VOCAL EXPRESSION AND PERCEPTION OF EMOTION IN KOREAN

Soo-Jin Chung
Institut de Phonétique, Université de la Sorbonne Nouvelle, CNRS/UPRESA 7018, Paris. *
Current address: Cognitive and Linguistic Sciences Department, Brown University, Providence, RI 02912, USA.

ABSTRACT

The present study investigates the vocal expression of emotions in Korean and the perception of the emotions by Korean, American, and French listeners. The emotions were taken from spontaneous speech of three female Koreans. The listeners rated the emotionality on the scales of emotional valence and activation. The results showed that the listeners agreed on the global emotional perception. However, they differed in the subtle distinction of emotionality of the speakers, due the different use of lexical, cultural, and acoustical information.

1. INTRODUCTION

The perception of emotions has been often explained in terms of the listener's use of acoustic cues. Scherer [4] reports that listeners attribute emotional meaning to synthesized audio stimuli on the basis of particular acoustic cues. They rate the emotionality on the Activation dimension based on pitch and amplitude variation and tempo, whereas they evaluate the emotionality on the Valence dimension in reference to pitch level and variation. While these kinds of perceptual behavior implies the universality of basic emotional perception [3 & 5], the study of McCluskey *et al*. [1] shows that cultural differences in emotional perception are established early in life. The present paper addresses to the question to what extent the perception of positive and negative emotions is universal or culture-dependant, including Korean, American, and French listeners.

2. EXPERIMENT 1

This experiment examines similarities or differences in the perception of emotional expressions in spontaneous speech by Korean and American listeners. The emotions, happiness and sadness, were taken as the prototypes of positive and negative emotions and listeners rated them on the scales of emotional valence and emotional intensity.

2.1. Speech Material

The data were collected from a series of TV interviews videotaped in a broadcasting studio. Three interviews were selected for our study, including three Korean female speakers in the thirties. Each interview involved one speaker and one interviewer. At the beginning of the interview, the speaker was in a good mood, talking about pleasant experiences with her lover or family. Later, when she talked about her personal problems and difficulties, she became sad and spoke in tears.

All the interviews were carried out in an extemporary manner and none of the speakers was a professional actor or a famous person known to the public. There was no apparent attempt to pretend certain emotions and their emotional upset seemed to be too strong to be controlled. For these reasons, we considered the emotions of the three speakers manifested in their face or voice as real and genuine ones, which could be distinguished from emotions simulated by actors.

According to facial expressions of the speakers and contextual information, eight passages were excised from each interview, so 24 passages in total. Each passage roughly corresponded to a single speaking turn and its mean duration was 28.6 seconds (SD 12.3s). The structure of the eight passages for a given speaker could be described as follows: the first and second passages were extracted from the beginning of the interview when the speaker talked with happy smiling face; during the following four passages, she did not show any particular emotional expression; and the last two passages were taken from the moment when she experienced an extreme sadness and distress, as seen in her weeping face and sobbing speech. These characteristics of the interviews allowed us to establish three categories of emotion, positive, neutral and negative emotions, for each speaker.

As stimuli of Experiment 1, 275 utterances of three speakers were excised from the 24 passages and digitized at a sampling rate of 22kHz using 12-bit linear quantization. Three sets of stimuli were established according to the speakers: one set including 100 utterances of speaker 1 (S1), the other 95 utterances of speaker 2 (S2), and the third 80 utterances of speaker 3 (S3). Mean durations (with SD) of the utterances were 1709ms (599ms), 1015ms (473ms), and 1311ms (524ms) for S1, S2, and S3 respectively.

2.2. Perception Test

Ten Koreans (5 M + 5 F) and Ten Americans (5 M + 5 F) participated in the test. Most of them were students at Brown University. Their age varied from 23 to 29. The listeners were asked to perform two tasks: one task was, after listening an utterance, to decide whether the utterance sounded emotionally positive, neutral, or negative, by choosing one of the labels 'positive,' 'neutral,' and 'negative'; the other task was to rate how much emotion is expressed in the utterance on a 5-point

scale, which runs from 'no emotion' to 'extreme emotion'. The subjects listened ten utterances of each speaker in practice and rated the stimuli in six sessions individually (3 sets of stimuli x 2 tasks). The stimuli were presented in a random order and a three-second silence was given after each stimulus for the subject to respond. The tests took about forty minutes in total.

