Cross-linguistic cross-modality perception of English sad and happy speech

Caroline Menezes [1], *Donna Erickson* [2], *Jonghye Han* [3]

[1] Department of Health and Rehabilitative Sciences, University of Toledo, Toledo, Ohio, U.S.A.
[2] Showa Music University, Kawasaki, Japan
[3] Korea University, Seoul, Korea

caroline.menezes@utoledo.edu,ericksondonna2000@gmail.com, jonghyehan@gmail.com

Abstract

This paper is a cross-cultural perception study of speech emotions of English utterances by American, Japanese and Korean listeners. The perception of sad and happy speech conveyed through linguistic modality (semantic) and affective modality (prosody) is tested to understand how native and non-native listeners comprehend the speaker's emotion. It is expected that native subjects would be better than non-natives at perceiving emotion expressed in both modalities because of their competence in accessing the semantic information as well as emotional prosodic information. Results reveal that in general, Americans perceive emotion in English better than Japanese and Koreans. However, native listeners and non-native Japanese listeners are more successful in discriminating emotion in affective and neutral utterances. Korean speakers are better at perceiving emotions in linguistic utterances. This could be due to English being taught as a second language in many countries. Our findings also indicate that listener's choice of modality processing is based on emotion types. *Happy* utterances are better perceived in the affective modality, while *sadness* is better perceived in the linguistic modality. Results also show that females are better at judging emotion by affective prosody, while males need to heed the semantic coding of emotion. *Happy* utterances are better perceived by males while *sad* utterances are better perceived by females. Females in general are better at perceiving emotions than males for all language groups.

Index Terms: perception of emotion, affective, semantic, cross-cultural, sad and happy, gender prototypes

1. Introduction

Speakers of a language use both linguistic (semantic) and paralinguistic (acoustic) modalities in vocally communicating emotional information. It is known that acoustic changes related to voice pitch, duration, intensity, tempo and voice quality affect listeners' perception of the emotional meaning of an utterance [1], [2], [3], [4]. In [2] we show that when Japanese speakers are instructed to read sentences using a "linguistic modality", (sentences that contain *happy* or *sad* words), the acoustic characteristics of their sentences are similar to their "affective" utterances (emotion not implied in the words but conveyed only through affective prosody). That is, both types of *happy* productions (linguistic and affective) are shorter in moraic duration, louder in intensity and higher in pitch compared to those of *sad* utterances, which are longer, softer and lower in pitch. For the paralinguistic (affective) expressions of *happy* and *sad* sentences however, the measured values are more extreme than those for the linguistic mode of expression.

The acoustical concomitants of emotion in both linguistic and affective utterances helped the naïve listeners of Japanese (American English subjects) to perceive the intended emotion of both linguistic and paralinguistic modalities of Japanese sentences. There was a definite advantage for native speakers (Japanese) who performed better in determining the vocal emotion. This finding conforms with a number of studies that report a universality in acoustic cues expressing vocal emotions, and that native listeners are better able to identify emotion in their own language than non-native [5], [6], [7].

However, there are studies that tend to show the opposite: that linguistic knowledge interferes with emotional processing. A study [8] compared English and Japanese perception of emotion in nonsense utterances, and reported that English listeners did better than Japanese ones, not only with the nonsense syllables, but also with Japanese utterances. It is also reported that Korean listeners did more poorly than Japanese and English listeners in identifying the emotions of very short vowels/words extracted from emotional Korean speech [9]. They reported that Korean listeners who did not know they were listening to Korean, performed similar to non-native listeners, and better than those Korean listeners who knew the utterances were in their native language. These studies suggested that native listeners used a linguistic-mode for identifying the emotional content of an utterance, while non-native listeners used prosody-only mode. It would seem thus, that there is an interaction between linguistic and paralinguistic expressions of emotion that needs to be explored further.

