Audio and Audio-visual Effects of a Short English Emotional Sentence on Japanese L2s' and English L1s' Cognition, and Physio-acoustic Correlate

Toshiko Isei-Jaakkola[1,3], Qinghua Sun[2] and Keikichi Hirose[1]

[1]Graduate School of Information Science and Technology, [2]Graduate School of Engineering
University of Tokyo, Japan, [3]Department of Speech Sciences, University of Helsinki, Finland
[1,2]{tijaakkola, qinghua, hirose}@gavo.t.u-tokyo.ac.jp, [3]toshiko.jaakkola@helsinki.fi

Abstract

The cognition test results of audio (A) and audio-visual (AV) effects on nine English emotions in a short sentence were compared to the physio-acoustic features of sound used for the cognition tests. Two groups of Japanese learners of English (JL2) and one group of English speakers (EL1) participated in these A and AV cognition tests. In the physio-acoustic analyses we used F0 and intensity contours and calculated the area of sentential patterns and three forms of distance: area-, average, and pattern-distance for each emotion. Similar patterns of the F0 and intensity contours might have been caused by the cognitive confusions among emotions; the relationships between the cognition tests and physio-acoustic analyses confirmed that there was not a strong correlation between them, intensity seeming to be more correlated to the cognition test results for A for both JL2 and EL1 than F0. EL1's correlation was higher than that of JL2.

1. Introduction

Emotions play an important role in speech communication. The acoustic realization of emotions in temporal and prosodic organization (see [4]) is extremely complicated. While a great number of experiments have been conducted to discover the acoustic parameters of emotions, how the cognition structures cope with acoustic parameters of even six basic emotions remain obscure ([7]). Even so, the research materials designed to identify the correlate between human cognition of emotions and the acoustic parameters and articulatory settings in producing emotions have been often limited to rather short utterances such as isolated words or emphatic words in a sentence, particularly in speech synthesis and recognition.

Emotions are also inevitably related to the culture (e.g., [6]) of the target language. Study [1] tested the cognition by Japanese learners of English (JL2) of ten English emotion words in dialogues recorded by four North Americans (EL1) using only sound. The overall correct ratio was 65% (Japanese *N*: 113, chance level 33.3%). Study [2] investigated JL2s' judgment (*N*: 259) both when they heard recorded sound (A) and when they heard and saw facial expressions (AV) simultaneously on the video, using a recorded short sentence uttered with nine emotions. The overall correct ratio was not very high in either test (A: 21%, AV: 53%), compared to that in the dialogues. Non-linguistic (basic) emotions had higher correct answer ratios than paralinguistic emotions. Like study [2], study [3] compared the cognition test by JL2 of English eight emotions, using English word (an interjection), using two methods. The results showed a significant difference from [2] between A and AV. Neither were patterns similar to those in [2] observed in these correct answer ratios of A and AV.

This study adds the results of English speakers' (EL1) cognition tests to the results for [2] for the purposes of comparison, questioning (1) whether there are differences between JL2s' and EL1s' cognition of English emotions uttered in a short statement. We shall also investigate (2) whether there are correlates between these correct answer ratios of the cognition tests and physio-acoustic features. In (2), we shall focus on two parameters: pitch and intensity contour patterns of English emotional sentences, and attempt to compare these with the results in (1), particularly by physio-acoustic distances between emotions.

Japanese is said to be a pitch-timed language, while English is stress-timed. We might predict that Japanese speakers depend on more F0 movement while their English counterparts rely more on intensity when judging English emotions.

2. Cognition Test

2.1. Methods

Nine emotion words were used for the cognition test: 'happiness', '(cold) anger', 'suspicion', 'surprise', 'sadness', 'fear', 'hatred', 'disappointment', and 'contempt'. The sentence used for recording purposes was "This is a pen". This was done to avoid the linguistic information affecting the results as much as possible, so that the subjects could concentrate on only non- and para-linguistic English emotions. The sentence was not written on the answer sheet. It was simultaneously recorded on both the DAT tape (A) and video tape (AV) and each emotion uttered twice in sequence by one British female informant (51 years old, university lecturer), based on her own reproduction of her emotions in the recording studio.

There were two groups numbering 149 and 110 for the test: JL2 and EL1. JL2, participating in Tests A and AV, consisted of male and female university students between 18 – 22 years old, majoring in various fields. There was no great difference in their English proficiency. Thirty-four EL1, consisting of both males and females aged between 18 – 64 years old, participated. They were from the U. S. A., U. K., Australia, Canada and New Zealand and their profiles varied. The testing method was a forced-choice (chance level 11.1%).

