The Demo / Kemo corpus: A principled approach to the study of cross-cultural differences in the vocal expression and perception of emotion

Martijn Goudbeek* and Mirjam Broersma**

*University of Tilburg, The Netherlands
**Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen,
and
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
m.b.goudbeek@uvt.nl, mirjam@mirjambroersma.nl

Abstract

This paper presents the Demo / Kemo corpus of Dutch and Korean emotional speech. The corpus has been specifically developed for the purpose of cross-linguistic comparison, and is more balanced than any similar corpus available so far: a) it contains expressions by both Dutch and Korean actors as well as judgments by both Dutch and Korean listeners; b) the same elicitation technique and recording procedure was used for recordings of both languages; c) the same nonsense sentence, which was constructed to be permissible in both languages, was used for recordings of both languages; and d) the emotions present in the corpus are balanced in terms of valence, arousal, and dominance. The corpus contains a comparatively large number of emotions (eight) uttered by a large number of speakers (eight Dutch and eight Korean). The counterbalanced nature of the corpus will enable a stricter investigation of language-specific versus universal aspects of emotional expression than was possible so far. Furthermore, given the carefully controlled phonetic content of the expressions, it allows for analysis of the role of specific phonetic features in emotional expression in Dutch and Korean.

1. Introduction

In this paper, we present a corpus of Dutch and Korean vocal emotion expressions: the Demo (Dutch Emotion) / Kemo (Korean Emotion) corpus. In contrast to existing corpora of vocal emotion expressions, the present corpus has been specifically developed for the purpose of cross-linguistic and cross-cultural comparison. Therefore, it is more balanced than any materials available so far. The corpus contains a comparatively large number of emotions (eight) uttered by a large number of speakers (eight Dutch and eight Korean). Further, the phonetic content of the expressions has been carefully selected to enable the analysis of the role of specific phonetic features in the expression and recognition of emotion in the two languages.

Basic emotions like joy, anger, fear, disgust and sadness have been shown to be recognized well above chance levels between cultures (Elfenbein and Ambady, 2002). The increasing body of evidence that shows recognition to be above chance between cultures, as well as certain invariant properties in the expression of emotion, have been taken as support for basic emotion theory (Ekman et al., 1969; Ekman, 1992). The line of reasoning is that if members from a radically different culture are able to understand which emotion is being expressed, then this expression can be said not to depend on cultural factors, but must be universal.

Most of this research investigates facial expression of emotion, although some notable efforts have been made in the past with respect to vocal expression (Albas et al., 1976; van Bezooijen et al., 1983). Recently, the effects of language and culture on the vocal encoding and decoding of emotion has been the topic of many studies (Scherer et al., 2001; Thompson and Balkwill, 2006; Pell and Skorup, 2008; Sauter et al., 2009; Pell et al., 2009). In general, these studies follow one of two design choices. They either use an existing corpus of emotional expression that has been developed and validated in one language and use this material to assess the ability of speakers from one or more other languages to successfully decode the emotional expressions (e.g., Van Bezooijen et al., 1983; Scherer et al., 2001; Pell and Skorup, 2008; Sauter et al., 2009). Alternatively, corpora are developed and validated in several languages and then presented to listeners from one language (e.g., Thompson and Balkwill, 2006; Pell et al, 2009).

In both types of studies, the emotions are usually expressed in meaningless phrases or nonverbal affect vocalizations. Both contain a minimum of semantic information while remaining linguistically valid. Nonverbal affect vocalizations have the advantage over meaningless phrases that they are more natural. However, they have the serious disadvantage of possibly being semantically loaded (e.g., yuck might be a universal expression of disgust, regardless of its nonverbal realization).

