Comparison between linguistic and affective perception of sad and happy – A cross-linguistic study

Caroline Menezes [1], *Donna Erickson* [2], *Clayton Franks*[1]

[1] Department of Health and Rehabilitative Services, University of Toledo, Toledo, Ohio, U.S.A.
[2] Showa Music University, Kawasaki, Japan
caroline.menezes@utoledo.edu, Clayton.Franks@rockets.utoledo.edu,
ericksondonna2000@gmail.com

Abstract

This paper is part of a larger study that examines cross-linguistic perception of sad and happy speech when the information is transmitted semantically (linguistic) or prosodically (affective). Here we examine American English and Japanese speakers' ability to perceive emotions in Japanese utterances. It is expected that native subjects will be better at perceiving emotion expressed semantically than non-natives because they have access to the semantic information. However, we see that Japanese listeners like American English listeners were not successful in discriminating emotion in the semantic content of the utterance. Both native speakers and non-native speakers could perceive that a speaker is sad or happy through the affective prosody. These results show that sad and happy are universally expressed the same way even in the auditory modality. Acoustic analysis showed differences in intensity, morae duration and F0 range for the linguistic, affective and neutral utterances and sad, happy and neutral emotions. Linguistic utterances revealed acoustic differences between the three emotional stages besides differences in the semantic context.

Index Terms: perception of emotion, affective, semantic, cross-linguistic, sad and happy

1. Introduction

Facial expressions of sad and happy were postulated to be universal from as early as Darwin (1872) and later methodically studied by Paul Ekman and his associates (e.g., Ekman, 1999). Banse and Scherer (1996) list the acoustic features that vary in speech depending on a particular emotion, indicating some universality even in vocal emotions. However, vocal emotion can be conveyed through various linguistic modalities: extra-linguistic - crying or laughing, linguistic - the use of words and language dependent prosody, and para-linguistic - the tone of voice and global/affective prosody (Banse and Scherer, 1996; Cosmides, 1983). When a listener perceives the emotional state of the speaker s/he attends to all of these modalities. However, most cross-cultural perception studies on vocal emotions have focused mainly on affective or paralinguistic communication, while extra-linguistic communication has been studied in the domain of visual communication. Little work has been done on the universal perception of linguistic emotion outside of irony or sarcasm where there is incongruity between the spoken word and affective prosody (Cappelli *et al.*, 1990). In a cross-linguistic study comparing English and Japanese perception of emotion in nonsense utterances Tickle (2000) found that Japanese subjects did not do as well as English subjects in determining vocal emotions in their native language. Interestingly, she also found that English listeners were better than Japanese listeners at perceiving emotions in Japanese utterances. She concluded that Japanese cultural codes bias them against perceiving vocal emotions.

Contradictory to Tickle (2009) other studies with cross-linguistic perception of paralinguistic information have shown that native speakers of the language generally perform better in identifying the emotional content of the utterance than non-native speakers who do not know the language, e.g., Sawamura et al., 2007. Along these lines, work by Huang et al. (2008) and Erickson et al. (2008a) suggest that there may be primary and secondary phonetic cues to emotion—the primary ones are those that are more or less equally shared by the different language groups, and the secondary ones, those that are peculiar to a certain language group. Specifically, increased F0 and increased intensity are significant cues for happy speech for Japanese, Chinese and American listeners. Phonetic cues associated with higher energy resonance frequencies (e.g., F4) may a type of secondary cue to emotion (e.g., Erickson et al.,(2008b) that are interpreted differently by different language groups.

In a study comparing the differences in brain activation for English and Chinese subjects perceiving Chinese semantically neutral utterances Gandour *et al*., (2003) reported brain lateralization in the processing of emotionally significant prosodic cues and also some dissociation in processing of linguistic prosody from emotional prosody in Chinese subjects. They also found differential brain activity for the different language groups which they interpret to be due to the functionally different processing of prosody by the two groups. Here linguistic prosody refers to intonation patterns that differentiate declarative sentences from questions, while emotional prosody refers to intonation patterns that differentiate sad, happy and angry.