2.2. Results

Figure 1 illustrates the correct ratio by audio and audio-visual effects and by JL2 and EL1 according to each emotion word. The overall correct ratio of AV was higher in EL1 (59%) than in JL2 (47%), but that of A was lower in EL1 (31%) than in JL2 (38%). These ratios by EL1 were lower than those in [5]. The range between A and AV was very large, being 28% in

EL1 but only 9% in JL2. In JL2, the basic emotions such as ‘anger’, ‘sadness’, ‘surprise’, ‘happiness’, ‘fear’, and ‘hatred’, had higher correct answer ratios than paralinguistic emotions such as ‘disappointment’, ‘contempt’, and ‘suspicion’, in both tests A and AV. However, this was not true in EL1; ‘contempt’ had a higher correct ratio than some basic emotions in A and ‘disappointment’, ‘contempt’, and ‘suspicion’ did not have markedly lower correct answer ratios than the basic emotions in AV. The answer patterns from the highest to the lowest were similar in JL2 in both tests except for ‘anger’ and ‘happiness’ (AV) and relatively so in EL1 except for ‘contempt’ (A) and ‘happiness’ (AV) between A and AV. ‘Happiness’ showed a significant difference between the two tests, particularly for EL1 (*R* 73%: A = 21%, AV = 94%). The *R* for JL2 was 26% (A = 50%, AV = 76%). ‘Disappointment’ (*R* 43%), ‘suspicion’ (*R* 38%), ‘sadness’ (*R* 26%), ‘fear’ (*R* 26%), and ‘hatred’ (*R* 20%) also showed rather large range differences for EL1, but not for JL2 except for ‘happiness’ (*R* 26%).

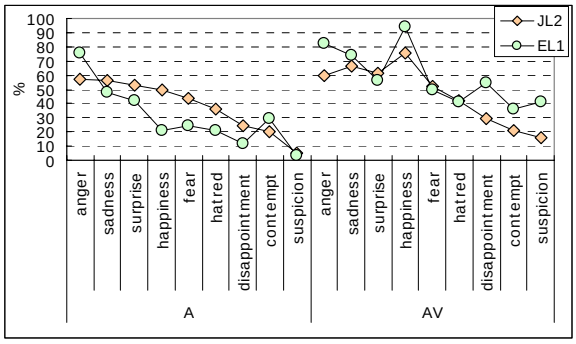

Figure 1: *Audio and Audio-visual effects.*

3. Test results and physio-acoustic features

We shall use only A to get physio-acoustic features. But, in comparing them with the cognition test results, we shall use the results from both A and AV.

3.1. Sentential patterns

Figures 2 and 3 illustrate pitch (hereafter F0) and intensity (dB) contours normalised in the time domain. ((In the figures, the values of Y-axis differ, but F0 height and intensity height, each which depends on each utterance for each emotion, were adjusted so that the height of the Y-axis looks the same on these figures. The duration was also normalised in X-axis so as to have 100 (%) for all contours.) The base value for F0 contours was 0 Hz, while it was 30 dB for intensity contours. These values will be used in the calculation of distances dealt with the sections below.) We added the values of two utterances for each emotion together and acquired the average values. A smoothing process was applied to F0 contours and to the intensity contours of these utterances, based on the piecewise 3rd order polynomials (cf. [5]), including the voiceless and silent part. These contours, F0 or intensity, show their own approximate patterns in accordance with each emotion, and thus a tendency for each emotion. In the sentences uttered for the test purposes, the F0 contours carries less voice information when part of the sentence is voiceless or silent but still the energy often remains. This implies that it is worth observing intensity movement as well as F0 movement. The cognitive confusions among emotions might have been caused by the similar patterns of these contours. The number of peaks might affect the subjects’ judgement as well.