Our study focuses on the cross-linguistic perception of linguistically and affectively expressed emotions. Specifically, we ask American, Japanese and Korean listeners to indicate what emotion (*happy*, *sad*, *neutral* or other) they perceive when listening to a set of English utterances which vary in linguistically (semantic only) expressing happiness, sadness or neutral emotion, and a different set of utterances produced with three different affective prosodies: *happy*, *sad*, and *neutral*. We would expect that native listeners will do better than non-native listeners in both the linguistic modality and paralinguistic modality because they can comprehend the emotion of the utterance both by the affective prosody as well as the semantic content of the utterances. Non-native speakers were predicted to fare better in the paralinguistic modality than the linguistic modality, given that there would be limited access to the semantic content of the linguistic utterances. English in the last few decades is being taught in both Japan and Korea and these non-native speakers have different levels of mastery. It is assumed that these non-native listeners might face difficulties from the linguistic modality, leading to a dichotomy in perception of linguistic utterances.

2. Method

2.1. Speech material and recording

The auditory stimuli presented in this perception study were produced by four English student actors who read a list of sentences that varied in modalities - *linguistic*, *affective* and *neutral*, and emotion types – *happy*, *sad* and *neutral*. The *linguistic* sentences were semantically coded and contained words like "hurt", "cried" to indicate sadness, and "happy", "laughed" to indicate happiness (e.g., I cried my heart out. I was happy to meet my friend.). 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., He is wearing a white shirt.). The *affective* sentences were neutral in content but read with a *sad* or *happy* affect (e.g., Soon it will be May.). Each speaker produced eight *linguistic happy*, *linguistic sad*, *affective happy*, *affective sad* and *neutral* utterances. The *neutral* list contained completely different sentences from the *affective* sentences in order to avoid a familiarity effect. 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. 63 native American English undergraduate students (52 females and 11 males), 57 Japanese undergraduate students (44 females and 13 males) and 42 Korean (31 females and 11 males) undergraduate students listened to 157 of the 160 English sentences on the computer using headphones (3 sentences were omitted from the study because of their audio quality). Listeners judged if the speaker was *sad*, *happy*, *neutral* or *other* in a four-way choice paradigm. The response category *other* prevents the inflation of correct guesses. Subjects had the chance to hear each sentence up to two times. Korean and Japanese subjects were students of English at the time of the experiment and had more than five years of English training. 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 included duration, intensity, pitch range, fundamental frequency and tempo measurements. Duration values were calculated as length of utterance in milliseconds divided by the number of syllables/utterance (syllable duration). This was done because all sentences did not contain the same number of syllables and there were no long pauses in any of the emotional states studied here. The expected small effect of pauses in this calculation is directly accounted into the syllable length. Mean intensity was measured as average intensity calculated over the duration of the syllable nuclei in an utterance. Likewise, average pitch was calculated as the mean pitch of all syllable nuclei over the utterance. This was done to avoid effects of variable consonants in each utterance. Minimum and maximum F0 were calculated as the minimum and maximum fundamental frequency within each utterance. Pitch range reported here is the difference between maximum and minimum F0 values calculated for each utterance. Tempo was calculated as the number of syllables produced per second. All acoustic measurements were made using PRAAT.

3. Results

3.1. Perception experiment

The perception of speaker emotion in English utterances was significantly different across all cultural groups according to Chi-square analysis, $\chi^2(2, N = 25920) = 426.63$, $p<.001$, with Americans (56% correct) performing slightly better than Koreans (45%) who were slightly better than Japanese (41%). Also, across all languages, females (50%) marginally outperformed males (42%) in correctly perceiving the speakers' emotion, $\chi^2(1, N = 25920) = 105.25$, $p<.001$. Cross tabulation with gender and modality revealed that females were significantly better at comprehending the emotion in affective utterances (Females=53%, Males=38%), while males tended to determine the speaker's emotion from the semantic coding in linguistic utterances (Males=48%, Females=40%) $\chi^2(1, N = 25920) = 149.75$, $p<.001$ (affective), $\chi^2(1, N = 25920) = 45.64$, $p<.001$(linguistic). Cross tabulation show significant effects for gender and emotion across all modalities, males (Males=53%, Females=45%) performing slightly better on happy utterances [$\chi^2(1, N = 25920) = 30.58$, $p<.001$] and females (Females=47%, Males=38%) better at perceiving sad utterances [$\chi^2(2, N = 25920) = 81.14$, $p<.001$]. Performance specific to each language group is discussed next in further detail.