2.3. Results

A total of 550 responses, 275 responses of Koreans and of Americans, collected from the first task. Three possible responses, positive, neutral and negative, were computed in the data as '1,' '0,' and '-1' respectively. According to the definition of three emotional dimensions by Schlosberg [6], these ratings were considered as the estimation of emotional pleasantness on the 'Valence' axis.

Mean values of the Valence rating were evaluated by analyses of variance (ANOVA) regarding three factors, 'Listener' ('Koreans' and 'Americans'), 'Emotional category' ('Positive,' 'Neutral,' and 'Negative'), and 'Speaker' ('S1', 'S2', and 'S3').

The results showed that the overall Valence ratings were similar between the Korean and American listeners ($F(1,548)=0.034$, $p>0.05$). The Valence values were, however, significantly different among the 'Emotional categories' ($F(2,532)=102.41$, $p<0.01$) and the 'Speakers' ($F(2,532)=17.84$, $p<0.01$). In other words, the 'Emotional categories' (positive, neutral, and negative) could be distinguished in terms of Valence values such as 0.13, -0.15, and -0,51, on the rating scale with two ends, +1 and -1 (positive – negative). The three speakers, S1, S2, and S3, were also perceived differently with mean Valence values of -0.10, -0.19, and -0.34, respectively, showing that overall the utterances of S3 were perceived as more negative than those of the others. Generally speaking, the absolute values of Valence rating was much higher for negative emotion than for positive emotion, which was expected given that the expression of negative emotion was more explicit than that of positive emotion in our data.

Speaker	S1		S2		S3	
Listeners	Koreans	Americans	Koreans	Americans	Koreans	Americans
Positive	+0.47	+0.25	-0.12	+0.07	-0.04	+0.14
	*		*	*	*	
Neutral	+0.03	+0.05	-0.37	-0.30	-0.18	-0.08
	*	*		*	*	*
Negative	-0.75	-0.52	-0.40	-0.64	+0.03	-0.55

Table 1. Mean values of Valence ratings by Korean and American listeners for three emotions expressed by 3 speakers. The sign, '+' or '-,' indicates that listeners perceived the emotion as 'more positive' or 'more negative'.
The sign '*' indicate that the difference is statistically significant at the level of $p<0.05$.

Concerning the second task, 550 responses (275 of Koreans and of Americans) were entered with values of '1,' '2,' '3,' '4,' and '5'. In terms of emotional dimension, these ratings concern the estimation of emotional intensity on the 'Activation' axis, which runs from 'no emotion' to 'extreme emotion' in our test. Analyses of variance were performed on the data regarding the same three factors, 'Listener,' 'Emotional category,' and 'Speaker' as described above.

The results showed that the Activation rating was different between Koreans and Americans ($F(1,532)=10.43$, $p<0.01$). In fact, Americans rated relatively higher Activation values than Koreans. The three emotional categories were also different with Activation ratings such as 2.17, 2.40, and 3.62. The Activation values were also different for three speakers ($F(1,532)=6.52$, $p<0.05$): the speaker S1 received higher Activation values than the other speakers, S2 and S3.

	S1		S2		S3	
	Koreans	Americans	Koreans	Americans	Koreans	Americans
Positive	2.55	2.81	1.79	2.08	1.82	2.17
		*	*		*	*
Neutral	2.12	2.34	2.43	2.61	2.34	2.81
	*	*	*	*	*	*
Negative	3.90	3.60	3.51	3.54	3.45	3.56

Table 2. Mean values of Activation ratings by Korean and American listeners for three emotions expressed by three speakers. Higher values indicate higher emotional intensity perceived by the listeners.

2.4. Discussions

Activation theory predicts that the stimuli of positive or negative emotion would receive higher Activation rating values than neutral ones, producing a quadratic correlation of Activation and Valence ratings. This prediction works only for the speaker S1 in our data, not for S2 and S3. This asymmetry might be due to the fact that the negative emotion of sadness was dominant overall in our data and that the expression of positive emotion of S2 and S3 was not as explicit as that of S1.

