Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

An Experimental Study of Emotional Speech in Mandarin and English

Ting Wang [1], *Yong-cheol Lee* [2], *Qiuwu Ma* [1]

[1] School of Foreign Languages, Tongji University, Shanghai, China
[2] Department of English Language and Literature, Cheongju University, Cheongju, Korea
2011ting_wang@tongji.edu.cn, soongdora@gmail.com, mqw@tongji.edu.cn

Abstract

This study reports our initial results on whether the use of pitch for expressing emotions differs between Mandarin and English. The production experiment was conducted using five emotions (anger, fear, happiness, sadness, and neutral) by comparing both prosodic cues and phonation cues between Mandarin and English emotional speech. Results demonstrated that within each language, each vocal emotion had specific acoustic patterns. Moreover, Mandarin and English showed different mechanisms of utilizing pitch for encoding emotions. The differences in pitch variation between neutral and other emotions were significantly larger in English than in Mandarin. However, the variations of speech rate and certain phonation cues (e.g., CPP and CQ) were significantly larger in Mandarin than in English. The differences in emotional speech between the two languages may be due to the restriction of pitch variation by the presence of lexical tones in Mandarin. This study reveals an interesting finding that when a certain parameter (e.g., pitch) is restricted in one language, other cues turned out to be strengthened for compensation. Therefore, we posit that the acoustic realizations of emotional speech are multidimensional.

Index Terms: emotional speech, pitch variation, Mandarin, English

1. Introduction

It has long been debated whether the mechanisms underlying the vocal expression of emotion are language-universal or language-specific. The former proposal is primarily suggested in cross-cultural perception studies, in which subjects from one language background were asked to identify the underlying emotions expressed by speakers from either the same or a different language background (e.g., [1-4]). These studies reliably reported that listeners could successfully recognize vocal emotions expressed in a foreign language and show no great variation in rates of identification. The only exception is that listeners may have an in-group advantage when identifying vocal emotions expressed in their native language. In addition, research focusing on the acoustic profiles associated with different vocal emotions also reveals that the acoustic characteristics of emotions remain similar across different languages [5]. For example, [6] investigated the acoustic patterns indicated by mean f0, f0 range, and speech rate for six vocal emotions in English, German, Hindi, and Arabic. Results showed that the vocal expression of emotions in these four distinct languages exhibited similar acoustic patterns, which seemed to be unaffected by language.

However, this may not be the case when it comes to tonal languages. A separate body of literature has investigated the influence of the lexical tone system in tonal languages on the acoustic realization of vocal emotions (e.g., [7-8]). The motivation to examine this possible influence arose from the hypothesis that the existence of a lexical tone system may restrict the degree to which pitch can be used for other intonational variations [9]. This idea is supported by a recent study, [10], which focused on the role of pitch variation for expressing emotions within a tonal language such as Mandarin. This study reported that the pitch variation within all high level tone sequences for vocal emotion expressions showed clear pitch restriction. This discovery was interpreted as a possible effect of the lexical tone system of tonal languages, which may have caused a restriction of the paralinguistic use of pitch for the expression of vocal emotions.

The current study follows this finding and proposes some further questions to be answered. The first question is whether the pitch restriction phenomenon also occurs in Mandarin mixed tone sentences when compared to a non-tonal language such as English. To explore this question, we gathered data using the same material of the mixed tone sentences from [10], paired with a parallel version in English. Secondly, another point to consider is that, although prosodic cues have been reported to be the most salient indicators of vocal expression of emotion, phonation cues also play important roles in differentiating emotions [11]. When pitch variation is restricted, the changes of other cue dimensions are worth examining. We hypothesized that the restricted use of pitch in a tonal language may result in a trade-off with different cues being used to signal and express vocal emotions. Thus, in the present study, we explored this issue by also considering phonation-related acoustic cues and EGG cues.

