JAPANESE EMOTIONAL SPEECH PRODUCED BY CHINESE LEARNERS AND JAPANESE NATIVE SPEAKERS: DIFFERENCES IN PERCEPTION AND VOICE QUALITY

Xinyue Li[1], Aaron Lee Albin[1], Carlos Toshinori Ishi[2], Ryoko Hayashi[1]

[1]Kobe University, [2]ATR Hiroshi Ishiguro Labs
lixinyue12200@163.com, albin@people.kobe-u.ac.jp, carlos@atr.jp, rhayashi@kobe-u.ac.jp

ABSTRACT

The present study leverages L2 learner data to contribute to the debate whether the perception and production of emotions is universal vs. language-specific. Japanese native speakers and Chinese learners of L2 Japanese were recorded producing single-word Japanese utterances with seven emotions. A different set of listeners representing the same two groups were then asked to identify the emotion produced in each token. Results suggest that identification accuracy was highest within groups (i.e., for learner+learner and for native+native). Furthermore, more confusions were observed in Japanese native speech, e.g., with 'angry' vs. 'disgusted' confused for Japanese native, but not Chinese learner, productions. Analyses of the electroglottography signal suggest these perception results stem from crosslinguistic differences in the productions themselves (e.g., Chinese learners using a tenser glottal configuration to distinguish 'angry' from 'disgusted'). Taken together, these results support the hypothesis that the encoding and recognition of emotions does indeed depend on L1 background.

Keywords: second language acquisition, Mandarin, overlap score, EGG, open quotient

1. INTRODUCTION

It has been proposed that the set of basic emotions shared by all humans is universal [4]. In the same way, some authors have argued that the ways emotions are phonetically realized in production and recognized in perception are also universal, i.e., do not vary based on native language (L1) background [1, 18, 20]. As one piece of evidence in favor of this view, similar confusion patterns have been observed to hold across various L1 groups. For example, in [6], a Japanese native speaker produced the emotionally-neutral word "banana" with five emotions (happy, angry, sad, surprised, and suspicious), and listeners from three language groups (American English, Korean, and Japanese) were asked to identify which emotion they heard in each token. Results suggested that confusion patterns were shared in common across all three groups: happy+anger+surprised were confused, as were sad+suspicious. In a similar design, [16] asked Japanese speakers to identify emotions in speech samples produced by native speakers of Swedish and Russian. The speaker L1 group factor (Swedish vs. Russian) did not have an impact on identification accuracy, nor was it linked to significant acoustic differences in production, suggesting universality at the level of both perception and production.

However, there is increasing evidence L1 background does in fact play a role. For example, in [14], native speakers of Japanese and American English were recorded producing the name "Pikachu" with four emotions (happy, sad, angry and calm), and native speakers from the same two language groups completed an emotion identification task. For both groups, identification accuracy was higher for tokens produced in the listener's L1 (e.g., for Japanese listeners, accuracy was highest with tokens from the Japanese speakers). Likewise, in [15], four native speakers of German produce 30 "pseudo-utterances", which were then presented to listeners from nine different L1 backgrounds (Indonesian plus eight languages spoken in Europe). Results suggested that the closer the typological distance between German and the listener's L1, the higher the identification accuracy. Under a strong universalist view, where L1 background is irrelevant, the results from both of these studies are hard to explain.

The cross-linguistic transfer observed in second language (L2) learners is another form of evidence that emotional speech is language-dependent. In particular, L2 learners have been shown to pattern differently from native speakers in how they perceive emotional speech. For example, in [12], only Russian learners of L2 Japanese (and not Japanese native speakers) had difficulty recognizing "surprise" and "reluctance" in lexically unaccented words.

While such results have been reported at the level of perception, at present, relatively little is known about L2 learners' phonetic encoding of emotional speech in production. The present study seeks to fill this gap, examining emotional speech by Chinese learners of L2 Japanese (as well as a control group of Japanese native speakers) in both perception and production. In particular, this study has two goals. The first goal is to test the hypothesis from [14] that, in a perception experiment, emotional speech would be recognized more accurately within (rather than across) L1 groups. The second goal is to document whether any phonetic cues (and if so, which ones) differentiate the native speakers from L2 learners in production, focusing on differences in voice quality as measured by electroglottography.

