Exploring Acoustic Differences between Cantonese (Tonal) and English (Non-Tonal) Spoken Expressions of Emotions

Chee Seng Chong, Jeesun Kim, Chris Davis

MARCS Institute, University of Western Sydney, Australia

L.chong@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

Abstract

It has been claimed that tone language speakers use less F0 related cues in the production of verbal expressions of emotions. This is because F0 is used in the production of lexical tones. This study investigated this claim by examining how F0 and various other acoustic parameters are used in the production of verbal emotion expressions in Cantonese (tone language) compared to English (non-tone language). Acoustic measurements (e.g., mean F0, F0 range) were extracted from the verbal expressions of five emotions (angry, happy, sad, surprise and disgust) and a neutral expression produced by five male native speakers of Cantonese and English. They were analyzed using K-means clustering to see how different acoustic properties are grouped and how this varies as a function of language. The results showed some difference between the two languages in how F0 related cues are used in the production of emotions. The results are discussed in terms of the general acoustic characteristics of spoken emotion expressions and in relation to behavioral data from perceptual studies.

Index Terms: vocal emotion perception, F0, acoustic measures, Cantonese, English

10.21437/Interspeech.2015-333

1. Introduction

Pitch or Fundamental Frequency (F0), has been identified as one of the most salient carriers of vocal emotion information in the English [1, 2]. However this may not be the case for tone languages because using F0 to express emotional prosody in expressive utterances may compromise the use of this code of lexical tone [3, 4, 5]. This idea is supported by a study that found that the variance of F0 in emotion expressions and in a neutral baseline was much smaller in Mandarin (a tone language with 4 tones) than in Italian (non-tone language) [6].

The current study followed up this finding because a number of issues were raised by the Mandarin/Italian study. The first concerned the notion of whether F0 would really be overloaded by having to code both emotion and lexical tone. In essence, it would depend on whether emotion and lexical tone was coded by the same type of variation in F0. In the Mandarin/Italian study [6], F0 variation was measured in terms of the range, minimum, mean, maximum and standard deviation of the F0. Of these measures, F0 range is quantified by the single point estimates of the minimum and maximum values and so is insensitive to the total variation in F0. Moreover the mean may not be a very good estimate because F0 is not normally distributed. Therefore, in the current study, we compared acoustic measures of emotion expression in a tone and non-tone language and included two additional measures, median F0 and number of F0 turning points. The number of turning points is defined as the number of peaks or troughs in the F0 contour over a single sentence. As duration may vary across sentences, each utterance was divided into 30 equal time points and the F0 value was sampled at each point giving a measure of F0 every 80 to 100 ms.

The second issue concerns whether the finding that the expressions of emotions in Mandarin have a smaller F0 variation than Italian applies to the comparison between tone and non-tone languages in general. As the authors in [6] have pointed out, the Italian language has a particularly expressive speech style, one that is rich in F0 variation, so Italian may not be a typical representative of a non-tone language. Moreover, there is a dearth of studies on the expressions of emotions in tone languages, so determining whether this restriction in F0 generalizes to other tone languages is a worthwhile endeavor. Here then, we examined the production of emotion in Cantonese, a tone language with 6 tones, with speakers of Australian English as the non-tone language comparison group. Cantonese is an interesting tone language to study as it has more lexical tones than Mandarin, so the linguistic system of this language may place greater demands on the use of F0.

Assuming that it is the case that all tone languages show restricted F0 variation in emotion expression, the third issue then becomes what is the best explanation for this? A recurring claim is that this restriction preserves the signal integrity of the lexical tones. However to our knowledge, this claim has never been verified, so we tested this idea by examining the differences in F0 variability of neutral expressions between tone and non-tone languages. The basic premise that underlies the emotion F0 restriction argument is that neutral expressions in Cantonese should have a larger F0 range than English, i.e., that the linguistic system has monopolized the use of F0. In other words, due to the constraints imposed by the linguistic system, there is a diminished opportunity for F0 to carry emotion information, thus leading to a restricted use of F0 in emotion expression.