3.1.1. English perceiving English

Table 1 reveals that neutral utterances were correctly perceived as *neutral* by most listeners (69%) across both genders. Chi-square analysis pooled across males and females also reveal that English listeners perform slightly better when attending to affective utterances (63% *vs* 40%) than linguistic utterances, $\chi^2(2, N = 10080) = 700.42$, $p<.001$, as seen in Table 1. Post-hoc Goodman & Kruskal tau conducted on the Pearson Chi-Square values confirm that gender differences seen in Table 1 for *neutral*, *affective sad*, and *linguistic happy*

Table 1. *Percentage of English listeners' response separated by linguistic modality, emotion and listener gender. Bold cells indicate high percentage values for each emotion.*

Listener Gender		Female				Male			
Modality	Emotion	H	N	O	S	H	N	O	S
Affective	Happy	**78**	11	11	0	**78**	13	8	1
	Sad	5	37	11	**47**	5	29	11	**54**
Linguistic	Happy	29	**57**	5	9	**45**	41	6	9
	Sad	1	40	10	**49**	4	36	10	**51**
Neutral	Neutral	3	**74**	3	19	9	**64**	5	21

were significant. Males were better at perceiving *linguistic* utterances, while females were better judging *neutral utterances* (all at p<.001). In general, English listeners were best at perceiving *affective happy* utterances (78%). *Affective sad* utterances were mostly perceived as *sad* but a high percentage of listeners erroneously judged these utterances as *neutral*. The reverse was true for the *linguistic* modality: *sad* utterances had a higher chance of being perceived correctly as *sad*, while *happy* utterances were mostly confused with *neutral* emotion, even though these utterances were semantically coded for the emotion.

3.1.2. Japanese perceiving English

Both Japanese males and females, like Americans, were best able to perceive *happy* emotion in the affective domain. They were also good at perceiving *neutral* utterances but the

Table 2. *Percentage of Japanese listeners' response separated by linguistic modality, emotion and listener gender. Bold cells indicate high percentage values for each emotion.*

Listener Gender		Female				Male			
Modality	Emotion	H	N	O	S	H	N	O	S
Affective	Happy	**59**	20	13	7	**52**	15	19	13
	Sad	14	**35**	15	**35**	16	**34**	20	30
Linguistic	Happy	34	**38**	10	18	**37**	28	12	24
	Sad	13	33	16	38	16	28	12	**44**
Neutral	Neutral	14	**46**	15	26	18	**43**	14	26

overall percentage values were lower than the Americans. Post-hoc Goodman & Kruskal tau conducted on the Pearson Chi-Square values confirm that as seen in Table 2 Japanese females were better than males in perceiving emotion in *neutral*, *affective sad*, and *linguistic sad*. *Affective sad* utterances were likely to be confused as ne*utral* by most Japanese listeners, especially males. Again similar to English listeners, Japanese listeners had less difficulty discriminating *linguistic sad* utterances as *sad*, but *linguistically happy* utterances were confused with *neutral* emotion (particularly by females). Pearson Chi-square analysis pooled across gender also reveal that Japanese listeners, like the Americans, perform slightly better when attending to affective utterances (46% *vs* 37%) than linguistic utterances, $\chi^2(2, N = 9120) = 69.23$, $p<.001$, as is evident also in Table 2.

3.1.3. Koreans perceiving English

In deviation from the Americans and Japanese, the Koreans performed best in the *neutral* modality, but note that still the Americans had the highest percentage correct responses in this category. Pearson Chi-square analysis pooled across gender

Table 3. *Percentage of Korean listeners' response separated by linguistic modality, emotion and listener gender. Bold cells indicate high percentage values for each emotion.*

Listener Gender		Female				Male			
Modality	Emotion	H	N	O	S	H	N	O	S
Affective	Happy	**56**	29	13	2	**45**	28	23	3
	Sad	3	36	12	**49**	6	**44**	16	34
Linguistic	Happy	30	**41**	9	20	34	**40**	12	13
	Sad	1	24	12	**63**	3	27	14	**57**
Neutral	Neutral	3	**67**	7	23	7	**69**	5	19

also reveal that, unlike the Americans and Japanese listeners, Korean listeners perform slightly better when attending to *linguistic* utterances (50% *vs* 34%) than *affective* utterances, $\chi^2(2, N = 2688) = 196.53$, $p<.001$. Of the *linguistic* utterances, *sad* utterances were more likely to be correctly perceived. Erroneously, *linguistically happy* utterances were most often perceived as *neutral* by these listeners with a smaller percentage of *happy* responses. *Affective happy* utterances were correctly perceived as *happy* but *affective sad* utterances showed a gender preference, such that female listeners heard them as *sad* but males as *neutral*. Post-hoc Goodman & Kruskal tau conducted on the Pearson Chi-Square values confirm that as seen in Table 3, Korean females were better than males in perceiving emotion in *affective sad*, *affective happy* and *linguistic sad*. Males performed better in *neutral* and *linguistic happy* utterances (all at p<.001).

