Prosodic and Acoustic Features of Emotional Speech in Taiwan Mandarin

Hsin-Yi Lin, Janice Fon

Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan

niceshelly46@yahoo.com.tw, jfon@ntu.edu.tw

Abstract

The present study investigated pitch and durational cues of emotional speech in Taiwan Mandarin. Acted connected speech in anger, joy, sorrow, fear and neutral emotions were recorded. Results showed that F_0 height and speech rate were more correlated with the arousal dimension, which differentiate emotions of high arousal, such as anger and joy, from low arousal ones. On the other hand, final lengthening was shown to distinguish emotions of different valences. Negative emotions, such as anger and sorrow, had longer lengthening than positive ones. In addition, instead of using the size code, speakers of Taiwan Mandarin mainly utilize arousal cues in cueing anger, which was shown through raised pitch register and earlier peak alignment.

Index Terms: prosodic cues, pitch, duration, emotional speech, Taiwan Mandarin.

1. Introduction

Of many ways to express emotional states, prosody is among the most important ones. However, since prosody is also highly determined by linguistic structures [1], the work of disentangling the acoustic correlates of emotion from prosody is rather challenging. In order to systematically observe the association between prosodic cues and emotions, the dimensional theories of emotion were broadly applied in many previous studies [2]. They conceptualize emotion as having two or more dimensions, and the two most commonly defined dimensions are arousal and valence. Arousal refers to the degree of excitement or engagement of the signaler with an emotion, while valence refers to the intrinsic positive or negative properties of an event that leads to an emotion. Many studies have tried to figure out possible cues that vary along the two dimensions, and some correlations have been established. For example, emotions with high arousal, such as anger and joy were found to be closely related with a higher mean F_0, a greater pitch range, and a faster speech rate, while ones with low arousal, such as sorrow, were found to be characterized by the opposite trend (e.g., [3]). As for the dimension of valence, positive emotions show smaller intensity variation, steeper spectral slope, while negative emotions show the opposite [4].

However, while most of the previous studies focused on emotional expressions in English, little has been done for Mandarin. As a tone language, Mandarin might be different from English since the tonal system should impose certain constraints on pitch patterning [5]. As a consequence, emotion in Mandarin might have different pitch patterns from English, or might even be conveyed by other cues. Of the few studies on emotional speech in Mandarin, they were mostly on Beijing Mandarin [6, 7], and the results showed that while F_0, phonation, and amplitude play important roles in differentiating different emotions, duration was less used. This might be due to the fact that since Beijing Mandarin is more of a stress-timed dialect of Mandarin Chinese [8], its syllable duration, under the need of lexical stress contrast, might be less elastic to emotional changes. This suggests that as a more syllable-timed dialect [8], Taiwan Mandarin may have more space to utilize duration cues. However, since the only relevant prosodic study on emotional speech in Taiwan Mandarin focused only on single syllables [9], it is hard to obtain a more global duration modulation pattern. Therefore, this study intends to fill the gap by investigating emotional cues in Taiwan Mandarin using acted connected speech.

2. Method

2.1. Speakers

Five male and five female native speakers of Taiwan Mandarin, aged from 22 to 35 years old, were recruited. They had no language impairment and training in performance according to self report. They were paid for their effort. It should be noted that since the analyses of male speakers' results have not been completed at this stage, only the results of female speakers were reported and discussed in this paper.

2.2. Equipment

Recording was made using a KORG MR-1000 digital recorder with a SENNHEISER HMD 25-1 600Ω head-mounted microphone. SONY MRD-7506 headphones were used by the experimenter for monitoring throughout the recording.

2.3. Speech Materials

Speech materials were short scripts containing daily conversations. There were five targeted emotion types in total, anger, joy, sorrow, fear, and neutral. Except for neutral, two levels of strength were also included for each emotion. Each script contained three target sentences, and four scripts were constructed for each emotion, resulting in [emotion (4) × strength (2) + neutral] × 4 = 36 scripts in total. The scripts were printed on index cards, one script per card. An example is shown in Table 1.