The final issue concerns the role of F0 in emotion expression. Despite the finding that there is a difference in F0 variation in tone language expressions of emotions, no clear evidence was provided concerning the role that F0 plays in the production of emotions. As the authors of [6] pointed out, it may be although F0 is still used in Mandarin to some extent, other cues such as intensity and speech rate may be used to supplement F0. Alternatively, due to the restriction in F0, tone language users may utilize F0 cues in a manner different from non-tone language users. For example, it has been shown that Cantonese speakers raise their mean F0 to convey sarcasm while English speakers tend to lower their mean F0 [7]. So there is a possibility that the same F0 cues may be used to express different emotions depending on the language. To investigate the mix of acoustic factors used in emotion expression, we used k-means cluster

analysis to examine at how the different acoustic properties are grouped together and whether there is a difference in how F0 related properties are used across the languages.

2. Methods

2.1. Participants

Cantonese participants: Five male native speakers of Cantonese who were born and raised in Hong Kong were invited to participate for monetary reimbursement. The average age of the participants was 29.1 years.

English participants: Emotion expressions of five male native speakers of Australian English were obtained from our lab database. The average age of the participants was 23.0 years.

2.2. Materials

2.2.1. Speech Materials

Fifty semantically neutral sentences were chosen from the Cantonese Hearing In Noise (CHINT) sentences list [8] on the basis that they had a good spread of different tones at the initial and final position in the sentences (see for detailed procedure). All 50 sentences were recorded as expressive speech stimuli for the five basic emotions and neutral sentences.

Ten sentences selected from the Semantically Unpredictable Sentences [9] were recorded as the English expressive speech stimuli for five emotions (anger, disgust, happy, sad, surprise) and the neutral expression.

2.2.2. Production Setup

While only the audio recordings are used in this paper, the entire recording procedure including video recording is reported for completeness.

Participants were seated in front of a 20.1" LCD video monitor (Diamond Digital DV201B) that is used to present the stimulus sentences to the participant. Directly above the monitor is a video camera (Sony NXCAM HXR-NX30p) where participants are requested to fixate at prior to expressing the sentences. The videos were recorded at 1920 x 1080 full HD resolution at 50 fps. To capture participants utterances a microphone (AT 4033a Transformerless Capacitor Studio Microphone) was placed about 20 cm away from the participants lips and out of the field of view of the camera. Audio captured using the microphone was fed into the Motu Ultralite mk3 audio interface with FireWire connection to a PC running CueMix FX digital mixer and then to Audacity which captured the sound at a sampling rate of 48kHZ. This audio feed as well as video feed from the video camera was monitored by the experimenter outside of the booth who provided the participants with feedback as well as displaying the next sentence on the monitor in front of the participants.

The recording details of English speakers were similar to the Cantonese recordings (see [10] for details).

2.3. Procedure

2.3.1. Production of emotions

Since verbal expressions of emotions are often consciously and deliberately produced to convey emotions, rather than focusing on the experience of the emotion itself, we were interested in emotional expression; the signals that people present to others to express emotion. Given this, participants were instructed to be as natural as possible in how they expressed themselves and were asked to produce the emotions with the intent of communicating their emotional feelings to an observer. Moreover, through the course of the recording, the experimenter did not interfere, guide or give examples as to how each emotion type should be expressed to avoid demand characteristics and to preserve the natural idiosyncratic variation in the expression of emotions.

During the recording sessions, each stimulus sentence was displayed one at a time in a random order on the computer monitor and the participants then produced the utterances when ready. Participants were given feedback via the screen if they had to repeat the sentence (e.g., they misread the sentence or did not fixate on the camera while producing the expressions). Participants were also given a short story as a form of mood induction and three practice trials prior to the start of each emotion block and asked to put themselves in the mode of expressing the emotion. As was mentioned, the emotions to be expressed were blocked, so participants would produce all 50 sentences expressing the same emotion giving a total of 350 sentences per speaker (50 sentences x 7 (six emotions plus neutral).

2.4. Analysis

Acoustic parameters (mean F0, minimum F0, maximum F0, duration, maximum velocity, and mean intensity) of whole sentences were automatically extracted using Prosody Pro [11], a Praat script [12] with the lower threshold of F0 set at 50 Hz and the upper at 350 Hz. We then extracted the median, number of turning points and speech rate (duration/ number of syllables) in Matlab [13].