2. Method

2.1. Speech materials

Speech materials were made up of 15 declarative sentences. They were first designed in Chinese and then were translated to English. Each target sentence is semantically neutral and is suitable to convey different kinds of emotions. They were embedded in a certain context to reflect five different emotions: anger, fear, happiness, sadness, and neutral. Table 1 lists the sample target sentences in two languages.

Table 1. *Sample target sentences in Mandarin and English.*

Language	Mandarin	English
Target	我不敢相信这是真的。	I can't believe this is true.

 doi: 10.21437/SpeechProsody.2016-88

sentences	这是李梅的男朋友。	This is Mary's boyfriend.
	老板今天给我打电话了。	My boss called me today.

2.2. Subjects

We recruited five native speakers (2 males and 3 females) for each language. All the speakers were graduate students at the University of Pennsylvania. Mandarin speakers had spent less than a year in the US at the time of recordings. They all had experience of acting and public speaking. Participants signed a consent form before the experiment and were offered ten dollars as compensation for their time. Participants reported no problems with their speech and hearing.

2.3. Recording procedure

Recordings were conducted in a sound-proof booth in the Department of Linguistics at the University of Pennsylvania. We obtained both simultaneous electroglottograph (EGG) and audio recordings from all speakers. Audio recordings were made electronically and saved directly on a computer as 16-bit wave files at a sampling rate of 44.1 kHz, using a Glottal Enterprises M80 omnidirectional headset microphone. EGG data were obtained using a two-channel Glottal Enterprises Electroglottograph, model EG2. During the recordings, we presented speech materials through PowerPoint Slides.

Different emotions were recorded in a separate block, and speakers were offered a break between blocks for a smooth transition from one emotion to another. Sentences with neutrality were produced in isolation, and those with anger, fear, happiness, and sadness were embedded in a dialogue setting, in which speakers conducted a dialogue with a native speaker. This dialogue setting enabled speakers to express different emotions in a natural way. Finally, we collected a total of 750 sentences (= 15 sentences × 5 speakers × 2 repetitions × 5 emotions) for each language.

2.4. Listening tests

Listening tests were conducted to confirm that the intended emotions were accurately produced using a online survey tool (Qualtrics) with Mandarin and English stimuli respectively. For each language, the stimuli were divided into five sets produced by each of the five speakers. Native Mandarin and English listeners were recruited online and the number of listeners for each set was at least 15. During the listening tests, participants were asked to listen to randomized audio stimuli carefully and select the most appropriate emotion on a five-choice task (i.e., anger, happiness, fear, neutral, and sadness) by pressing a computer mouse.

Recordings were excluded whose identification rate was less than 60%. This is three times the 20% chance level permitted in our study (Pell et al., 2009). Finally, in order to compare five emotions with parallel texts in each language, we chose 530 perceptually valid recordings (106 sentences × 5 emotions) in Mandarin, and 300 perceptually valid recordings (60 sentences × 5 emotions) in English.

2.5. Measurements

Automatic segmentation of recordings was made using SPPAS [12]. Three tiers (phoneme tier, syllable tier and sentence tier) were generated and manually corrected afterward. A sequence of pitch target points was detected by Momel algorithm [13] using Praat. Pitch contours were modeled as continuous smooth curves, interpolated quadratically from pitch target points of each utterance in order to eliminate microprosodic effects. Based on the Momel outputs, a series of prosodic and phonation parameters were measured to quantify voice quality. In what follows, we first describe the details of the prosodic measurements, followed by the remaining measurements.