Table 1. *An example of the scripts. The underlined sentences are the target sentences.*

（有點開心）(joy at the weak level)

A: 什麼事這麼開心啊？
(What makes you so happy?)
B: 剛才老闆跟我說，有人告訴他我很努力，覺得我上進。
(The boss just told me that someone told him that I am hard-working, and (he/she) thinks I am very aggressive.)
A: 哇，是哪位同事人這麼好啊？
(Wow, who is that kind colleague?)
B: 一定是陸依娜。她常常說我很棒。應該是她沒錯。
(It must be Lu Yi Na. She often says I rock. It should be her with no doubt.)

2.4. Procedure

Recording was done individually in a sound-treated room. As shown in Table 1, the lines of Role A were read by the experimenter, and those of Role B were read by a participant. Participants were given a practice trial before the real experiment began. They were shown the index cards in a semi-randomized order and were asked to act out the designated emotions as natural as possible. Recording was divided into four blocks, each of which contained nine trials. Participants were allowed short breaks between blocks. The whole recording took about thirty to forty minutes.

3. Measurements

3.1. Final lengthening

Final lengthening is the lengthening of the final syllable at the boundary region. It has been shown to be an important cue in indicating linguistic structures [1]. In view of its importance in prosody, the present study attempted to see whether speakers also utilize it in differentiating emotions. It was quantified as the duration ratio between the final and the initial syllables of each target sentence.

3.2. Speech rate

Speech rate is one of the most studied parameters in emotional speech. It was usually calculated as the number of uttered syllables divided by the duration of a given chunk (e.g. a sentence). In the present study, in order to avoid its possible co-variation with final lengthening mentioned above, it was calculated as the number of uttered syllables per second before the final syllable in a target sentence.

3.3. Pitch height and pitch range

Pitch height and pitch range were also widely studied in previous research. In the present study, one wants to see whether and in what pattern they are used by Taiwan Mandarin speakers. Pitch height was quantified by maximum and minimum F_0 of a sentence, and the semitone difference of the two was taken as the measure for pitch range, whose formula is shown in (1).

$$\textit{pitch range} = 12 \cdot \log_2 (F_{0\max} / F_{0\min}) \qquad (1)$$

3.4. Peak alignment

Peak alignment is a recently proposed parameter [10], which is suggested to reflect the use of size code [11] by speakers to differentiate anger and joy. In [10], it was found that speakers reach the F_0 peak in accented syllable at significantly later points in time in anger than in joy, and the authors suggested that it was related to speakers' laryngeal lowering to exaggerate body size in anger speech. However, since the studied languages in [10] were English, German, French, Spanish, and Slovenian, none of which were tonal languages, their modulation of F_0 should be freer than that in Mandarin, the peak position variation of which might be inhibited by the dense occurrence of lexical tones. Therefore, the present study attempts to see whether this pattern of peak delay also exists in Taiwan Mandarin.

This cue was calculated as the proportion of time taken to reach F_0 peak relative to syllable duration of the last syllable in all the target sentences, which were all in Tone 4, a high-falling tone, and only those spoken with stress level 3, as defined by pan-Man-ToBI [12], were included.

4. Screening Test

In order to make sure that the collected utterances were representative enough for the designated emotions, a screening test was performed prior to acoustic analyses. Forty judges were instructed to listen to assigned utterances carefully through earphones, and then chose among the five emotion labels (anger, joy, sorrow, fear, and neutral). They were also asked to rate the emotional strength on a 5-point Likert scale, from 1, representing a weak degree, to 5, representing a strong degree. Each utterance was judged by five listeners for three times, which added up to 15 votes. Only those utterances receiving the same emotion label as the intended for at least 8 votes were included. In addition, their rated strengths needed to correlate positively with the intended level of emotion. In the end, there were 1117 utterances that fulfilled the criteria. Table 2 shows their distribution.

Table 2. *Distribution of utterances passing the screening test.*

	Neutral	Anger	Joy	Sorrow	Fear
neutral	183				
weak		129	108	162	0
strong		225	121	189	0

As can be seen from Table 2, after screening, none of the utterances in fear passed the test. As a matter of fact, difficulties in finding representative utterances for fear were also encountered by other studies [6]. It might have to do with the fact that it is not an emotion that speakers actively encode in speech. According to Ohala [11], fear is a state which "leaks out" when we are preparing ourselves to effectively deal with some threats, such as excessive perspiration and tremor in the voice when we are facing danger. This may be why the non-actresses in the present study had trouble in acting out this emotion, as their physiological states were not alert for any threats during recording.