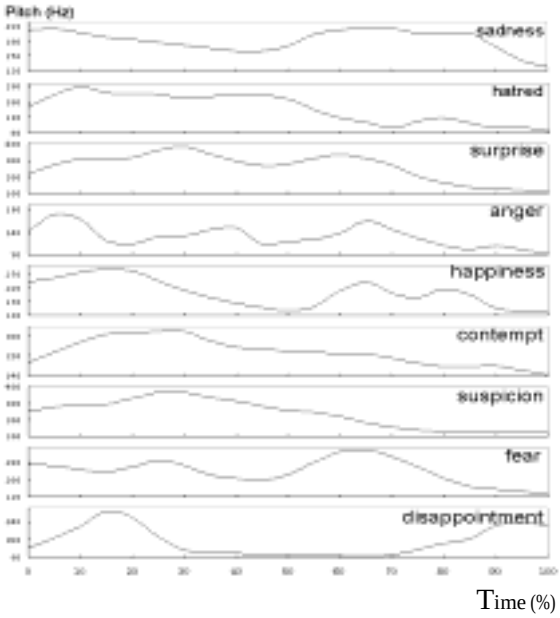

Figure 2: *The F0 (Hz) contour of each emotion, normalised in the time domain.*

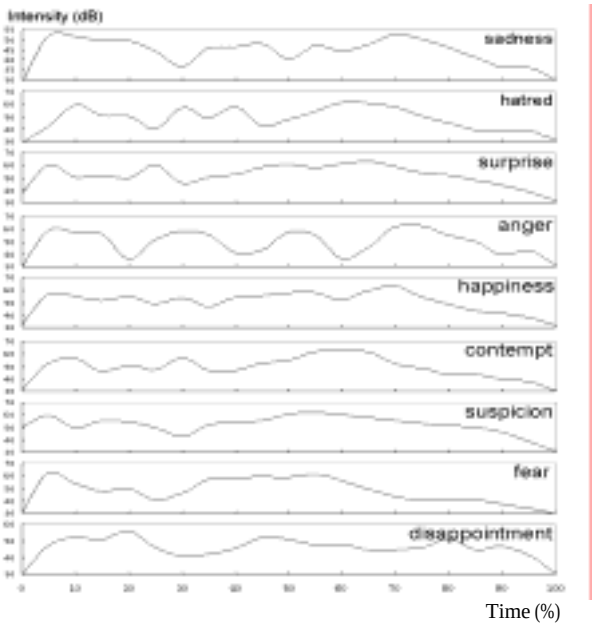

Figure 3: *The intensity (dB) contour of each* Emotion, *normalised in the time domain.*

3.2. Area of sentential patterns

It is very difficult to compare the relationship between emotions because the duration of emotional sentences differs. Thus we calculated the area of each contour corresponding to each emotion, including the sentence duration. It was found that F0 patterns did not correspond to the correct answer ratio patters at all, but the intensity patterns did so to those of A and AV by both JL2 and EL1 as a rough tendency. So, we do not report the details here.

3.3. Distance

We presupposed that the larger the value of the distance between emotions, the clearer the difference between emotions is, which might thus lead to higher correct answer ratios. In other words, the more different one emotion is judged to be, the longer the distance. Thus, if F0 or intensity plays a very important role in the cognition of each emotion, there could be a large correlation between the correct answer ratios and distance. We thus decided to calculate area-distance, average distance and pattern distance between emotions for F0 and intensity respectively.

3.3.1. *Area-distance between emotions*

Area-distance is the difference in the areas of F0 or intensity (dB) contours. We calculated the area-distance (D_a) between two F0 contours ($F_i(t)$ and $F_j(t)$) of each emotion using equation (1). First, we acquired the difference value between each emotion of nine emotions, e.g., between 'anger' and 'sadness', between 'anger' and 'surprise', between 'anger' and 'happiness' and so forth, according to each emotion, for both F0 and intensity contours respectively. And, we then obtained the average values (e.g., 'anger') of the differences (e.g., the other eight emotions than 'anger') in each emotion for both F0 and intensity respectively, according to each emotion. These average values were translated into Figures 4 (F0) and 5 (intensity) respectively (as bars in the Figs.), comparing the correct answer ratios (lines in the Figs). There seemed to be no correspondence between them in F0, while there was a slight correspondence between them in intensity for both A and AV by both JL2 and EL1.

$$D_a = \left| \sum_{t=0}^{T_i} F_i(t) - \sum_{t=0}^{T_j} F_j(t) \right| \quad (1)$$

where T is sentence duration.

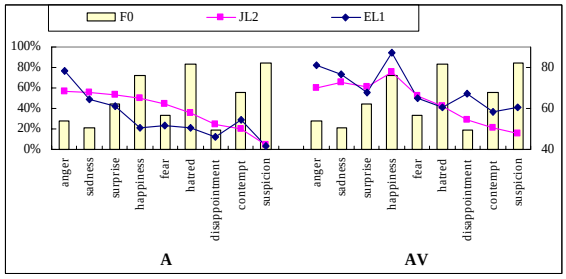

Figure 4: *A comparison between correct answer ratios and average area-distance of F0 calculated by equation (1).*