2. PERCEPTION EXPERIMENT

2.1. Speech stimuli

The present study examines the following seven emotions: happy, angry, sad, surprised, afraid, disgusted, and neutral. "Neutral" serves as a baseline to the other six emotions, which correspond to the six basic emotions proposed by [4]. A survey of the existing literature confirmed that these seven are indeed the most commonly used in previous studies examining emotional speech from a cross-linguistic and/or L2 perspective.

The present study used the following 11 words as target words: *e* ('eh'), *mé* ('eye'), *úmi* ('ocean'), *momo* ('peach'), *bánana* ('banana'), *niói* ('smell'), *Manami* (female given name), *ímanimo* ('any moment now'), *namámono* ('raw things'), *menomée* ('before one's eyes'), *aóyama* (Tokyo place name). These words were chosen for having meanings that lack connection to any particular emotion, and as such as easy to produce with a variety of emotions. Moreover, these words represent a variety of word lengths (counted in terms of number of moras) and pitch accent patterns. (In the transcriptions above, an acute accent mark indicates lexical pitch accent.)

Eight Japanese native speakers and eight Chinese learners of L2 Japanese were recruited as speakers to produce emotional speech in Japanese. Each group had 4 males and 4 females, with an overall average age of 25.9 (*SD* = 3.71) across all 16 speakers. All Chinese learners spoke Mandarin as their native language and had passed the highest level ("N1") of the standardized Japanese Language Proficiency Test (JLPT). In total, 1,232 tokens of emotional speech were recorded: 7 emotions × 11 target words × 16 speakers.

2.2. Procedure

A different set of 24 listeners (i.e., not overlapping with the speakers who produced the speech stimuli) were recruited for the perception experiment: 12 native speakers of Japanese and 12 Chinese learners of L2 Japanese. Each group had 6 males and 6 females, with an overall average age of 28.96 (*SD* = 6.92) across all 24 listeners. Like the speakers who produced the speech stimuli, all Chinese learners spoke Mandarin as their native language and had passed the highest level ("N1") of the Japanese Language Proficiency Test.

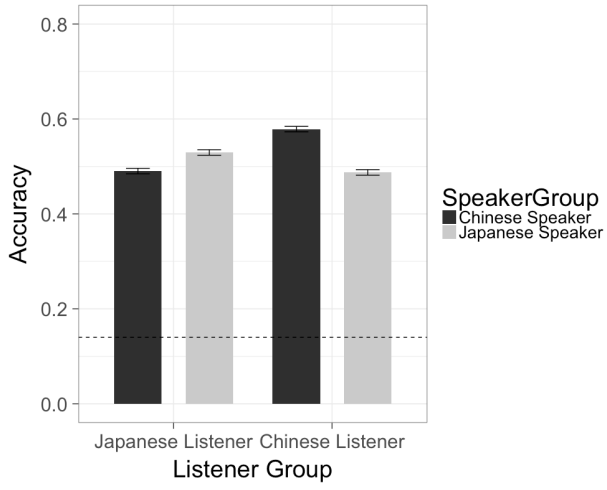
In a quiet language lab, listeners performed a forced-choice perception experiment using the ExperimentMFC functionality in the phonetics software "Praat". Each stimulus was presented once over headphones, and then listeners were presented seven emotions on the computer screen and asked to identify which emotion they had just heard. The 1,232 stimuli were presented in two testing sessions (always in the same order) - first, the 616 stimuli from the Japanese native speakers, then the 616 stimuli from the Chinese L2 learners. Within each session, the stimuli were presented in a randomized order.

2.3. Results and discussion

2.3.1. Identification accuracy

A response was counted as "correct" when the emotion selected by the listener matched the emotion originally intended by the speaker. Identification accuracy, coded using this criterion, is shown in Figure 1 below. (The horizontal line indicates chance level, i.e. 100/7 = 14.3%.)