5. Results

5.1. Final lengthening

Figure 1 shows the result of final lengthening. As can be seen, joy and neutral had the shortest final lengthening, while anger and sorrow the longest. Two ANOVAs were conducted for statistical analyses. The first was a one-way ANOVA comparing across neutral and different emotions at the weak level. Results showed that the main effect of emotion was significant [$F(3, 83) = 4.63$, $p < .01$, $\eta\hat{}^2 = .143$]. Post-hoc tests with Bonferroni adjustments showed that final lengthening was significantly shorter in neutral than in anger and sorrow ($p < .05$). The second ANOVA was a two-way emotion (3) × strength (2) analysis. Results showed that only the main effect of emotion was significant [$F(2, 138) = 4.46$, p

< .05, $\hat{\eta}^2$ = .061], and post-hoc tests with Bonferroni adjustments showed that final lengthening in joy was significantly shorter than that in anger and sorrow (p < .05). No significant effects involving strength were found.

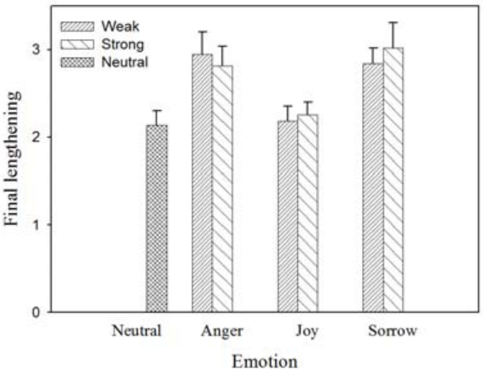

Figure 1: Results of final lengthening for each emotion.

5.2. Speech rate

Figure 2 shows the results of speech rate. As can be seen, neutral and sorrow had the slowest speech rate among all the emotions. Two ANOVAs were conducted for statistical analyses. The first one was a one-way ANOVA comparing across neutral and different emotions at the weak level. Results showed that the main effect was significant [F(3, 83) = 3.13, p < .05, $\hat{\eta}^2$ = .102]. Post-hoc tests with Bonferroni adjustments showed that sorrow were significantly slower in speech rate than anger (p < .05). The second ANOVA was a two-way emotion (3) × strength (2) analysis. Results showed that only the main effect of emotion was significant [F(2, 138) = 11.906, p < .01, $\hat{\eta}^2$ = .147]. Post-hoc tests with Bonferroni adjustments showed that speech rate in sorrow was significantly slower than that in anger and joy (p < .01).

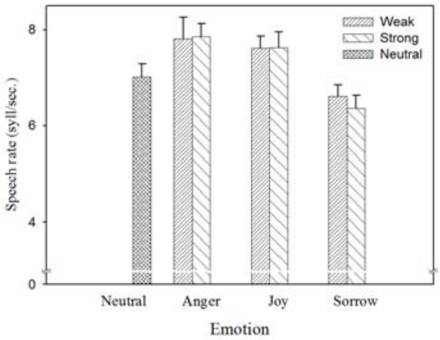

Figure 2: Results of speech rate for each emotion.