3.2. Acoustic measurements

Table 4 shows mean and standard deviation values for the acoustic parameters measured. Looking at RMS or intensity values, we see that *neutral* utterances have the lowest mean value. Univariate analysis shows significant effects for emotion type $F(1,157)=10.52$ (sig<.001). Tukey HSD test also revealed that the RMS values were significantly higher in *affective* utterances when compared to *neutral* utterances ($p<.023$) in our stimuli. *Linguistic* utterances were not significantly different from *affective* and *neutral* utterances. This could be the result of speakers emphasizing the emotional word in the *linguistic* utterances, which possibly affected the mean values. Post-hoc Tukey HSD test reveal that RMS values are significantly higher for *happy* compared to *sad* and *neutral* utterances ($p=.003$, $p=.002$ respectively). There were no significant difference in intensity between *sad* and *neutral* utterances.

Observing the mean values of syllable duration in Table 4, *affective* syllables had generally shorter durations when compared to *linguistic* and *neutral* utterances (not significant). *Sad* speech had the longest syllable duration, while *neutral* and *happy* syllables were approximately equal in length.

Table 4 also reveals that pitch range was different across all modalities. Univariate analysis predicts that these differences were significant ($F(1,157)=5.27$, $p=.023$) with *affective* utterances showing higher pitch range than *linguistic* and *neutral* utterances. There was also significant interaction between modality and emotion type ($F(1,157)=3.97$, $p=.05$). *Affective sad* utterances had smaller pitch range than *happy* utterances ($p=.015$).

As seen in Table 4, *happy* utterances are spoken faster than *sad* utterances for both linguistic and affective utterances (greater number of syllables/second). However, these differences were not significant.

Univariate analysis of average pitch shows significant effects for modality and emotion type ($F(1,157)=10.7$, $p<.001$; $F(1,157)=15.9$, $p<.001$; respectively). The interaction effect of modality and emotion was also significant ($F(1,157)=4.93$, $p=.028$). Post hoc Tukey HSD tests further show that a*ffective* utterances had significantly higher mean F0 when compared to *linguistic* and *neutral* utterances ($p=,003$, $p=.018$ respectively). Mean F0 for *linguistic* utterances was not significantly different from *neutral* utterances. Post-hoc Tukey HSD test also revealed that *happy* utterances were significantly higher in pitch when compared to *sad* and *neutral* utterances (both p<.005). Mean F0 of *sad* and *neutral* utterances were not significantly different.

These results show that *sad*, *happy* and *neutral* speech are clearly differentiated by acoustic features of intensity, average fundamental frequency, pitch range and syllable duration. Speech rate or tempo did not vary across modality and/or emotion types. In general, *affective happy* utterances were higher in intensity, shorter in syllable duration; higher pitch range, higher average pitch than *affective sad* utterances. *Linguistic happy* utterances differed from *sad* utterances in being shorter in syllable duration and lower in pitch range. *Neutral* utterances were similar to *affective* and linguistic *sad* utterances in having lower intensity and lower mean fundamental frequency, but were similar to *linguistic happy* utterances in having higher pitch range. Thus, acoustically,

English speakers clearly differentiated the linguistic and affective utterances. Linguistic sentences were similar to neutral sentences, except for the higher intensity and pitch range values due most likely from emphasizing the emotive word.

4. Discussion

Our results show that *sad*, *happy* and *neutral* emotions in English utterances are produced by varying the acoustic parameters of intensity, pitch range and pitch values. *Happy* speech was produced with larger intensity, larger range and higher fundamental frequency values when compared to *sad* speech. Neutral utterances were generally produced with the least intensity, lowest range, and lowest pitch values.