Figure 1: Identification accuracy for each combination of speaker group and listener group



A three-way ANOVA (listener group × speaker group × emotion) revealed a significant three-way interaction: $F(6, 132) = 2.77$, $p < .05$. To further investigate this interaction, a separate two-way ANOVA (listener group × speaker group) was performed. For the Japanese native stimuli, the identification accuracy of the Japanese listeners was significantly higher than the Chinese listeners (52% vs. 48%, $p < .01$). Conversely, for Chinese learner stimuli, the identification accuracy of the Chinese listeners was significantly higher than Japanese listeners (57% vs. 51%, $p < .05$). Post-hoc analyses confirmed that this result depended to some extent on specific emotions (hence the three-way interaction). For example, Chinese listeners identified emotions more accurately in Chinese learner stimuli only for 'afraid', 'angry', 'neutral', 'sad', and 'surprised'.

These results suggest that identification accuracy was highest within groups. This is consistent with the results of [14], where listeners performed significantly better if they shared the same L1 as the speakers who produced the stimuli. This is what [13] calls the 'in-group advantage', and also reminiscent of the 'interlanguage speech intelligibility benefit' discussed in [2].

2.3.2. Overlap scores

The raw data from the perception task is a confusion matrix: 7 emotions intended by the speaker × 7 emotion response options listeners could choose. Following [10], these confusion matrices were converted into 'overlap scores' using formula (1). Information about directionality (A identified as B vs. B identified as A) is discarded, and the resulting values range from 0 (no overlap/confusion) to 1 (complete overlap/confusion).

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (1)$$

The overlap score for all logically possible pairs of emotions is shown in Table 1. Rows correspond to emotion pairs, and columns represent different pairings of speaker group and listener group. Bolding indicates cases in which the overlap score exceeds 0.50. The final column indicates the average of the four overlap scores in each row. Rows are placed in descending order according to these averages.

These results suggest that pairs of emotions can be arranged on a scale from extremely confusable (e.g., Sad+Afraid, mean=0.60) to clearly perceptually separate (e.g., Surprised+Sad, mean=0.18). Comparing the four different combinations of groups, an overall greater level of confusion was observed for the Japanese native stimuli. For instance, Disgusted+Angry and Sad+Neutral (third and fourth rows) had overlap scores exceeding 0.50 for stimuli from the Japanese speakers but not the Chinese speakers. This suggests that, at least for certain pairs of emotions, the speech produced by the Japanese speakers is perceived to be more ambiguous than that of the Chinese speakers.

Table 1: Overlap Scores for each emotion pair (JS=Japanese speaker, CS=Chinese speaker, JL=Japanese listener, CL=Chinese listener)

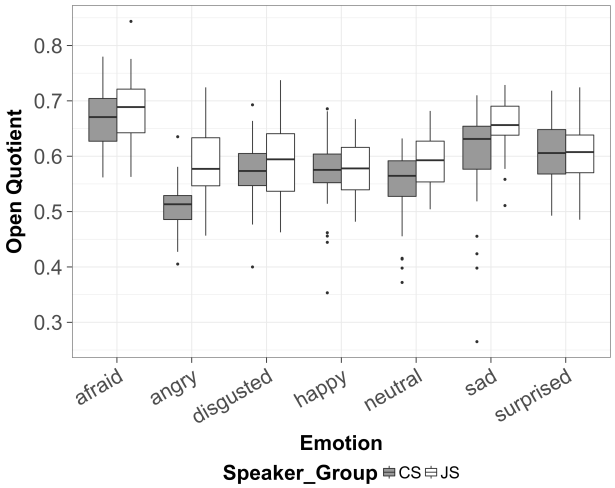
	JS-JL	**JS-CL**	**CS-JL**	**CS-CL**	*Avg.*
Sad+Afraid	**0.62**	**0.68**	**0.55**	**0.56**	*0.60*
Surprised+Happy	**0.56**	**0.65**	**0.52**	**0.63**	*0.59*
Disgusted+Angry	**0.58**	**0.53**	0.40	0.46	*0.49*
Sad+Neutral	0.45	**0.55**	0.35	0.44	*0.45*
Happy+Disgusted	0.44	0.47	0.41	0.48	*0.45*
Happy+Angry	0.41	0.42	0.44	0.47	*0.44*
Disgusted+Afraid	0.40	0.49	0.35	0.40	*0.41*
Surprised+Disgusted	0.36	0.40	0.39	0.45	*0.40*
Neutral+Afraid	0.42	0.50	0.29	0.37	*0.40*
Sad+Disgusted	0.45	0.45	0.27	0.30	*0.37*
Neutral+Disgusted	0.41	0.44	0.27	0.34	*0.36*
Neutral+Angry	0.33	0.35	0.34	0.43	*0.36*
Happy+Afraid	0.35	0.35	0.27	0.26	*0.31*
Surprised+Angry	0.31	0.35	0.26	0.31	*0.31*
Sad+Angry	0.34	0.35	0.25	0.28	*0.31*
Angry+Afraid	0.33	0.36	0.26	0.27	*0.30*
Neutral+Happy	0.27	0.28	0.32	0.33	*0.30*
Surprised+Afraid	0.30	0.30	0.23	0.25	*0.27*
Sad+Happy	0.27	0.27	0.27	0.27	*0.27*
Surprised+Neutral	0.18	0.20	0.15	0.20	*0.18*
Surprised+Sad	0.19	0.21	0.16	0.16	*0.18*
Overall mean:	***0.38***	***0.41***	***0.32***	***0.36***	