Our study focuses on the cross-linguistic perception of sad and happy and compares the universality of linguistic and paralinguistic modalities in communicating these emotions. We chose sad and happy as they are two universal emotions. For the lack of space, we restrict our study here to understanding how American English speakers and Japanese speakers perceive emotion in Japanese utterances. Linguistic utterances are differentiated for sad, happy and neutral only in their semantic content while controlled for linguistic and affective prosody. Para-linguistically utterances differentiate emotions through the tone of voice and prosody (affective prosody). We expect that native listeners will do better than non-native listeners in both the linguistic modality and paralinguistic modality, while non-native speakers will fare better in the paralinguistic modality only given that they do

not have cues to the semantic content of the linguistic utterances.

2. Method

2.1. Speech material and recording

The auditory stimuli presented in this perception study were produced by four Japanese voice actors who read a list of sentences that made up our *linguistic* stimuli, *paralinguistic* stimuli and *neutral* stimuli. The *linguistic* sentences were semantically coded and contained words like "saddened", "burdened" to indicate sad, and "happy", "laughed" to indicate happiness (e.g., I am burdened with lots of work; I laughed till my sides split.). To control for prosodic effects in these utterances speakers were asked to read the sentences in a neutral voice with emphasis on the emotive word. The *neutral* sentences were neutral in content and were produced in a neutral voice, (e.g., I see the house with the blue tiles). The *paralinguistic* sentences were neutral in content but read with a sad or happy voice (e.g., Soon it will be May.). The *neutral* list contained completely different sentences from the *paralinguistic* sentences in order to avoid familiarity effect. There were eight different sentences that were produced twice to form 16 stimuli for the *linguistic happy, linguistic sad, affective happy, affective sad* categories. All four speakers did not produce all the sentences for these categories. 8 different sentences were repeated four times (each speaker produced each sentence) to form 32 *neutral* sentences. In total 160 sentences were used for the perception study. The recordings were made using Marantz PMD 660 at 48 kHz sampling rate (16bit accuracy) and saved onto a Compact Flash memory card. Later they were down-sampled to 16 kHz before conducting the acoustic analysis.

2.2. Perception test

The perception experiment was administered using the PRAAT program. 37 native American English undergraduate students (24 females and 13 males) and 39 Japanese undergraduate students (24 females and 15 males) listened to the 160 Japanese sentences on the computer and judged if the speaker was *sad*, *happy*, *neutral* or *other* in a four-way choice paradigm. We used the category *other* so that subjects were not forced to choose between the three categories presented. Subjects had the chance to hear each sentence up to a maximum of two times. All the American English subjects had no exposure to the Japanese language. Pivot tables were created to show the percentage of correct responses to incorrect response separated by gender, modality types, and emotion types.

2.3. Acoustic analysis

Acoustic analysis includes duration, intensity, pitch range, and fundamental frequency measurements. Duration values reported here are mora length, and were calculated by dividing the duration of the utterance by the number of morae. This was done because all sentences did not contain the same number of morae and there were no significantly long pauses in any of the emotional states studied here. The expected small effect of pauses in this calculation is directly accounted into the mora length. Mean intensity was measured across the entire duration of the utterance. Minimum F0 and maximum F0 were calculated for each utterance. Pitch range reported here is the difference between maximum and minimum F0 values calculated for each utterance. All acoustic measurements were made using PRAAT.

3. Results

3.1. Perception experiment

3.1.1. Japanese perceiving Japanese

Table 1shows the percentage of *sad, happy neutral* and *other* responses made by the English subjects' perception of emotion on Japanese utterances. We have separated the results by listener and speaker gender, and linguistic modality and emotion types. Japanese listeners were better at perceiving emotion in the *affective* utterances when compared to the *linguistic* utterances. *Linguistic* utterances were mostly judge to be *neutral* regardless that these utterances were semantically coded for the emotion. When not judged to be neutral they erred to *other* than any of the emotional categories. Interestingly, listeners sometimes tended to perceive men's *linguistically happy* utterances as *happy* but this was not the case for the female speakers. *Neutral* utterances were well perceived as *neutral*, however, when not sure Japanese subjects judged this category as *other*. Therefore, Japanese listeners did not pay attention to the meaning of the sentence but rather to the prosody of the utterances.