Seven prosodic parameters were generated by a Praat script: (1) Number of syllables per second (Speech Rate); (2) Mean intensity of each sentence (Mean Intensity); (3) Mean pitch of each sentence (Mean Pitch); (4) Pitch range of each sentence (Pitch range) (5) Average absolute difference between two adjacent pitch target points divided by distance in seconds (Mean Absolute Slope, MAS hereafter), reflecting the frequency of pitch movements [14]; (6) Average pitch difference between two adjacent pitch target points for rise and fall separately in each sentence (Rise, Fall), which determine the degree of pitch raising and pitch falling [14]; (7) Average slope for rise and fall separately in each sentence (Rise Slope, Fall Slope), indicating the speed rate of pitch raising and pitch falling [14]. These parameters were captured by both global and local pitch movements of each sentence. Pitch-related parameters were normalized to eliminate individual differences using the OMe (Octave-Median) scale [15] by applying the following equation:

$$OMe = log_2\left(\frac{Hz}{Median}\right)$$

where Hz is a raw pitch value, and Median indicates the median of a speaker's pitch range.

In addition, we obtained a set of phonation measurements using VoiceSauce [16] through the voicing parts of each sentence. They are: (1) Cepstral Peak Prominence (CPP), indicating harmonics-to-noise ratio and periodicity [17]; (2) Amplitudes of the first, second and fourth harmonics (H1, H2 and H4); (3) Amplitude difference between the first and second harmonics (H1-H2), reflecting the relative breathiness or creakiness of phonation [18]; (4) Amplitude difference between the first harmonic and harmonics nearest to F1, F2, and F3 (H1-A1, H1-A2 and H1-A3), as measures of spectral tilt [19]. This set of phonation parameters was often used as standard in measuring different phonation properties [20]. Next, three kinds of EGG measures were extracted using EggWorks [21], including (1) Contact quotient (CQ), illustrating the duration of the vocal fold contact during one vocal fold period [22]; (2) Peak Increase in Contact (PIC), the peak positive value in the EGG derivative, indicating the highest speed of vocal fold contact [23]; (3) Speed Quotient (SQ), the ratio between closing duration and opening duration, reflecting the asymmetry of the EGG pulses [24]. These parameters are indicators of the physiological mechanism of vocal folds vibration during speech production.

In summary, in order to assess the vocal expression of emotion in Mandarin and English, we measured a total of 20 parameters which were then converted to z-scores combining all the emotions separately for each speaker.

3. Results

3.1. Prosodic cues for encoding emotions in Mandarin and English

To overcome evident biases arising from the absolute value in two languages, data normalization was first made according to the following equation [25-26]:

$$Normalized\ X = (x - N)/N$$

where x is the absolute value of each parameter for anger, fear, happiness and sadness, respectively. N is the absolute value of each parameter for neutral. Thus this equation produces the relative value of each parameter in each four emotions compared to the neutral baseline. The derived data have a positive or a negative value, depending on the relative difference with neutral.

These derived data was analyzed in a set of mixed repeated measures ANOVAs, with Emotion (anger, fear, happiness, sadness) as a within-subject factor, and Language (Mandarin, English) as a between-subject factor.

As presented in Figure 1, the variations of pitch-related measures around 0 are larger in English than in Mandarin. The main effect of Emotion is significant on all seven pitch-related measures (mean pitch: $F[3, 24] = 19.905$, $p < 0.001$; pitch range: $F[3, 24] = 31.567$, $p < 0.001$; mean absolute slope: $F[3, 24] = 50.404$, $p < 0.001$; rise: $F[3, 24] = 33.377$, $p < 0.001$; fall: $F[3, 24] = 23.507$, $p < 0.001$; rise slope: $F[3, 24] = 16.980$, $p < 0.001$; fall slope: $F[3, 24] = 27.416$, $p < 0.001$). In terms of Language, Mandarin and English have significant differences on pitch range ($F[1, 8] = 7.768$, $p < 0.05$), mean absolute slope ($F[1, 8] = 34.760$, $p < 0.001$), rise ($F[1, 8] = 12.769$, $p < 0.01$), fall ($F[1, 8] = 30.419$, $p < 0.01$), rise slope ($F[1, 8] = 22.950$, $p < 0.01$), and fall slope ($F[1, 8] = 22.086$, $p < 0.01$). The interaction effect of Emotion × Language is significant on mean absolute slope ($F[3, 24] = 7.734$, $p < 0.01$), rise ($F[3, 24] = 5.904$, $p < 0.01$), fall ($F[3, 24] = 5.158$, $p < 0.01$), rise slope ($F[3, 24] = 3.210$, $p < 0.05$), and fall slope ($F[3, 24] = 5.095$, $p < 0.01$), suggesting that English vocal emotions show more dynamic pitch-related movements than Mandarin vocal emotions.