An example of the first (one to many) approach is the study by Scherer and colleagues (Scherer et al., 2001). They presented a German corpus of vocal emotion expressions (Banse and Scherer, 1996) to judges from nine different languages and cultures ranging from Europe and North America to Asia. An example of the second (many to one) approach is the work by Pell and colleagues (Pell et al., 2009). They presented monolingual Argentinian Spanish listeners with emotional expressions in Argentinian Spanish, English, German, and Arabic.

Both studies showed that vocal expressions of (basic) emotions can be accurately decoded when listening to expressions in a foreign language. Similarly, recent work by Sauter and colleagues (2009) showed that even judges of the isolated Himba community manage to accurately decode emotional vocalizations of British speakers. This suggests that it is certainly likely that some aspects of vocal emotion expression are present in most if not all languages.

Nevertheless, both Scherer et al. and Pell et al. also conclude that there are language specific elements in the decoding of vocal emotion expression. Pell et al. label this an 'in-group' advantage (participants perform better in their native language than in a nonnative language), and Scherer et al. formulate the bolder language distance hypothesis: based on their finding that listeners of a closely related language like English were better at categorizing German emotional expressions than listeners of a more remote language like Indonesian, Scherer et al. conclude that linguistic similarity plays an important role when decoding emotional expressions in a foreign language.

Taken together, these studies strongly suggest that expression and understanding of emotion is based on a combination of universal and language- and culture-specific processes. These simultaneous contributions of language and culture independent and language and culture dependent processes appear to be much stronger for vocal than for facial expression of emotion (Elfenbein and Ambady, 2002).

As explained above, most studies use either the many-to-one or the one-to-many design (a notable exception being the work of Alblas et al. (1976) that combines both designs). While these studies are carefully executed, their lopsided design makes it difficult to untangle universal and language specific effects. To better understand the relative contributions of universal and language specific effects on the recognition of vocal emotion expression, a design where the languages of encoders and decoders are fully balanced is needed.

The Demo / Kemo corpus is a balanced corpus of posed vocal emotion expressions in a number of ways. First, it contains expressions by Dutch and Korean speakers and judgments by Dutch and Korean listeners. Second, in the recording phase, the same elicitation technique was used by the Korean and Dutch stage directors. Third, the speakers of both languages used the same verbal expression that was carefully constructed to contain phonemes present in both languages in combinations permissible in both languages. Finally, the emotions in the corpus are balanced in terms of valence, arousal, and dominance characteristics.

In this paper, we describe the recording of the materials and the judgment studies that were carried out in order to select the tokens included in the Demo / Kemo corpus. Further we address the benefits of the corpus and our aims for it.

2. Corpus recording

In the recording of the corpus we adhered to the methods developed by Scherer and colleagues (Banse and Scherer, 1996; Bänziger and Scherer, 2007). This approach uses posed emotional expressions by (semi-) professional actors. While acted portrayals are in principle not spontaneous, the approach aims at ensuring the naturalness of these expressions by using the method acting principles put forward by Stanislawski (1936). In Stanislavski's approach a director coaches the actors to produce full-blown emotional reactions by remembering and reliving a personal episode in which the target emotion occurred or by very vividly imagining such episodes (Stanislavski, 1936). In addition, in this study, the actors were given three possible scenarios illustrating the emotion for them (Banse and Scherer, 1996).

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

Table 1: The emotions in the corpus in a valence by arousal grid.

Eight Dutch actors (four males and four females) and eight Korean actors (four males and four females) participated in exchange for a small payment. All were or had been engaged in a full-time professional drama school at college level in their own country. Both directors were professionals well acquainted with the Stanislawski technique. Table 1 lists the emotions that were posed by the actors. The order in which the emotions were enacted was counterbalanced between actors.

The actors had to express the emotion using a fixed phrase [nuto hɔm sɛpikaŋ]. This phrase was constructed according to the following three criteria. First, the phrase contains only phonemes that occur in both Dutch and Korean, in phonotactic combinations that are legal in both languages. Second, it is meaningless in both languages. Third, the phrase does not contain any clearly embedded words.