Table 1. *Percentage of Japanese subjects response separated by linguistic modality, emotion and listener and speaker gender*

Listener Gender			Female				Male			
Speaker Gender	Speech Modality	Emotion Type	h	n	o	s	h	n	o	s
Female	Affective	h	81	4	14	1	80	4	14	2
		s	2	6	14	78	0	8	13	80
	Linguistic	h	13	56	29	2	13	56	16	3
		s	4	62	24	10	3	72	19	7
	Neutral	n	.2	75	123	13	0	77	91	14
Male	Affective	h	85	4	11	.5	84	5	10	1
		s	.5	8	252	66	.4	13	23	64
	Linguistic	h	31	42	23	3	30	48	18	5
		s	12	40	33	15	9	46	28	16
	Neutral	n	.5	75	22	3	.4	84	13	3

3.1.2. English perceiving Japanese

Table 2, shows the percentage of *sad, happy neutral* and *other* responses made by the English subjects' perception of emotion on Japanese utterances. Here we see that Americans both males and female listeners were also better at judging *happy* and *sad* in the *affective* utterances than in the *linguistic* utterances. However, they did not do as well as Japanese listeners. Like the Japanese listeners, American listeners also perceived l*inguistic* utterances mostly as *neutral.* When *linguistic* utterances where misjudged American English subjects tended to be perceive them more as *happy* especially for the *happy* utterances than sad utterances. *Neutral* utterances were judged to be *neutra*l in emotion, however, it could also be perceived to be *sad* especially if the speaker was female.

Table 2. *Percentage of English subjects response separated by linguistic modality, emotion and listener and speaker gender.*

Listener Gender			Female				Male			
Speaker **Gender**	**Speech Modality**	**Emotion Type**	**h**	**n**	**o**	**s**	**h**	**n**	**o**	**s**
Female	Affective	h	70	5	20	6	67	8	24	.5
		s	.8	13	6	80	2	13	20	65
	Linguistic	h	41	42	14	2	27	57	11	6
		s	22	60	15	3	10	47	31	12
	Neutral	n	5	52	3	40	3	72	10	1
Male	Affective	h	72	7	10	11	73	8	17	3
		s	4	11	4	82	4	1	9	76
	Linguistic	h	52	35	6	7	27	54	7	11
		s	30	50	8	12	18	53	14	15
	Neutral	n	13	39	4	43	10	63	6	22

3.2. Acoustic measurements

Table 3 shows mean and standard deviation values for all the acoustic parameters measured. Looking at RMS or intensity values we see that *neutral* utterances have the lowest mean value. Univariate analysis show significant differences in RMS across speaker gender (F=10.146), speech modality (F=7.230), and emotion type (F=6.519) all at (sig<.05). The interaction between speech modality and emotion type was also significant (F=14.697, sig<.05). Tukey test also revealed that the RMS values were significantly different for all modalities (p<.05) – *linguistic*, *affective* and *neutral*. While Table 1 indicates higher RMS values for *affective happy* when compared to *affective sad*, post-hoc Tukey test show that RMS values are significantly different for all emotion types (p<.05) – *sad*, *happy* and *neutral*. Sad speech was found to have lower intensity even in Swedish (Abelin & Allwood, 2000). In general, happy speech had higher intensity followed by sad and then neutral,

Observing the mean values of mora duration in Table 1 we note that females generally produce longer morae than males which is confirmed by univariate analysis (F=24.572, p <.05). Though no significant difference is found for modality type there is significant difference among emotion types (F=4.285, p<.04). Generally, *linguistic happy* is longer in duration when compared to *linguistic sad* for both genders. Abelin and Allwood (2000) also show longer word durations for happy when compared to sad for Swedish speakers. Gender differences can be seen and univariate analysis reveals significant interaction between gender and modality type (F=7.485, p=.007). Females have longer durations for *neutral* while males have longer durations for *affective sad*. However, in general, sad speech had the longest morae duration, followed by neutral then happy speech.