3. VOICE QUALITY OF SPEECH STIMULI

In previous research, voice quality has been closely linked to the encoding and decoding of emotional speech [21]. For instance, breathiness and tenseness have been found to be relevant for the identification of emotions [11, 17]. The glottal open quotient is one measurement of voice quality that is useful for discriminating tense vs. lax voice [7, 8]. Originally documented in [19], open quotient can be defined as the duration of the glottal open phase normalised to the local glottal period [9]. Open quotient can be directly derived from an electroglottography (EGG) signal. Lower open quotient values (closer to 0) indicate tenseness (e.g., creaky voice), and higher values (closer to 1) indicate laxness (e.g., breathy voice).

When recording the speech stimuli as described in section 2.1 above, a Glottal Enterprises EG2-PCX2 unit was used to record the EGG signal from all male speakers and a subset of the female speakers. The average open quotient value across each entire word token was then calculated, and these values were grouped by emotion and speaker group. Figure 2 represents the results of this analysis, using data from the male speakers only (4 Japanese native speakers and 4 Chinese L2 learners of Japanese).

Figure 2: Distribution of Open Quotient values for each emotion and speaker group



A two-way ANOVA (speaker group × emotion) revealed a significant main effect of emotion ($F(6, 36) = 23.76$, $p < .001$) as well as a significant two-way interaction ($F(6, 36) = 2.98$, $p < .05$). Post-hoc tests on the Chinese speakers ($F(6,18) = 30.45$, $p < .001$), corrected for multiple comparisons, indicated six specific pairwise contrasts were different for the Chinese speakers: angry < afraid, disgusted < afraid, disgusted > angry, sad < neutral, surprised < afraid, and surprised < disgusted (all *p*s < .05). These are summarized visually in Table 2 below.

Table 2: Significant differences in Open Quotient values between pairs of emotions for Chinese Speakers

CS	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Afraid							
(2) Anger	<						
(3) Disgusted	<	>					
(4) Happy							
(5) Neutral							
(6) Sad					<		
(7) Surprised	<		<				

Crucially, no similar significant differences for open quotient were observed for the Japanese speakers. This implies that the Chinese speakers use glottal tenseness/laxness as a phonetic cue when encoding emotions in production to a greater extent than Japanese speakers.

To take up one example, recall from Table 1 above that Disgusted+Angry is one example of an emotion pair that is highly confused (overlap score > 0.50) in speech from the Japanese speakers but not the Chinese speakers. In the results from Table 2, it can be seen that Angry was significantly tenser than Disgusted in the speech of Chinese speakers. Thus, Chinese speakers may be using vocal tension (as signalled by open quotient) as one cue to distinguish these two emotions, whereas the same is not true for Japanese speakers.

4. CONCLUSION

The results of the present study can be summarized as follows. First, identification accuracy was highest within L1 groups, i.e., when a listener heard Japanese emotional speech produced by a speaker from his/her same language background (Japanese native speaker vs. Chinese learner of L2 Japanese). This entails that the identification of emotions in speech perception is dependent on the listener's L1 background.