Each actor was recorded individually, in the presence of a (Dutch or Korean) stage director. Director and actor worked on reliving one emotion at a time. Actors were free to improvise, silently or using any speech desired. Once the actor felt confident expressing the emotion, (also) using the crucial phrase, recording started. Recording for that particular emotion finished when the actor had produced at least five good expressions of the emotional phrase, as determined by the director. Most recording sessions, including all eight emotions, took about two hours two complete. The final four recorded utterances of each emotion of each actor that were acceptable from an acoustic point of view were included in the judgment study. This resulted in 256 utterances in each language set (8 actors * 8 emotions * 4 repetitions).

3. Judgment studies

Two judgment studies were conducted to investigate the quality and naturalness of the emotional expressions as judged by listeners sharing the native language of the actors.

3.1. Method

3.1.1. Participants

Two groups of listeners participated in the experiment: 24 Dutch listeners (11 males, 13 females) recruited from the Radboud University Nijmegen in the Netherlands, and 24 Korean listeners (12 males, 12 females) recruited from Korea University in Seoul, Korea. All participants were students and participated in exchange for a small payment or course credits. All were native speakers of Dutch and Korean respectively and none of them reported any hearing or speech problems.

	Speaker		Emotion								
			Anger	Fear	Irritation	Joy	Pride	Relief	Sadness	Tenderness	**Total**
Dutch	1	Min	0.12	0.01	0.00	0.13	0.03	0.00	0.15	0.07	**0.00**
		Max	0.41	0.19	0.06	0.60	0.13	0.01	0.36	0.22	**0.60**
	2	Min	0.03	0.12	0.02	0.02	0.03	0.00	0.36	0.00	**0.00**
		Max	0.81	0.19	0.40	0.17	0.17	0.01	0.45	0.02	**0.81**
	3	Min	0.60	0.01	0.11	0.00	0.01	0.00	0.24	0.00	**0.00**
		Max	0.74	0.19	0.46	0.13	0.13	0.16	0.45	0.07	**0.74**
	4	Min	0.03	0.09	0.01	0.68	0.06	0.08	0.32	0.45	**0.01**
		Max	0.18	0.69	0.22	0.96	0.17	0.21	0.60	0.75	**0.96**
	5	Min	0.18	0.12	0.35	0.07	0.00	0.01	0.24	0.00	**0.00**
		Max	0.66	0.54	0.58	1.06	0.13	0.47	0.55	0.11	**1.06**
	6	Min	0.47	0.29	0.18	0.01	0.00	0.16	0.21	0.16	**0.00**
		Max	0.60	0.54	0.46	0.86	0.09	0.47	0.60	0.87	**0.87**
	7	Min	0.60	0.24	0.22	0.10	0.03	0.33	0.40	0.02	**0.02**
		Max	0.74	0.69	0.58	0.27	0.36	0.95	0.50	0.64	**0.95**
	8	Min	0.60	0.19	0.09	0.10	0.06	0.12	0.28	0.00	**0.00**
		Max	0.89	0.47	0.35	0.45	0.36	0.47	0.45	0.45	**0.89**
Korean	1	Min	0.06	0.01	0.08	0.00	0.00	0.00	0.00	0.14	**0.00**
		Max	0.79	1.16	0.59	0.02	0.01	0.03	0.58	2.06	**2.06**
	2	Min	0.00	0.00	0.02	0.00	0.01	0.00	0.05	0.00	**0.00**
		Max	0.55	0.22	0.15	1.26	1.10	0.37	0.28	0.47	**1.26**
	3	Min	0.09	0.00	0.04	0.02	0.00	0.08	0.28	0.00	**0.00**
		Max	0.88	0.56	0.35	0.63	0.01	0.31	0.53	0.06	**0.88**
	4	Min	0.01	0.49	0.00	0.04	0.00	0.00	0.05	0.00	**0.00**
		Max	0.35	0.97	0.35	0.44	0.09	0.00	0.25	0.02	**0.97**
	5	Min	0.02	0.00	0.35	0.00	0.00	0.01	0.25	0.00	**0.00**
		Max	0.63	0.26	0.44	0.00	0.02	0.31	0.44	0.56	**0.63**
	6	Min	0.01	0.02	0.00	0.04	0.01	0.05	0.13	0.02	**0.00**
		Max	0.35	0.97	0.27	0.28	0.45	0.44	0.58	0.88	**0.97**
	7	Min	0.00	0.11	0.02	0.00	0.02	0.00	0.09	0.19	**0.00**
		Max	0.35	0.26	0.21	0.02	0.81	0.52	0.32	0.47	**0.81**
	8	Min	0.00	0.01	0.00	0.00	0.00	0.00	0.16	0.04	**0.00**
		Max	0.88	0.14	0.10	1.12	0.09	0.01	0.44	0.88	**1.12**
	Min		**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Max		**0.89**	**1.16**	**0.59**	**1.26**	**1.10**	**0.95**	**0.60**	**2.06**	**2.06**	**2.06**