As a consequence, a linear unidirectional correlation of the Activation and Valence ratings was found in our data (Pearson's Correlation Coefficient = -0.59, $p<0.01$), indicating that the more negative Valence value was rated, the higher Activation value was given.

The perceptual ratings of emotion on two major dimensions, the 'Valence' and 'Activation' dimensions, seems to be a useful method to examine a general tendency in the perception of emotion. Especially, when the perception test involves listeners with different cultural backgrounds, this general dimensional approach allows to avoid the potential problems of definition or labeling of specific emotions.

Some practical problems of this approach were also found in our data. As the listeners relied on subjective scaling of Activation and Valence level of emotion, there existed certain inconsistency among their ratings. Regarding the perception of neutral emotion, some listeners attributed a Activation values of '1' (labeled as 'no emotion') to the utterances, which they identified as neutral on the Valence dimension, whereas others rated a Activation value of '3' (labeled as 'some emotion') for the neutral utterances, in the sense that they sounded 'not very emotional'. Concerning the rating of Activation level of the positive and negative emotions, the choice of listeners varied from '1' to '5,' which means that the listeners indeed attributed the value '1' for the utterances of 'little positive or little negative emotion,' in spite of its label, 'no emotion'. This kind of conceptual confusion in the ratings seems to be responsible for the high Activation values of neutral emotion, which were

sometimes even higher than that of positive emotion. By the way, the perceptual rating behavior of listeners, the attribution of Activation level, '3,' to the utterances identified as neutral, seems to suggest the existence of emotionality even in neutral speech, as noted by Fónagy [2]. According to his notion of 'Double Coding,' every utterance contains linguistic aspect as well as emotional or stylistic aspect of the speaker. In this concept, neutral speech means just 'emotionally unmarked speech,' and emotions expressed in spontaneous speech, can be studied on a gradual scale, instead of using artificial labels.

3. EXPERIMENT 2

The purpose of this experiment is to extend the generality of the previous results on the perceptual differences among the listeners using different languages. The perceptual ratings of three groups of listeners, Koreans, French, and Americans, are compared on the Valence dimension. This experiment also examines the possible influence of lexical meaning on the Korean perceptual ratings and that of acoustic cues on the native and foreign ratings.

3.1. Speech Material

The stimuli of this experiment were extracted from the data of Experiment 1. Three sets of 15 stimuli were constructed, each set consisting of five utterances of three emotions, positive, neutral, and negative, of each speakers, S1, S2, or S3. The total number of stimuli was 45 (5 utterances x 3 emotions x 3 speakers). The selection of utterances was based on the mean Valence values of Korean and American ratings in the results of Experiment 1. In other words, five utterances, which received the most high values in positive Valence ratings for each speaker were chosen as the utterances of positive emotion. The utterances of negative emotion were selected in the same manner with the most high values of negative Valence ratings. The neutral utterances were the ones with the smallest Valence values, close to '0'. Mean durations (with SD) of utterances of the positive, neutral, and negative emotions were 1946ms (472ms), 1759ms (393ms), and 1881ms (526ms), respectively.

3.2. Perception Test

Ten of each Korean, American, and French subjects participated in the perception test. The Koreans and Americans were different from those in the previous test. They were students at Brown university in the age of twenties. Each of the three groups of listeners, Korean, American, and French, consisted of half male and half female subjects.

They were asked to perform the task of Valence rating, which was the same as in Experiment 1. After listening an utterance, the subjects decided which kind of emotion was expressed in the utterance, by choosing one of three emotions, positive, neutral, and negative. They responded after each utterance during three-second-silence. In this time, they did not rate the degree of emotional intensity (Activation rating).

The subjects performed the task individually in six sessions (3 sets of stimuli x 2 repetition) ; the utterances of one speaker were presented in each session and every utterance was rated twice in different sessions. The order of presentation of the stimuli was randomized within and across the sessions. Total duration of the test was about half an hour.

3.3. Results

450 responses were collected from each group of Korean, French, and American listeners, so 1350 responses in total (45 Stimuli x 10 Listeners x 3 Language groups). Three possible responses, positive, neutral, and negative, were computed as '+1,' '0,' and '-1' in the same way as in Experiment 1, and they were called 'Valence rating'.