3. Results

3.1. Do Cantonese neutral expressions differ from English in F0 related measures?

Using a multivariate analysis of variance, Pillai's trace showed a significant effect of language on the mean, minimum, maximum, median and number of turning points of F0, V = 0.91, $F(4,5) = 7.72$, $p < 0.5$. However, separate univariate ANOVAs revealed non-significant effects of languages on each of the F0 measures, minimum, $F(1,8) = 0.61$, *ns*, maximum, $F(1,8) = 0.02$, *ns*, mean, $F(1,8) = 0.03$, *ns*, median, $F(1,8) = 0.01$, *ns*, mean, $F(1,8) = 0.03$, *ns*, number of turning points $F(1,8) = 0.92$, *ns*. Table 1 below shows the mean values for all of the measures.

Table 1: *The mean values of minimum, mean, maximum, median (in Hz) and number of turning points of neutral expressions.*

Language	F0 mean	F0 min	F0 max	F0 median	T.points
English	132.68	96.71	182.62	127.04	10.94
Cantonese	130.75	93.18	180.55	128.14	11.52

The MANOVA was followed up with a discriminant analysis which revealed that the language groups can be differentiated by mean F0 (b=1.47) and median F0 (b=-1.58), i.e., although there was no difference in F0 range, English neutral expressions had higher mean F0 but lower median F0 than the Cantonese expressions.

3.2. Do Cantonese emotion expressions use a smaller range of F0 related cues?

Separate MANOVAs revealed that except for happy, V = 0.84, $F(4,5) = 4.26$, *ns*., and surprise, V = 0.90, $F(4,5) = 7.22$, *ns*,

there was a significant difference between Cantonese and English on the F0 measures in expressions of angry, V = 0.97, $F(4,5) = 24.3$, $p < 0.01$, disgust, V = 0.96, $F(4,5) = 17.47$, $p < 0.01$, and sad, V = 0.98, $F(4,5) = 33.57$, $p < 0.01$. All significant effects were followed up with ANOVAs. Table 2 lists the results of the ANOVAs.

Table 2: *F and significance values for the difference between English and Cantonese emotion expressions on F0 measures.*

Emotion	min	max	mean	median	T.points
Angry	.16	.37	.10	.22	27.24 ***
Disgust	.83	4.3	.01	21.25 **	.42
Happy	*ns*	*ns*	*ns*	*ns*	*ns*
Sad	.69	.02	.25	.45	41.8 ***
Surprise	*ns*	*ns*	*ns*	*ns*	*ns*

Note: The values in the table are for F(1,8), *p < .05, **p < .01 ***p < .001.

In both angry and sad, English expressions had a significantly higher number of F0 turning points than Cantonese expressions, (14.24 vs. 11.23 and 12.66 vs. 10.35, respectively). As for disgust, median F0 was higher in English expressions (168.52 Hz vs. 120.38 Hz).

In summary, with regards to F0, the difference between English and Cantonese neutral expressions is captured by median F0 and mean F0, indicating that the difference between the two lie in the contour shapes of F0. The measures of median F0 and F0 turning points were able to capture the difference between English and Cantonese expressions of angry, disgust and sad. When compared with the F0 measures of the neutral expressions, it is clear that the median F0 and number of turning points in Cantonese emotion expressions did not vary much from the neutral expressions. On the contrary, English emotion expressions had a larger median F0 and number of turning points than the neutral expressions.

3.3. K-means clustering

The data consisted of 10 acoustic measures, (duration, speech rate, finalF0, minimumF0, maximumF0, maxF0velocity, meanF0, mean Intensity, median F0 and number of F0 turning points), extracted from 10 speakers by 5 emotion and 1 neutral expression giving a total of 1800 utterances. Separate k-means analyses were conducted for each language.