Second, analysis of overlap scores indicated that, overall, more confusions were observed in Japanese native speech (e.g., Angry+Disgusted frequently confused when produced by the native speakers but not the Chinese learners). An analysis of word-level average open quotient values suggested that this crosslinguistic difference in confusability may stem from differences in the productions themselves. More specifically, Chinese learners of L2 Japanese use a tenser glottal configuration to distinguish six different pairs of emotions (including Angry+Disgusted), whereas Japanese native speakers did so for zero pairs. This finding implies that the phonetic encoding of emotions in speech production also varies depending on the speaker's L1 background. More broadly, this entails that different sets of emotional contrasts are distinguished phonetically in different languages.

Taken as a whole, the results of the present study support the hypothesis that the phonetic encoding of emotions in speech production, as well as the recognition of emotions in speech perception, both depend on L1 background. This conclusion backs up similar findings reported in several recent previous studies [12,14,15] and adds further evidence against the claim that the principles of emotional speech are universal and language-independent.

5. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 17H02352 and partly supported by JST ERATO Grant Number JPMJER1401, Japan.

6. REFERENCES

[1] Albas, D., McCluskey, K., Albas, C. 1976. Perception of the emotional content of speech: A comparison of two Canadian groups. *Journal of Cross-Cultural Psychology*, *7*(4), 481–489.

[2] Bent, T., Bradlow, A. 2003. The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610.

[3] Campbell, N., Erickson, D. 2004. What do people hear? A study of the perception of non-verbal affective information in conversational speech. *Journal of the Phonetic Society of Japan*, *8*(1), 9-28.

[4] Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200.

[5] Erickson, D. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, *26*, 317–325.

[6] Erickson, D. 2006. Some gender and cultural differences in perception of affective expressions. *In Proceedings of Speech Prosody 2006*, Paper 029.

[7] Hanson, H., Stevens, K., Kuo, H., Chen, M., Slifka, J., 2001. *Towards models of phonation*. Journal of Phonetics, *29*, 451- 480.

[8] Henrich, N., d'Alessandro, C., Doval, B. 2001. Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data. *In Proceedings of EUROSPEECH 2001*, 47-50.

[9] Kane, J., Scherer, S., Morency, L., Gobl, C. 2013. A comparative study of glottal open quotient estimation techniques. *In Proceedings of INTERSPEECH 2013*, 1658-1662.

[10] Levy, E. S. 2009. On the assimilation-discrimination relationship in American English adults' French vowel learning. *The Journal of the Acoustical Society of America*, *126*(5), 2670–2682.

[11] Lugger, M., Yang, B. 2007. The relevance of voice quality features in speaker independent emotion recognition. *In Proceedings of ICASSP 2007*, *4*, 17-20.

[12] Nakabayashi, R. 2011. Recognition Emotion from Japanese Utterances: Individual differences among russian learners of Japanese language. *Journal of the Phonetic Society of Japan*. *15*(3), 14–25.

[13] Pell, M. D., Monetta, L., Paulmann, S., Kotz, S. A. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*, 107–120.

[14] Sakuraba, K., Imaizumi, S., Kakehi, K. 2004. Emotional expression in "Pikachuu". *Journal of the Phonetic Society of Japan*. *8*, 77-84.

[15] Scherer, K. R., Banse, R., Wallbott, H. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*, 76–92.

[16] Shigeno, S. 2014. Effects of Lexical Contents on Expression of Emotional Speech. *The Journal of the Japan society of Logopedics and Phoniatrics*, *55*(3), 233–238.

[17] Tahon, M., Degottex, G., Devillers, L. 2012. Usual voice quality features and glottal features for emotional valence detection. *In Proceedings of Speech Prosody 2012*, 693–696.

[18] Thompson, W., Balkwill, L-L. 2006. Decoding speech prosody in five languages. *Semiotica*, *158*, 407–424.

[19] Timcke, R., von Leden, H., Moore, P. 1958. Laryngeal vibrations: Measurements of the glottis wave. *AMA Archives of Otolaryngology*. *68*(1), 1–19.

[20] Van Bezooijen, R., Otto, S., Heenan, T. 1983. Recognition of vocal expression of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, *14*, 387–406.

[21] Wilson, D., Wharton, T. 2006. Relevance and prosody. *Journal of Pragmatics*, *38*, 1559–1579.