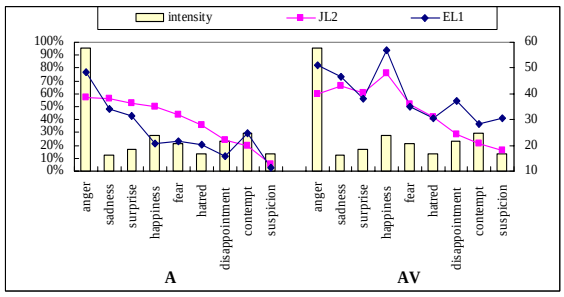

Figure 5: *A comparison between correct answer ratios and average area-distance of intensity calculated by equation (1).*

3.3.2. *Average distance between emotions*

Here we observed the distance between the patterns of each emotion, normalised in the time domain. The distance (D) between the two F0 contours ($F_i(t)$ and $F_j(t)$) of all emotions was calculated by equation (2). As in the calculation in area-distance, we obtained the average values of F0 and intensity according to each emotion, which were translated into Figures 6 (F0) and 7 (intensity) respectively (as bars in the Figs.), comparing the correct answer ratios (lines in the Figs). There seemed to be no correspondence between them in F0, while there was a slight correspondence between them in intensity for both A and AV by both JL2 and EL1.

$$D = \frac{\sum_{t=0}^{T} \left| F_i(t) - F_j(t) \right|}{T} \quad (2)$$

where T is sentence duration.

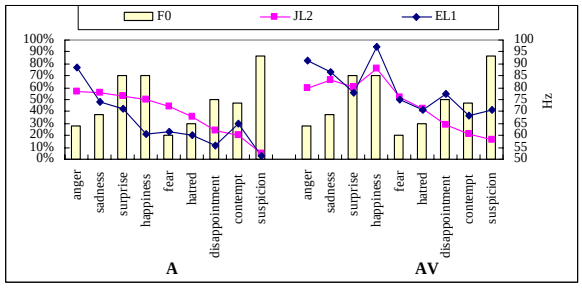

Figure 6: *A comparison between correct answer ratios and average distance of F0 calculated by equation (2).*

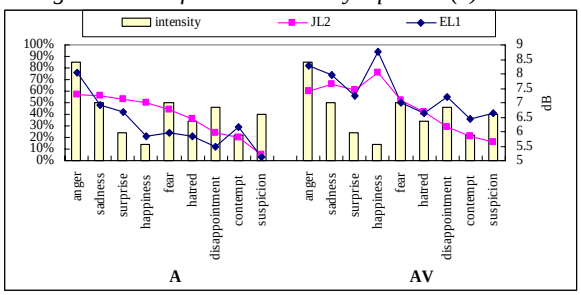

Figure 7: *A comparison between correct answer ratios and average distance of intensity calculated by equation (2).*

3.3.3. *Pattern-distance between emotions*

Pattern-distance means only the difference between the patterns of each emotion, ignoring the average height of the F0 contour. The pattern-distance (D_p) between the two F0 contours ($F_i(t)$ and $F_j(t)$) of all emotions was calculated using equation (3). As in the above two sections, the average values of F0 and intensity according to each emotion were translated into Figures 8 (F0) and 9 (intensity) respectively (as bars in the Figs.), comparing the correct answer ratios (lines in the Figs). There seemed to be no correspondence between them in F0, while there was a slight correspondence in intensity for both A and AV by both JL2 and EL1.

$$D_p = \frac{\sum_{t=0}^{T-1} \left| (F_i(t+1) - F_j(t+1)) - (F_i(t) - F_j(t)) \right|}{T-1} \quad (3)$$

where T is sentence duration.

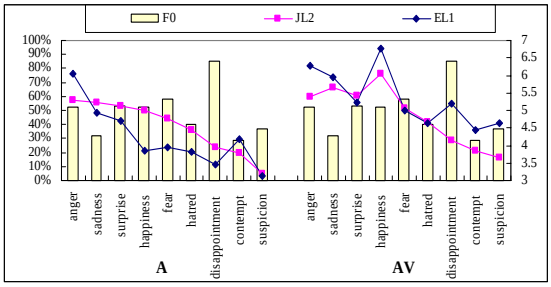

Figure 8: *A comparison between correct answer ratios and average pattern-distance of F0 calculated by equation (3).*

