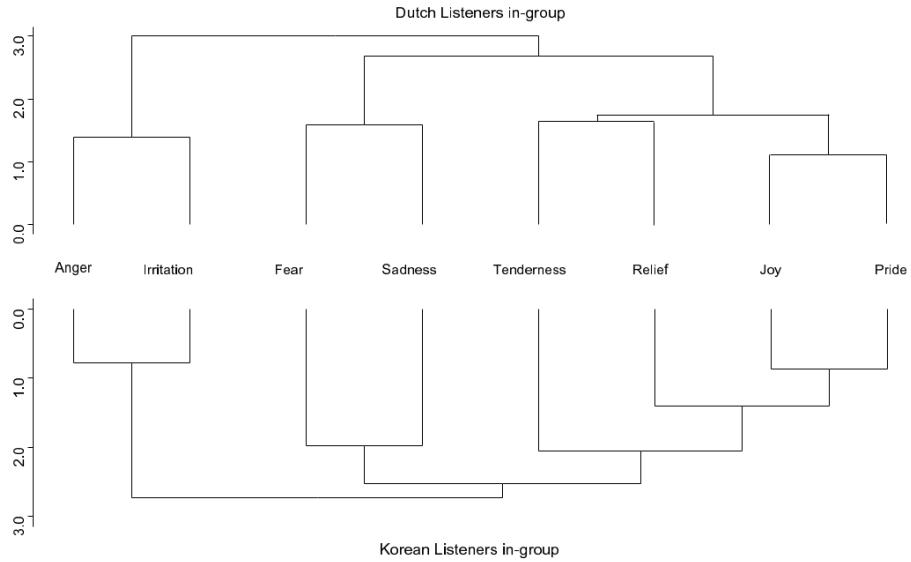

Figure 6.1. Hierarchical clustering dendrograms (Ward's method) for in-group identification of vocal emotions by Dutch (upper tree) and Korean (lower tree) listeners.

6.3.3 In-group advantage

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

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

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

6.3.4 Prosodic structure and vocal emotion recognition

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

6.3.5 Acoustic parameters identifying vocal emotions

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

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

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

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

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

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

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

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

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

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

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

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

6.4 General conclusions

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

6.4.1 The Cultural Proximity hypothesis

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

6.4.2 The Prosodic Proximity hypothesis

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

6.4.3 Acoustic parameters of vocal emotions

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

6.4.4 Dimensionality of vocal emotions

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

6.5 Limitations

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

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

6.5.1 Phonological legitimacy

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

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

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

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

6.5.2 The eight emotions

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

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

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

6.5.3 The neutral category

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

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

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

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

6.5.4 Acoustic correlates of emotional intensity

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

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

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

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

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

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

6.5.5 The impact of stimulus order on recognition accuracy

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

6.5.6 The ecological validity of stimuli

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

6.6 Future research

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

6.6.1 Second language acquisition

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

6.6.2 Acoustic manipulations of stimuli

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

6.6.3 Neuroimaging

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