5.3. Pitch height and pitch range

Figure 3 shows the results of pitch height and pitch range in Hz. The general trend shows that both neutral and sorrow had lower register than anger and joy, but the two were different in range in that sorrow had a wider one. There is also a trend for the overall pitch register to become higher at the strong emotion level. An ANOVA comparing across neutral and different emotions at the weak level, and a two-way emotion (3) × strength (2) ANOVA were conducted for pitch range in semitone, maximum F_0, and minimum F_0, respectively. For pitch range, when comparing across neutral and different emotions, the main effect was near-significant [F(3, 83) = 2.53, p =.06, $\hat{\eta}^2$ = .084], but the post-hoc test with Bonferroni adjustments did not show any significant or near-significant results. The two-way emotion (3) × strength (2) analysis did not show any significant main effect. For maximum F_0, when comparing across neutral and different emotions, the main effect of emotion was significant [F(3,83) = 8.562, p < .01, $\hat{\eta}^2$ = .236]. Post-hoc tests with Bonferroni adjustments showed that the maximum F_0 in neutral was lower than that in anger (p < .05) and sorrow was significantly lower than that in anger and joy (p < .01). The two-way emotion (3) × strength (2) analysis on maximum F_0 showed that both emotion and strength had main effects [emotion: F(2,138) = 22.50, p < .01, $\hat{\eta}^2$ = .246; strength: F(1,138) = 14.00, p < .01, $\hat{\eta}^2$ = .092]. Post-hoc tests with Bonferroni adjustments concerning the emotion effect showed that anger and joy were higher than sorrow (p < .01), and anger was also marginally higher than joy (p = .07). For minimum F_0, when comparing across neutral and different emotions, the main effect was significant [F(3,83) = 5.72, p < .01, $\hat{\eta}^2$ = .171], where post-hoc tests with Bonferroni adjustments showed that the minimum F_0 in sorrow was significantly lower than that in neutral and anger (p < .01). The two-way emotion (3) × strength (2) analysis on minimum F_0 showed that both the main effects of emotion [F(2,138) = 20.82, p < .01, $\hat{\eta}^2$ = .232] and strength [F(1,138) = 6.66, p < .05, $\hat{\eta}^2$ = .045] were significant, and the interaction effect was also significant [F(2,138) = 3.61, p < .05, $\hat{\eta}^2$ = .050]. Post-hoc analyses concerning the interaction effect showed that for anger, the minimum F_0 at the strong level was significantly higher than that at the weak level (p < .01). When at the weak level, the minimum F_0 of anger was significantly higher than sorrow (p < .05); when at the strong level, the minimum F_0 in sorrow was significantly lower than that in anger and joy (p < .01).

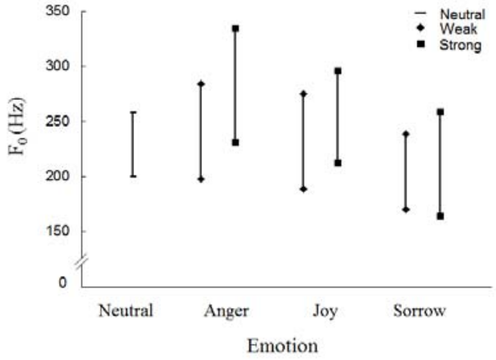

Figure 3: Results of pitch range, maximum F_0 and minimum F_0 for each emotion.

5.4. Peak alignment

Figure 4 shows the results of peak alignment. As can be seen from the graph, the amount of peak delay in neutral and anger is smaller than that in joy and sorrow. Also, if compared across different strength levels, it showed that the delay amount is bigger at the strong level than the weak level. A one-way ANOVA comparing across neutral and different emotions at the weak level was conducted, and the result showed no significant main effect. A two-way emotion (3) × strength (2) ANOVA was also conducted, and the result showed that there was a significant main effect of emotion (p < .05, $\hat{\eta}^2$ = .09). Post-hoc comparisons with Bonferroni adjustments showed that anger was near-significantly less in peak delay than sorrow and joy (p = .08). No effect concerning strength was found.

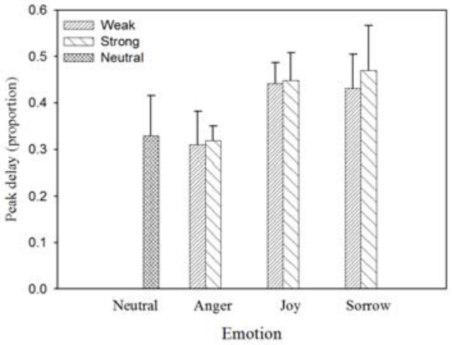

Figure 4: Results of peak delay for each emotion.

6. Discussion and Conclusions

The present study aimed to explore the prosodic and acoustic features of emotional speech in Taiwan Mandarin. Table 3 shows a summary of the current findings.