In contrast, the variations of speech rate around 0 are larger in Mandarin than in English. Statistical results show that there is a significant main effect of Emotion on both speech rate ($F[3, 24] = 29.519$, $p < 0.001$) and mean intensity ($F[3, 24] = 47.668$, $p < 0.001$). Language only has a significant main effect on speech rate ($F[1, 8] = 12.342$, $p < 0.01$). There is a significant interaction effect between Emotion and Language on speech rate ($F[3, 24] = 5.291$, $p < 0.01$).

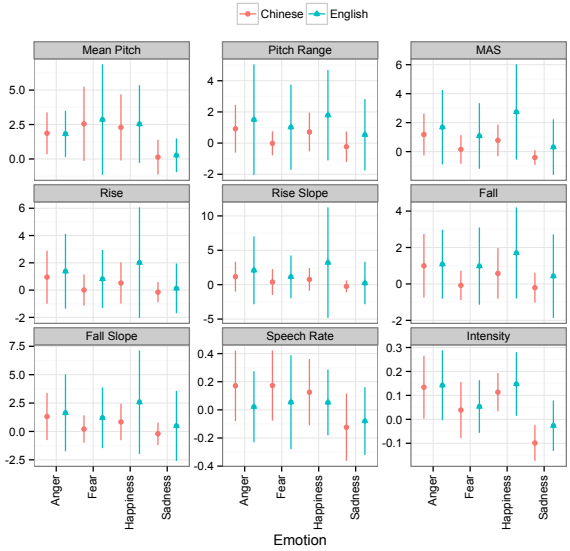

Figure 1: *Prosodic cues of four emotions in Mandarin and English. Points indicate mean values and error bars 95% confidence intervals.*

3.2. Phonation cues for encoding emotions in Mandarin and English

Figure 2 illustrates the effects of Emotion and Language on comprehensive phonation cues. It is clear that the mean differences of CPP among four emotions are larger in Mandarin than in English. Mixed repeated measures ANOVAs results show that Emotion has a significant main effect on CPP ($F[3, 24] = 18.852$, $p < 0.001$), H1 ($F[3, 24] = 3.595$, $p < 0.05$), and H2 ($F[3, 24] = 7.918$, $p < 0.01$). Language has a significant main effect only on H1-A1 ($F[1, 8] = 6.522$, $p < 0.01$). The interaction between Emotion and Language is significant on CPP ($F[3, 24] = 5.101$, $p < 0.01$), meaning that the two languages have different CPP patterns.

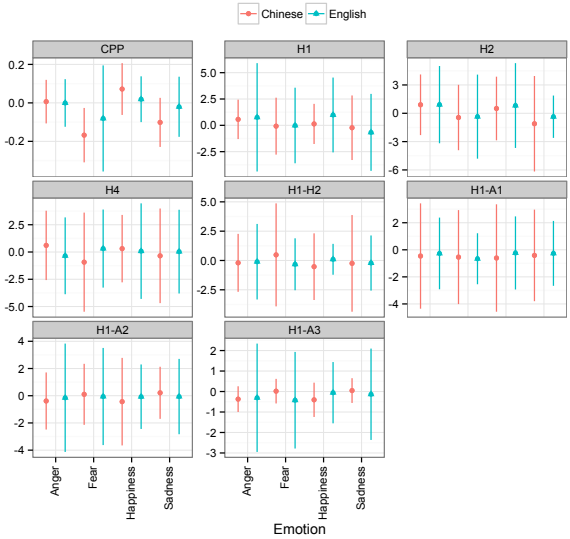

Figure 2: *phonation-related acoustic measures of four emotions in Mandarin and English. Points indicate mean values and error bars 95% confidence intervals.*