Table 1 shows that *affective sad* has the smallest range of F0, followed by *neutral*. *Linguistic* utterances were produced with emphasis on the emotive word and these utterances tend to show the largest range of fundamental frequency movement. However, even though speakers were asked to emphasize the emotive word they showed smaller F0 range for *sad* utterances when compared to *happy* utterances across gender. In general, *sad* utterances were produced with a lesser F0 range than *happy* utterances for both *affective* and *linguistic* modalities. Further, univariate analysis show that the range of fundamental frequency was significantly different across gender (F=155.80), modality type (F=15.040), and emotion type (F=93.252) all at p<.001. Also there is significant interaction between modality and gender, and modality and emotion type (F=15.0, p<.05). Of the modality types, the range values for *linguistic* utterances were significantly different from *affective* and *neutral* utterances (Tukey test p<.001). Post-hoc tests comparing differences between emotion types show that happy utterances were produced with significantly larger F0 range than sad or neutral (p<.001). Generally, our results show that happy speech had the largest pitch range, followed by sad and then the unemotional neutral speech.

In general, *happy* utterances have higher min. F0, which is especially true in *affective* speech (see Table 1). *Neutra*l had the low min. F0 values for both genders. In Table 1 we see that *happy* speech also shows higher max. F0 values when compared to *sad* for both modalities even though largest variations (between emotion types) are seen for *affective* speech. We also see gender differences for max. F0 values where *neutral* speech is generally lowest for males but for females *affective sad* has the lowest max. F0 values. In general, we find that *happy* speech is produced with higher fundamental frequency than *sad* or *neutral* speech.

These results show that *sad*, *happy* and *neutral* speech are clearly differentiated by acoustic features like intensity, fundamental frequency, range and morae duration. Happy speech in general had higher intensity, pitch range, and pitch values with shorter morae durations when compared to sad speech across the different linguistic modalities. We also see that *linguistic* and *paralinguistic* information varies along these same acoustic parameters. Sad and happy linguistic utterances were produced with the same prosodic structure and were supposed to be differentiated only semantically. However, the surprising results show that speakers still manipulated prosodic features like intensity, duration, F0 range, and maximum F0 values depending on emotion type.

4. Discussion

Our results show that sad, happy and neutral utterances are produced by varying the acoustic parameters of intensity, morae length, pitch range and pitch values. Happy speech was produced with larger intensity, shorter morae, larger range and higher pitch values in general, when compared to sad speech. Neutral utterances were generally produced with the least intensity, lowest range, minimum pitch values and intermediate morae length. While affective utterances showed the most variation in these acoustic parameters linguistic utterances showed similar manipulation but to a lesser extent.

On the perception of emotions, Japanese subjects were better at perceiving emotion in their own utterances than American English, which contradicts the earlier findings of Tickle (2000). In this study we used real sentences and this could be the reason for the different findings. It was surprising however, to find that Japanese listeners did equally bad as their American counterparts in judging emotion in the linguistic utterances. American listeners had no exposure to the Japanese language and therefore, no clue to the semantic context of the linguistic utterances, but it appears that Japanese listeners also do not pay attention to sentence meaning when judging emotion. They tend to rely mainly on voicing cues. This is even more surprising since in the acoustic analysis we saw that Japanese speakers produced the

linguistic utterances acoustically different depending on the emotion.

However, looking into the error responses of the linguistic utterances revealed interesting cultural difference. Japanese listeners tended to choose the unknown category "other" when they misjudged the linguistic utterances. American English subjects were better at perceiving the linguistically happy utterances than the linguistically sad utterances especially for male voices. They also erroneously judged female neutral utterances as sad probably due to the lower intensity and softer voice associated with sadness. English listeners who did not have access to the semantic content of the utterances focused their attention on the prosodic content. The linguistic utterances differed from neutral utterances in the placement of emphasis on the emotive word. Emphasis often involves increase in pitch range, and intensity which is similar to happy speech. Speakers rely on prosody in situations where listener's processing has limitations (Capelli, Nakagawa, and Madden, 1990).

In conclusion the results of this study tend to indicate that listeners judge vocal emotion by attending to the affective prosody of the utterance rather than meaning of the utterances regardless of cultural groups. However, this need to be further tested by having Japanese and English listeners judge emotion on English utterances and also by extending this study to other cultural groups.