Table 3. *A summary of findings for each emotion. The "n.s." means no significant results in statistics.*

	Neutral	Anger	Joy	Sorrow
Final lengthening	↓	↑	↓	↑
Speech rate	n.s.	↑	↑	↓
F_0 max	↓	↑	↑	↓
		strong ↑ ; weak ↓		
F_0 min	n.s.	↑	↓	↓
F_0 range	n.s.			
Peak alignment	n.s.	↓	↑	↑

As can be seen, for anger, almost all the parameters were expanded or enhanced. This pattern is basically consistent with what was found in angry speech in previous studies. However, unlike what was proposed in [10], which found that the amount of peak delay is greater in anger due to the lowering larynx, results of the present study indicated the opposite. One possibility is that instead of using the size code to make themselves sound bigger, speakers of the present study cared more about making themselves sound more aroused in anger. Therefore, they applied the strategy of reaching F_0 peak higher and faster in anger. The need to maintain lexical tonal shapes might also be a pressure for speakers to abandon using the size code. However, more studies on the interaction between lexical tones and emotional speech would be needed to confirm the observation.

For joy, results showed that it had a similar pattern as anger expression, except that its final lengthening is shorter, maximum F_0 is lower, and F_0 peak was reached later. These suggest that though the two emotions were notoriously similar in many ways [3, 11] due to their high arousal natures, cues mentioned above can still be potentially good indicators in differentiating them.

For sorrow, it almost had a completely opposite trend in all the parameters from those in anger, except for longer final lengthening, compared to joy. This suggests that final lengthening may serve as a cue to differentiate emotions of different valences. However, since final lengthening was less studied previously, it is unclear whether this pattern also holds for other languages. More studies would be needed to see how robust the cue is for both speakers and listeners on the valence dimension.

In addition to examining prosodic patterns across different emotions, the present study also investigates the patterning of different emotional strengths. Results showed that maximum F_0 reflected this dimension. Different from other emotions, anger also had higher minimum F_0 at the strong level than its weak counterpart, indicating that raising overall pitch register is an important strategy for Taiwan Mandarin speakers to signal anger. This again suggests that the speakers in the present study cared more about making themselves sound more aroused than big in anger. To see whether the found patterns are robust and whether more patterns could be found, examination on more data would be needed in the future.

7. Acknowledgement

This study was supported by the National Science Council of Taiwan [NSC99-2221-E-002-204-]. The authors would like to thank the reviewers for their valuable comments. Naturally, all the faults are ours.

8. References

[1] Hirschberg, J., and Nakatani, C., "A prosodic analysis of discourse segments in direction-giving monologues", In Associal for Computational Linguistics, 1996.

[2] Russell, J. A. "A circumplex model of affect", J. Pers. Soc. Psychol., 39: 1161–1178, 1980.

[3] T. Johnstone and K.R. Scherer, "Vocal communication of emotion", in Lewis M. and Haviland-Jones J.M. [Ed], Handbook of Emotions, 220-235, New York: Guilford Press, 2000.

[4] Goudbeak, M. and Scherer, K., "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion", J. Acoust. Soc. Am., 128(3):1322-1336, 2010.

[5] E.D. Ross, Edmondson J.A. and Seibert G.B., "The effect of affect on various acoustic measures of prosody in tone and non-tone language: A comparison based on computer analysis of voice", Journal of Phonetics, 14: 283-302, 1986.

[6] J. Yuan, L. Shen, and, F. Chen, " The acoustic realization of anger, fear, joy and sadness in Chinese," Proc. of the 7th International Conference on Spoken Language Processing, Denver, USA, 2025-2028, 2002.

[7] S. Zhang, P.C. Ching, and F. Kong, "Acoustic analysis of emotional speech in Mandarin Chinese", International Symposium on Chinese Spoken Language processing, Kent Ridge, Singapore, 2006.

[8] C.-C. Tseng, "Prosodic Properties of Intonation in Two Major Varieties of Mandarin Chinese: Mainland China vs. Taiwan," International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, China, 28-31, 2004.

[9] I ChangLiao, and W.-Y. Chiang, A Study of the Influences of Emotion on Mandarin Tones, Unpublished Master Thesis. National Taiwan University, Taiwan, 2004.

[10] S. Chuenwattanapranithi, Y. Xu, B. Tipakorn, and S. Maneewongvatana, "Expressing anger and joy with size code", Proc. of the 3rd Internat. Conf. on Speech Prosody, Dresden, Germany, 487-490, 2006.

[11] J. J. Ohala, "An ethological perspective on common cross-language utilization of F_0 of voice," Phonetica, 41: 1-16, 1984.

[12] Peng, S. et al. A Pan-Mandarin ToBI. Online: http://deall.ohio-state.edu/chan.9/MToBI.htm, 2000.