According to two-factor mixed ANOVA, the Valence rating differed significantly by the within-subject factor, 'Emotion' ($F(2,174)=195.39$, $p<0.01$). The between- subject factor, 'Language,' also produced a significant effect on the Valence rating ($F(2, 87)=7.65$, $p<0.01$).

The Valence rating values were different for the positive, neutral, and negative emotions, 0.26, 0.02, and–0.72, respectively. Across all the three groups of listeners, the three emotions were distinguished in the same way that the highest, the middle, and the lowest Valence values were attributed to positive, neutral, and negative emotions respectively.

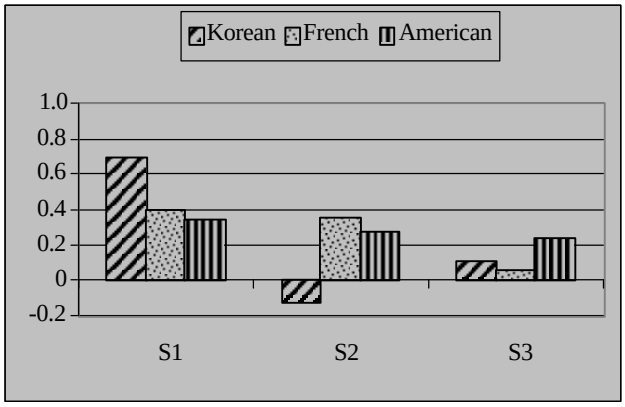

Figure 1. Mean values of Valence ratings by Korean, French and American listeners for Positive Emotion of three speakers, S1, S2, and S3.

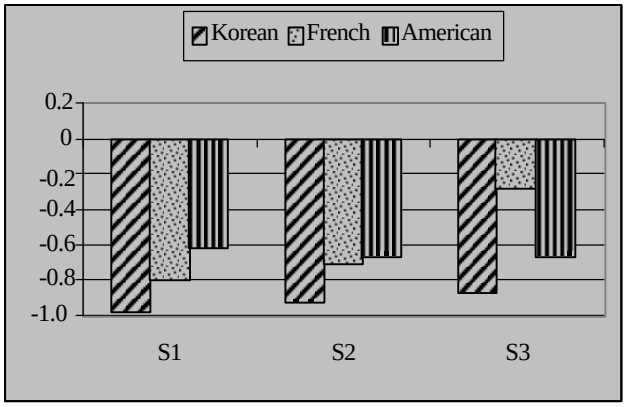

Figure 2. Mean values of Valence ratings by Korean, French and American listeners for Negative Emotion of three speakers, S1, S2, and S3.

However, the overall Valence rating of Korean, American, and French listeners were different with mean values of -0.26, -

0.11, and 0.06 respectively. The Korean and American listeners rated more negatively than the French listeners in general. This difference was, in fact, due to the least negative rating of French listeners in the case of S3, as seen in Figure 2. Again, as expected from the result of Experiment 1, the dominant emotional rating was negative, because of the high intensity of negative emotion expressed in our original data.

The three speakers, S1, S2, and S3, were perceived differently with mean values of Valence rating, -0.07, -0.13 and -0.22, respectively ($F(2, 87)=3.51$, $p<0.05$). A post-hoc analysis showed that only the difference between S1 and S3, was statistically significant. The Valence value of S2 was not significantly different from that of S1 or that of S3.

There was a significant interaction of the factors, 'Emotion' and 'Language,' ($F(4, 162)=2.65$, $p<0.05$); the listeners differed considerably in the ratings of positive emotion whereas they more or less agreed on the ratings of negative emotion. The interaction of the factors, 'Language' and 'Speaker,' was significant ($F(4, 81)=3.00$, $p<0.05$), indicating that the perceptual pattern of the two groups of listeners was different depending on the type of speaker.

3.4. Supplementary Test with Written Stimuli

In order to verify whether lexical meaning influenced the perceptual ratings of Korean listeners, the same utterances of stimuli, in a written form instead of in speech,. were presented to ten Korean subjects. The subjects made a judgement whether the written phrases might have been produced with a positive, neutral or negative emotion, and rated on the Valence scale of '1,' '0,' or '-1'. Analyses of variance showed that there was no significant effect of 'Emotion' on the Valence ratings ($F(2, 54)=0.44$, $p>0.05$) The mean Valence values were the same for positive and negative emotions, indicateing that the Koreans were not able to identify the emotions from the written phrases. This result indicates that the previous identification of positive and negative emotions was based on the acoustic cues of speech rather than the lexical meaning.