Table 3 *Mean (SD) values for acoustic measurements separated by gender, modality and emotion.*

	Modality	**Emotion**	**Intensity in db**	**Mora Duration in ms**	**Range F0 in hz**	**Minimum F0 in hz**	**Maximum F0 in hz**
Female	Linguistic	Happy	77.99(1.2)	147.25(21.4)	222.88(40.8)	155.31(35.0)	378.19(39.4)
		Sad	78.38(1.3)	148.19(22.8)	200.56(54.1)	140.69(43.1)	341.25(25.2)
	Affective	Happy	80.45(2.8)	135.43(17.1)	218.81(72.0)	200.63(41.5)	419.44(63.9)
		Sad	78.66(2.0)	144.94(16.2)	91.94(32.6)	151 (28.6)	243 (25.7)
	Neutral	Neutral	75.59(2.1)	155.06(19.0)	149.25(55.3)	166.5(31.5)	315.75(61.2)
Male	Linguistic	Happy	77.13(2.7)	120.25(16.7)	105.25(32.8)	85.56(10.3)	190.81(41.4)
		Sad	77.79(2.9)	125.75(23.1)	80.38(21.5)	89.94(14.5)	170.31(27.8)
	Affective	Happy	80.08(3.6)	128.56(30.9)	146.13(29.8)	105.44(27.0)	251.56(36.1)
		Sad	76.57(2.0)	144.5(24.8)	39.44(15.7)	83.44(8.2)	122.88(21.1)
	Neutral	Neutral	73.81(1.2)	129.50(21.1)	83.63(17.3)	82.19(7.5)	165.81(12.7)

5. Acknowledgements

This study was funded by University of Toledo – URAF grant for Visiting Faculty Research Program awarded to the first author. We are grateful to the undergraduate students of clinical phonetics for their assistance in data analysis.

6. References

Abelin, Å., Allwood, J., "Cross linguistic interpretation of emotional prosody". ISCA workshop on Speech and Emotion. Newcastle, Northern Ireland,110–113, 2000.

Banse, R., Sherer, K.R., "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology 70 (3), 614–636, 1996.

Bryant, G.A., and Fox Tree, J.E., "AcousticRecognizing verbal irony in spontaneous speech", Metaphor and Symbol, 17(2): 99-117, 2002.

Capelli, C.A., Nakagawa, N., & Madden, C. M., "How children understand sarcasm: The role of context and intonation". Child Development, 61:1824-1841, 1990.

Cosmides, L., "Invariances in the acoustic expression of emotion during speech". Journal of Experimental Psychology: Human Perception and Performance, 9, 864-881, 1983.

Darwin, C., "The expression of the emotions in man and animals." London: John Murray, 1872.

Ekman, P., "Facial expressions." In Dalgleish, T., & Power, M.,Handbook of Cognition and Emotion. New York: John Wiley & Sons Ltd, 1999.

Erickson, D., Rilliard, A., Shochi, T., Han, J., Kawahara, H., and Sakakibara, K., "A cross-linguistic comparison of perception to formant frequency cues in emotional speech." *COCOSDA, Kyoto, Japan,* 163-167, 2008b.

Erickson, D., Huang, C-F., Shochi, T., Rilliard, A., Dang, J., Iwata, R., and Lu, X., "Acoustic and arituclatory cues for Taiwanese, Japanese and American listeners' perception of Chinese *happy* and *sad* speech." *Proc. ASJ '2008 Fall Meeting,* 1-Q-14, 2008a.

Gandour, J., Wong, D, Dzemidzic, M., Lowe, M., Tong, Y., and Li, J. X., "A cross-linguisitc fMRI study of perception of intonation and emotion in Chinese", Human Brain Maping, 18:149-57, 1999.

Huang, C. F., Erickson, D., and Akagi, M., "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners." *Acoustics2008, Paris,* 2317-2322, 2008.

Sawamura, K., Dang, J., Akagi, M., Erickson, D., Li, A., Sakuraba, K., Minematsu, N., and Hirose, K., "Common factors in emotion perception among different cultures." *Proceedings of International Conference of Phonetic Science, Saarbrucken, German,* pp.2113-2116, 2007.

Tickel, A., "English and Japanese speaker's emotion vocalization and recognition: A comparison highlighting vowel quality", SpeechEmotion-2000, 104-109, 2000.