On the first pass using k=6 (emotions), instead of emotion type, the speakers themselves were identified as the most salient clusters, accounting for 86.3% (Cantonese) and 79.4% (English) of the variance. So we conducted separate k-means analysis for each of the speakers. Given the scope of the paper and that the only F0 measures that are meaningful in discriminating between Cantonese and English expressions are median and turning points; we focused the analysis on these two factors. Here we examined how these measures were clustered, specifically, we examined which emotion was classified to have the highest and lowest centroid means for our measures and if this differed between Cantonese and English. The cluster solution for each speaker is listed in Table 3.

Table 3: *Centroid means for the cluster solution for each speaker.*

Speaker	Median		T.points	
	highest	lowest	highest	lowest
Eng1	angry	happy	happy	neutral
Eng2	angry	happy	happy	neutral
Eng3	angry	happy	happy	neutral
Eng4	surprise	neutral	angry	surprise
Eng5	sad	neutral	surprise	neutral
Cant1	happy	neutral	neutral	angry
Cant2	angry	neutral	neutral	sad
Cant3	happy	disgust	disgust	sad
Cant4	sad	neutral	disgust	surprise
Cant5	sad	surprise	disgust	angry

From Table 3, it is clear that 3 out of the 5 English speakers in our study produce angry, happy and neutral quite similarly (Eng1, 2, and 3). Although there are only 5 speakers per group, the Cantonese speakers provided a stark contrast. First, among the Cantonese speakers, there was little agreement as to what a high median (2 happy, 2 sad, 1 angry) or low number of turning points correspond (2 sad, 2 angry, 1 surprise). Second, other than neutral having the lowest median value, there was little agreement between English and Cantonese speakers. However it is interesting to note that there was some similarity between Cantonese and English expressions such that, an emotion that had the lowest median F0 also tended to have a highest number of F0 turning points. This suggests that although expressions of emotions in Cantonese and English may utilize similar clusters of emotions, they do not convey the same emotion.

4. Discussion

In this study, we aimed to examine the differences between tone and non-tone languages, specifically in terms of the differences in F0 use. We followed up a study that examined the differences between Mandarin and Italian expressions of emotions by addressing 4 issues. First of all, we were concerned with how F0 variance may best be captured. Instead of relying on single point estimates such as range and mean F0, we included two additional measures that we judged would better capture the relevant aspect of variance, namely, the median F0 and F0 turning points. The median is a better estimate of the F0 because F0 is not normally distributed. Moreover it is more meaningful to define F0 in terms of fluctuations or contour change over time. Therefore an estimate of the turning points allowed us to capture this change over time at the sentential level. Indeed throughout the analysis conducted in this study, we found the median F0 scores and turning points to be the important F0 features that discriminated emotions and the language in which these were expressed. On this point we propose that it is essential for future studies on verbal expressions of emotions to use more sophisticated measures that take the contour of the F0 and its fluctuations over time into consideration.

Concerning the second issue of whether a restriction of F0 variation can be generalized to other tone and non-tone languages, we found partial support. This restriction was not observed across all emotions and what restriction we found was only for the median F0 and number of turns. Whether this constitutes as a restriction of variation is debatable because no dif-

ferences were observed on the other measures such as range and mean. Our results do however suggest that the F0 contour may be different, i.e., Cantonese expressions of emotions may be flatter or more monotonous than English emotion expressions. The third issue concerned how the linguistic properties of Cantonese as a tone language may affect F0 use in emotion expression. We examined and found no evidence that Cantonese use more F0 related cues resulting in a diminished capacity for F0 to carry emotion information. This finding is interesting as it is goes against the idea that tone languages may have a larger range or a larger number of turning points given the nature of its linguistic property. Discriminant analysis however suggests that a combination of F0 measures (mean and median) can distinguish Cantonese from English neutral expressions. This once again emphases the point that single measures of F0 may be uninformative, as it is the variation of F0 over time that appear to be more important.

Given that in terms of F0 use, Cantonese neutral expressions were similar to those in English, it follows that Cantonese should have the freedom to use F0 for emotion prosody in a manner similar to English. Yet, the data showed that instead of using a fuller F0 space (or at least as much as English emotion expressions do), Cantonese expressions of emotions used less F0 variation. This seems likely due to the need for Cantonese speakers to modulate their use of F0 to preserve the signal properties of lexical tones, although here a more sophisticated measure of the mix and interaction of tonal acoustic properties are needed.