There still existed some influence of lexical information on the Korean perceptual rating although it was not statistically significant. Lexical meaning of the phrases of S1 provided at least the correct distinction of positive and negative emotions, while it induced inverse ratings for the phrases of S2 and S3. The negative Valence ratings by the Korean listeners for the speech stimuli of positive emotion in the case of S2 and S3 could be explained by this negative effect of lexical meaning, which was irrelevant to the American and French listeners.

	S1	S2	S3	Mean
Positive	+0.06	-0.18	-0.06	-0.06
Neutral	-0.26	-0.08	-0.06	-0.13
Negative	-0.08	-0.16	+0.06	-0.06
Mean	-0.09	-0.14	-0.02	-0.08

Table 2. Valence ratings for Positive, Neutral, and Negative emotions in the written phrases. * None of the differences between the values were statistically significant ($p=0.05$).

3.5. Acoustic information

In order to see the influence of acoustic cues on the perception of emotions, we measured 'Mean Fo' and 'Tempo,' of the utterances. Analyses of variance showed that the 'Mean Fo' was significantly different for three emotions with mean values of 220Hz, 212Hz, and 269Hz, for positive, neutral, and negative emotion respectively ($F(2, 42)=9.18$, $p<0.01$). Negative emotion had always significantly higher mean Fo than positive and neutral emotions. The mean Fo of positive emotion was higher than neutral emotion for S1 and S3, but not for S2. The relatively low mean Fo of positive emotion for S2 and S3, seemed to be responsible for their low Valence and Activation ratings of positive emotion in the previous results.

Mean values of 'Tempo' were 6.31, 5.86, and 6.23, for positive, neutral, and negative emotions; their differences were not statistically significant ($F(2,42)=0.78$, $p>0.05$).

	Mean Fo			Speaking Rate		
Speakers	S1	S2	S3	S1	S2	S3
Positive	252.3	200.5	208.8	6.78	6.54	5.62
Neutral	212.7	228.3	196.9	5.42	6.42	5.74
Negative	284.7	238.1	285.7	6.60	6.12	5.98

Table 3. Mean Fo (Hz) and tempo (measured by the number of syllables per second) for three emotions expressed by the three speakers, S1, S2, and S3.

Still, these acoustic measurements are not sufficient to account for the perceptual differences of American and French listeners. According to the visual inspection of data, the voice quality of emotional expression seems to play an important role. For example, the negative emotion often involvs creaky voice for the speakers, S1 and S2, but not for S3, which could explain the especially low Valence values by French listener for the negative emotion of speaker S3. It remains, however, unknown why the Americans were not influenced by the same reason, which needs further analyses with other parameters.

4. CONCLUSION

The present study showed that Koreans and American and French listeners agreed on the overall identification of emotions but that their perceptual ratings differed significantly depending on the type of speakers. The linguistic knowledge of native Korean listeners seemed to facilitate the use of acoustic cues in the identification of emotion. In the absence of the lexical information, the foreign listeners, French and American listeners, were still different in the use of acoustic cues, indicating the importance of cultural factor in the identification of emotions.

REFERENCES

[1] McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., and Ferrer, C. A. (1975), "Cross-cultural differences in the perception of the emotional content of speech," Dev. Psychol. 11,551-555.

[2] Fónagy, I. (1983), *La vive voix: Essais de psycho-phonétique*, Payot, Paris.

[3] Kramer, E. (1964), "Elimination of verbal cues in judgement of emotion from voice," *J. Abn. Soc. Psychol.* 68, 390-396.

[4] Scherer, K. R. (1974), "Acoustic concomitants of emotional dimensions : Judging affect from synthesized tone sequences," In Weitz, S.(Ed.), *Nonverbal communication*, New York: Oxford University Press.

[5] Scherer, K. R. & Oshinsky (1977), "Cue Utilization in Emotion Attribution from Auditory Stimuli", *Motivation and Emotion*, Vol. 1, N. 4, 331-346.

[6] Schlosberg, H. (1954), "Three dimensions of emotion", *Psychological Review* 61(2), 81-88.