Table 2: Minimum (top row) and maximum (lower row) unbiased hit rates for the Dutch and Korean sample for each speaker/emotion pair. The mean unbiased hit rates are based on four portrayals per cell.

3.1.2. Materials

As described above, the materials were the 256 Dutch and 256 Korean selected utterances (8 actors * 8 emotions * 4 repetitions). They were segmented into separate wave files (mono, 44.1 kHz, 16 bit, uncompressed) that were not normalized with respect to intensity.

3.1.3. Procedure

The participants classified each of the 256 stimuli from their native language, that were presented to them in pseudo-random order. Stimuli were classified as one of the eight emotions (anger, fear, sadness, irritation, joy, pride, tenderness, relief), or as neutral. All response options were shown in written form on a computer screen, each in a separate square (all equally sized), at the same position (that reflected the valance and arousal properties of the stimulus) as shown in Table 1 and with the response option neutral in the middle. Participants indicated their response with a mouse click on the square that contained the name of the emotion category. The experiment was run using the Praat MFC object (Boersma, 2001) on a standard laboratory computer. After each categorical rating, participants had to indicate the naturalness of the expression on a scale ranging from 1 (very unnatural) to 4 (very natural).

3.2. Results

We computed unbiased hit rates for each portrayal (Wagner, 1993). Table 2 displays the minimum and maximum unbiased hit rates scores for each speaker and emotion for Dutch and Korean respectively.

The results show that for most basic emotions there was at least one portrayal per actor that was sufficiently recognized (unbiased hit rate > 0.1). For Dutch, only relief and tenderness had more than one speaker (i.e., two in both cases) who did not succeed in expression that emotion adequately. For Korean, five actors had difficulty expressing pride, while also joy, relief and tenderness were difficult for some actors (i.e., three, three, and two actors, respectively). An analysis of variance with the average unbiased hit rate as dependent variable, emotion (anger, fear, irritation, joy, pride, relief, sadness, tenderness) as within-subjects variable, and language (Dutch, Korean) as between-subjects variable revealed a significant effect of emotion ($F(1,7) = 3{,}596$, $p < 0.002$, $\eta^2 = 0.814$), showing that some emotions are better recognized than others. Importantly, there was no significant effect of language ($F(1,14) = 3.34$, *n.s.*), nor was there a significant interaction between language and emotion ($F(1,14) = 1{,}343$, *n.s.*), indicating that, on average, the Dutch and Korean actors expressed the emotions equally well.

4. Corpus selection

For the final corpus, the two portrayals of each actor-emotion pair with the highest unbiased hit rate were selected. When there was a tie, the portrayal with the higher naturalness rating was selected. When there still was a tie, portrayals that were confused with portrayals of the same emotion family were favored (e.g., when a pride portrayal was confused with joy it would be included before a portrayal confused with irritation). If all this failed, a portrayal was randomly selected.