As in the earlier study [2], these results also indicate that native listeners were best when it came to perceiving speech emotion in their native language. Here the Americans were significantly better than Koreans who were significantly better than Japanese in perceiving English emotions. Surprisingly, we find that the American speakers, like the Japanese in the earlier study [2], were best at perceiving emotion in the affective modality. The affective modality conveyed the speaker's emotion only in the affective prosody. When the speaker's emotion was expressed as words, American listeners had the greatest difficulty understanding the correct emotion. The Japanese subjects in this experiment behaved similar to the Americans, showing better responses when judging affective emotions when compared to the semantically coded emotions. Korean listeners, however, showed the reverse pattern; they performed better in the linguistic modality than the affective modality. It is not clear why these two non-native groups performed differently, since English is taught to both groups. However, comprehension of the meaning of the English sentences definitely influenced the Koreans performance on the linguistic utterances. Neutral utterances were well perceived by all groups. It is necessary to point out at this time that the neutral utterances were coded for neutrality in both the linguistic and affective domains. With regard to which emotion elicited most of the correct responses, we see that the Americans were best at perceiving *happy* utterances, followed by *neutral* utterances. They were least accurate in judging *sad* utterances. Acoustic analysis of the utterances used in the perception study show that the *happy* utterances were different from *sad* and *neutral* in having higher intensity, higher pitch range and higher mean fundamental frequency. *Neutral* utterances however, were more similar to *sad* utterances in having lower intensity and lower mean fundamental frequency, which could have led subjects to erroneously choosing neutral responses for the linguistic utterances, regardless of the emotion.

Finally, this study indicates some interesting trends in perception of emotions across all groups. We find that females were better than males across emotions and modality types. Females were also more responsive to the affective prosody;, men, on the other hand, were better in the linguistic utterances where the speaker indicated by word his/her emotion. Another interesting finding of this paper was significant correlation between gender and emotion. We find that men are better at perceiving *happy* emotion but women are better at perceiving *sad* emotion. It would be interesting to see if this pattern is true across other cultures.

5. Acknowledgements

We thank the undergraduate students at the University of Toledo, Showa Music University, and University of Korea for their assistance in data analysis and for participating in this experiment. This work was supported in part by the Japanese Ministry of Education, Science, Sport, & Culture,
Grant-in-Aid for Scientific Research (C)#22520412

Table 4 *Mean (SD) values for acoustic measurements separated by gender, modality and emotion*.

Modality	Emotion	N	Intensity (db)	Syllable Duration(ms)	Pitch Range (hz)	Tempo (syllables/sec)	Mean F0 (hz)
Linguistic	Happy	30	65.1(10.7)	224,56(83.45)	121.53(73.16)	4.9(1.44)	164.9(35.91)
	Sad	30	63.23(2.6)	218.34(43.3)	136.77(107.19)	4.8(.92)	152.57(34.92)
Affective	Happy	33	67.48(2.76)	211(55.27)	196.30(86.30)	5.0(1.13)	203.06(43.53)
	Sad	32	63.25(3.38)	222.1(45)	142.1(108.29)	4.7 (.90)	159.88 (37.80)
Neutral	Neutral	32	62.38(1.43)	212.34(37.62)	170.34(106.42)	4.8(.74)	158.66(40.69)

6. References

[1] Erickson, D. (2005) Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology* 26.4, pp. 317-325.

[2] Menezes, C., Erickson, D., & Franks, C. "Comparison of linguistic and affective perception of happy and sad: A cross-linguistic study." Speech Prosody2010, Chicago.

[3] Burkhardt, F., & Sendlmeier, W.F. Verification of acoustical correlates of emotional speech using formant synthesis. Proceedings of the ISCA Workshop on Speech and Emotion Northern Ireland, 151-156, 2009.

[4] Braun, A., & Oba, R. Speaking tempo in emotional speech – a cross-cultural study using dubbed speech. Interantional workshop on Paralinguistic Speech between models and data. Saarbrucken, Germany, 77-82, 2007.

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

[6] Nakamichi, A., Jogan, A., Usami, M. and Erickson, D. (2002). Perception by native and non-native listeners of vocal emotion in a bilingual movie. *Gifu City Women's College Research Bulletin,* **52**, 87-91.

[7] 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.

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

[9] Erickson, D., Menezes, C., Rilliard, A., Shochi, T. (2011). Effect of language knowledge on perception of emotional utterances, Acoustical Society of Japan. spring meeting, pp. 257-260.