In order to understand the physiological mechanisms of vocal emotions production in Mandarin and English, three EGG measures were analyzed. Figure 3 displays the effects of Emotion and Language on these EGG measures. Again, we can see that the mean differences of these three cues among four emotions are larger in Mandarin than in English. In terms of Emotion, there are significant differences among four emotions on CQ (F[3, 24] = 20.772, $p < 0.001$), SQ (F[3, 24] = 20.408, $p < 0.001$), and marginal significant differences on PIC (F[3, 24] = 2.932, $p = 0.054$). However, the main effects of Languages on these three cues were not significant. The interaction between Emotion and Language is significant on CQ (F[3, 24] = 9.524, $p < 0.001$) and SQ (F[3, 24] = 11.031, $p < 0.001$).

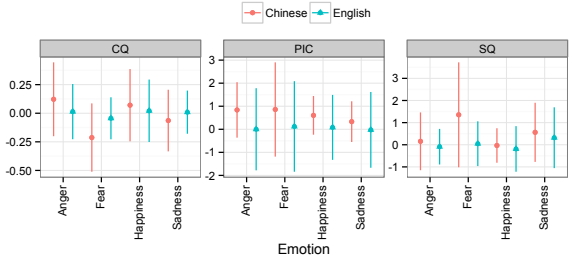

Figure 3: *Physiological measures of phonation of four emotions in Mandarin and English. Points indicate mean values and error bars 95% confidence intervals.*

4. Discussion and conclusions

The present study aimed to examine whether a tonal language shows restricted pitch variation when encoding vocal emotions, as compared to a non-tonal language. If this is indeed the case, then which acoustic cues are used to compensate for the restricted pitch variation? To discover the truth of this matter, the effects of emotion and language on comprehensive prosodic cues, phonation cues, and EGG cues were assessed through production experiments.

Based on the acoustic analysis on pitch-related cues, we observed that Mandarin and English showed different mechanisms when utilizing pitch to express vocal emotions. There were significant interactions between emotion and language on mean absolute slope, rise, fall, rise slope, and fall slope. The pitch variations, with neutral as the baseline, were significantly larger in English compared to those in Mandarin. We posit that this difference is due to the restriction of pitch variation by the existence of lexical tones in Mandarin. Although the present study is not the first to propose this idea [7-8], our study expands upon previous work in this area and provides additional objective measures. For example, this study included more sophisticated pitch measures, including rise, fall, rise slope, fall slope and mean absolute slope, which are able to better reflect the variation of pitch over time. Additionally, instead of comparing absolute values between Mandarin and English, we used derived values to reflect the magnitude of change certain cues showed in emotional speech given neutral as the baseline. That is to say, we measured the relative contribution of certain cues when differentiating among emotions. Therefore, we believe our study provides new evidence regarding the pitch restriction hypothesis from a fresh perspective.

With regard to other prosodic cues and phonation cues, the differences between neutral and other emotions on speech rate, CPP, CQ, and SQ were significantly larger in Mandarin than in English. These cues were enhanced in Mandarin emotional speech in order to compensate for the suppressed pitch variation.

To sum up, this study revealed evidence supporting the idea that the acoustic realizations of emotional speech are multidimensional. When a certain dimension (for example, pitch) is restricted within a language, other dimensions may be strengthened in compensation. Given the limited number of tonal and non-tonal languages examined by the present study, this conclusion needs to be further investigated by analyzing a greater and more diverse pool of languages.

5. Acknowledgements

Our thanks go to all the participants in the production and perception experiments. We also thank Professor Jianjing Kuang at University of Pennsylvania for help with EGG recording facility.

6. References

[1] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology,* vol. 32, pp. 76-92, 2001.