For some emotions (notably relief, pride, and tenderness for Dutch, and joy, pride, and tenderness for Korean) portrayals had to be selected that were recognized at or below chance levels. We decided to include these portrayals despite their low recognition rate in order to retain a balanced set with two portrayals per emotion by each actor and to ensure a wide enough range of quality in the portrayals for the cross-cultural experiments (e.g., to prevent ceiling effects from obscuring differences between the two cultural groups).

5. Corpus availability

While still in the development stage, the corpus is intended to be shared with researchers working on cross linguistic factors in vocal emotion production and perception. When the corpus has been sufficiently developed and annotated, the raw materials and annotations in the corpus will be freely available to the research community.

6. Conclusion

The Demo / Kemo corpus has specifically been developed to investigate he role of language and culture in decoding vocal expression of emotion. This has been achieved by taking into account the relevant properties of the concerned languages from the outset. The carrier phrase is consistent with the phonetic and phonotactic properties of both Dutch and Korean. Furthermore, the recruitment of the actors, the recording of the utterances, the rating procedure and the selection of the portrayals has been kept constant or as similar as possible over all recordings. This way, the possible differences in decoding accuracy will be entirely due to linguistic and cultural factors.

We are currently employing the corpus in studies on cross-cultural emotion perception. In those studies, instead of the categorical judgment of the study reported here, participants give continuous ratings of emotions (e.g., indicate on a continuous scale to what extent an emotion is present in an expression). This way more subtle differences in the decoding of vocal emotion expression between Dutch and Korean listeners can be detected. Furthermore, this procedure explicitly recognizes the importance of so called blended emotions (see, for example, Banse & Scherer, 1996) and the possibility that the difference between languages might not lie in the first emotion attribution of listeners to a certain vocalization but in the perceived relation with other emotions.

We strongly feel that the balanced way this corpus is set up is an important step forward in the study of cross-cultural emotion research. We hope that more studies will employ a design in which the properties of the languages to be investigated are entered into the study design from the very beginning.

7. Acknowledgments

Martijn Goudbeek was supported by funding from the VICI project "Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions", funded by the Netherlands Organisation for Scientific Research (NWO grant 277-70-007, awarded to Emiel Krahmer). Mirjam Broersma was supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO). We thank Jiyoun Choi for her help recording the Korean corpus and testing the Korean participants, and Sammie Tarenskeen for testing the Dutch participants.

8. References

Daniel C. Albas, Ken W. McCluskey, and Cheryl A. Albas. 1976. Perception of the emotional content of speech: a comparison of two canadian groups. *Journal of Cross-Cultural Psychology*, 7:481–489.

Rainer Banse and Klaus R. Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70:614–636.

Tanja Bänziger and Klaus R. Scherer. 2007. Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *ACII*, pages 476–487.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345.

Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science*, 164:86 – 88.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6:169–200.

Hillary A. Elfenbein and Nalimi Ambady. 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128:203–235.

Marc D. Pell and Vera Skorup. 2008. Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, 50(6):519–530.

Marc D. Pell, Laura Monetta, Silke Paulmann, and Sonja Kotz. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33:107–120.

Disa Sauter, Frank Eisner, Paul Ekman, and Sophie K. Scott. 2009. Universal vocal signals of emotion. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society. Amsterdam, The Netherlands.*

Klaus R. Scherer, Rainer Banse, and Harald G. Wallbott. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32:76–92.

Constantin Stanislavski. 1936. *An actor prepares*. London: Methuen.

William F. Thompson and Laura-Lee Balkwill. 2006. Decoding speech prosody in five languages. *Semiotica*, 158:407–424.

Renée van Bezooijen, Stanley A. Otto, and Thomas A. Heenan. 1983. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, 14:387–406.

Hugh L. Wagner. 1993. On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, 17:3–28.