Finally with regards to the role that F0 plays, we examined how the clusters of acoustic properties may differ in Cantonese and English expressions of emotions. In this paper we only examined median F0 and turning points. The solutions for our k-means cluster analysis showed that these measures were similarly grouped in Cantonese and English expressions of emotions. However, interestingly these clusters encoded different emotions. This was particularly the case if we considered expressions that had low median F0 (neutral for Cantonese and Happy for English). As for high median F0, the clusters corresponded to the expression of anger in English but did not fit any particular emotion class for Cantonese expressions. This is also interesting because this finding fits the previous observation that Cantonese speakers may use less F0 variation. So instead of coding very active expressions like angry that is usually associated with high F0, it may be that Cantonese uses F0 mainly for conveying less active emotions like happy or neutral, and may use a more moderate or mid-range F0 that can convey emotion without distorting the boundaries of lexical tones. Whereas for emotions like anger or sadness that is usually coded by the maximum or minimum F0, features such as speech rate and intensity may be used as the most salient feature instead.

The next step forward is to properly test some of these ideas and predictions. Firstly, we plan to fully explore the k-means clustering solution and how differences in F0 measures affect other properties such as speech rate. We will also investigate the possibility of using other methods of quantifying F0 measures to better capture the variance of F0. Finally, we intend to correlate the production data and the k-means solution with perception data. In the literature, it is a common finding that people are poorer at recognizing emotions produced in a language that one is not familiar with [14, 15]. This effect often produces systematic perceptual confusion between emotions. We propose that part of this confusion may be due to linguistic effects on emotion production. By combining the results of production and perception studies, we aim to provide a more unified understanding of how linguistic properties affect the production of emotion expressions.

5. Conclusions

In conclusion, our results gave some support for the proposal that emotion expression in tone languages has less F0 variation than in non-tone languages. However, a rather complex picture emerged in which the effects of the linguistic system on emotion expression still remain elusive.

6. References

[1] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?" *Psychological bulletin*, vol. 129, no. 5, pp. 770–814, Sep. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12956543

[2] K. R. Scherer and J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and Emotion*, vol. 1, no. 4, pp. 331–346, Dec. 1977. [Online]. Available: http://link.springer.com/10.1007/BF00992539

[3] E. D. Ross, A. E. Jerold, and G. 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.

[4] S. Luksaneeyanawin, "Intonation in Thai," *D. Hirst and AD Cristo, Intonation Systems A Survey of Twenty Language*, pp. 376–394, 1998.

[5] 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, no. 1, pp. EL117–EL123, 2015.

[6] L. Anolli, F. Mantovani, and a. De Toni, "The Voice of Emotion in Chinese and Italian Young Adults," *Journal of Cross-Cultural Psychology*, vol. 39, no. 5, pp. 565–598, Sep. 2008. [Online]. Available: http://jcc.sagepub.com/cgi/doi/10.1177/0022022108321178

[7] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in Cantonese and English." *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1394–405, Sep. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19739753

[8] L. L. N. Wong and S. D. Soli, "Development of the Cantonese Hearing In Noise Test (CHINT)." *Ear and hearing*, vol. 26, no. 3, pp. 276–89, Jun. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15937409

[9] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[10] J. Kim and C. Davis, "Perceiving emotion from a talker: How face and voice work together," *Visual Cognition*, vol. 20, no. 8, pp. 902–921, 2012.

[11] I. Xu, "ProsodyPro A Tool for Large-scale Systematic Prosody Analysis.itle," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, France, 2013, pp. 7–10.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," 2014. [Online]. Available: http://www.praat.org/

[13] "Matlab R2013a."

[14] C. S. Chong, J. Kim, and C. Davis, "The effect of expression clarity and presentation modality on non-native vocal emotion perception," in *The 17th conference of the Oriental chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment /CASLRE (Conference on Asian Spoken Language Research and Evaluation)*. Phuket, Thailand: IEEE, 2014.

[15] K. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76–92, Jan. 2001. [Online]. Available: http://jcc.sagepub.com/cgi/doi/10.1177/0022022101032001009