[2] W. F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica,* vol. 2006, pp. 407-424, 2006.

[3] C. Graham, A. W. Hamblin, and S. Feldstein, "Recognition of emotion in English voices by speakers of Japanese, Spanish and English," *IRAL-International Review of Applied Linguistics in Language Teaching,* vol. 39, pp. 19-37, 2001.

[4] A.-J. Li, Y. Jia, Q. Fang, and J.-W. Dang, "Emotional intonation modeling: A cross-language study on Chinese and Japanese," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA),* 2013, pp. 1-6.

[5] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology,* vol. 70, p. 614, 1996.

[6] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics,* vol. 37, pp. 417-435, 2009.

[7] E. D. Ross, J. A. Edmondson, and G. B. Seibert, "The effect of affect on various acoustic measures of prosody in tone and non-tone languages-a comparison based on computer-analysis of voice," *Journal of phonetics,* vol. 14, pp. 283-302, 1986.

[8] C. S. Chong, J. Kim, and C. Davis, "Exploring Acoustic Differences Between Cantonese (Tonal) and English (Non-Tonal) Spoken Expressions of Emotions," in *INTERSPEECH,* pp. 1522-1526.

[9] A. Cruttenden, *Intonation*: Cambridge University Press, 1997.

[10] T. Wang and Y.-c. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in Mandarin Chinese?," *The Journal of the Acoustical Society of America,* vol. 137, pp. EL117-EL123, 2015.

[11] T. Wang, H. Ding, J. Kuang, and Q. Ma, "Mapping Emotions into Acoustic Space: the Role of Voice Quality," in *INTERSPEECH,* 2014, pp. 1978-1982.

[12] B. Bigi and D. J. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody," in *Speech Prosody*, 2012, pp. 1-4.
[13] D. J. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of speech Sciences,* vol. 1, pp. 55-83, 2011.
[14] D. J. Hirst, "Melody metrics for prosodic typology: comparing English, French and Chinese," in *Interspeech*, 2013, pp. 572-576.
[15] C. De Looze and D. J. Hirst, "The OMe (Octave-Median) scale: a natural scale for speech prosody," in *Speech Prosody*, 2014, pp. 910-914.
[16] Y.-l. Shue, P. Keating, C. Vicenik and K. Yu, "VoiceSauce: A program for voice analysis," In *ICPhS*, 2011.
[17] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research,* vol. 37, pp. 769-778, 1994.
[18] K. Johnson, "The auditory/perceptual basis for speech segmentation," *Working papers in linguistics-Ohio State University Department of Linguistics,* pp. 101-113, 1997.
[19] K. N. Stevens, "Diverse acoustic cues at consonantal landmarks," *Phonetica,* vol. 57, pp. 139-151, 2000.
[20] J. Kuang and P. Keating, "Vocal fold vibratory patterns in tense versus lax phonation contrasts," *The Journal of the Acoustical Society of America,* vol. 136, pp. 2784-2797, 2014.
[21] H. Tehrani. 2012. EggWorks: http://www.linguistics.ucla.edu/faciliti/facilities/physiology/EGG.htm
[22] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech, Language, and Hearing Research,* vol. 31, pp. 338-351, 1988.
[23] P. Keating, C. Esposito, M. Garellek, S. Khan, and J. Kuang, "Phonation contrasts across languages," *UCLA Working Papers in Phonetics,* vol. 108, pp. 188-202, 2010.
[24] K. Marasek, "Glottal correlates of the word stress and the tense/lax opposition in German," in *ICSLP,* 1996, pp. 1573-1576.
[25] L. Anolli, L. Wang, F. Mantovani, and A. De Toni, "The voice of emotion in Chinese and Italian young adults," *Journal of Cross-Cultural Psychology,* vol. 39, pp. 565-598, 2008.
[26] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition, Technology & Work,* vol. 6, pp. 4-14, 2004.