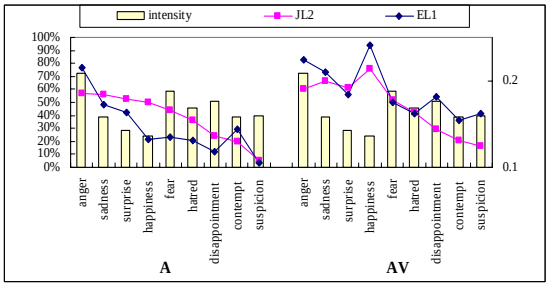

Figure 9: *A comparison between correct answer ratios and average pattern-distance of intensity calculated by equation (3).*

Comparing the distances in F0 with those of intensity, intensity had a similar tendency in pattern-distance and average distance concerning the correct answer ratios of the four basic emotions: 'anger', 'sadness', 'surprise', and 'happiness'. Thus, it might be possible to say that intensity plays more important role than F0 in judging emotions.

3.4. Correlation between cognition test and physio-acoustic features

In order to corroborate these findings, we examined whether there was a correlation between the correct answer ratios for A and AV by JL2 (= J) and EL1 (= E) in the cognition test and the overall average values for F0 and intensity in the physio-acoustic analyses (AD = area distance, D = average distance, PD = pattern distance). Table 1 shows the results. Coefficiency between them was not high. However, it seems that it is higher in intensity than in F0 for A for both JL2 and EL1. In addition, it is higher for EL1 than for JL2 in both F0 and intensity in A in all three forms of distances.

Table 1: *Coefficiency between cognition test results and physio-acoustic distances.*

		F0			Intensity		
		AD	D	PD	AD	D	PD
A	J	0.07	0.06	0.07	0.09	0.26	0.24
	E	0.12	0.13	0.14	0.41	0.59	0.55
AV	J	0.13	0.22	0.08	-0.13	0.06	0.08
	E	0.11	-0.11	0.20	-0.04	0.21	0.26

4. Conclusions

In terms of the cognition test results, the difference in overall correct answer ratios between the A and AV tests were much smaller in JL2 than in EL1. Three para-linguistic emotions used for the tests seem to be more difficult than non-linguistic emotions (basic six) for JL2 in both test conditions. The finding was that the correct answer ratios of A were the lowest of the tests using a short statement, a word ([3]), and dialogues ([1]). The order of the correct answer ratios was: dialogues > short statement > word in A, and word > short statement in AV. As far as A is concerned, this implies that the more linguistic information there is, the higher the correct answer ratios are. However, the converse was not true in AV. These results thus indicate that the cognition test results might differ according to the materials and methods used.

In the comparison of sentential patterns, it was suggested that similar patterns of the F0 and intensity contours might have been caused by the cognitive confusions among emotions. The relationships between these cognition tests and physio-acoustic analyses (in three forms of distance: area-distance, average distance, and pattern-distance) confirmed that there was not a strong correlation between them, intensity seeming to be more correlated to the cognition test results for A for both JL2 and EL1 than F0. EL1's correlation was higher than that of JL2, which suggests that EL1 depend more on intensity in judging emotions with a stress-timed language.

For further studies, the relationship between the distributions of the correct answer ratios in the cognition test and these distances should be compared for each emotion, and the third parameter: duration should be compared as well.

5. Acknowledgements

We would like to thank Ms. Yoshino Umegaki and Ms. Mariko Kanetou, English lecturers, of Rikkyo University and their students who supplied part of the Japanese data for this study.

6. References

[1] Isei-Jaakkola, T.; Neff, P., 2002. Japanese L2 Learner's Emotional Cognition of English Intonation. *Proc. of 7th Annual Congress of EPSJ.*

[2] Isei-Jaakkola, T.; Soga, S.; Barat, R., 2004. Audio and visual effects on English emotions by Japanese L2. *Handbook for the EPSJ Kanto Branch 6th Meeting*, 63-68.

[3] Isei-Jaakkola, T.; Sun, Q.; Hirose, K., 2005. Audio and Audio-visual Effects of English Emotional Word on Japanese L2's Cognition and the Acoustic Correlate. *Proceedings of SPECOM 2005*, 455-458.

[4] Laver, J., 1994. *Principles of Phonetics*, Cambridge University Press.

[5] Narusawa, S.; Minematsu, N.; Hirose, K.; and Fujisaki, H., 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proceedings of IEEE ICASSP*, Orlando, 509-512.

[6] Scherer, K. R., 2000. Cross-cultural investigation of emotion inferences from voice and speech: Implications of speech and technology. *Proceedings of ICSLP Beijing, 2*, 379-382.

[7] Shigeno, S., 2004. Recognition of vocal expression of emotion and its acoustic attributes. *The Japanese Journal of Psychology*, Vol., No. 6, 540-546.