# Body size projection by voice quality in emotional speech—Evidence from Mandarin Chinese

*Xiaoluan Liu, Yi Xu*

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
xiaoluan.liu.12@ucl.ac.uk, yi.xu@ucl.ac.uk

## Abstract

This study attempts to extend the line of research on using body size projection theory to account for emotional speech. It is predicted by the theory that anger is expressed by projecting a large body size with low pitch, rough voice and long vocal tract; happiness is expressed by projecting a small body size with high pitch, breathy voice and short vocal tract. Ten native speakers of Mandarin with drama training background recorded sentences in happy, angry, disgust and neutral emotions. We used multiple measurements to assess voice quality, formant dispersion (as an indicator of vocal tract length) and pitch. The results show clear support for the body size projection theory in voice quality, with anger and disgust associated with pressed and rough voice while happiness with breathy voice. But the results of formant dispersion and pitch demonstrate no clear directions. While the study is the first to show clear speech production support for the body size projection theory with voice quality data, the equivocal results of formant and pitch call for improvement in the method of emotion elicitation in the laboratory.

Index terms: emotional speech, body size projection theory, Mandarin Chinese

## 1. Introduction

Speech is one of the most important means of expressing emotions. While there have been many studies on the acoustic characteristics of emotional speech, the findings have generally been mixed [12]. One particular line of research, based on the body size projection theory, however, has taken a rather different theoretic approach to emotional speech. Originally proposed by Morton [8] for explaining animal calls, and later extended by Ohala [11] to human speech, the key idea is that vocal emotional expressions are a mechanism evolved under a selection pressure to influence the behaviour of other individuals in social interactions. For example, an angry expression is to project a large body size to scare off the opponents in case of confrontation; a happy expression is to project a small body size to attract the listener. Recently, this idea has seen support from a series of perception research in which the speech stimuli are synthetically manipulated in terms of pitch, vocal tract length and voice quality to simulate body size projection [3, 17, 18]. It is shown that listeners hear speech with synthetic parameters that project a large body size both as being spoken by a large person and as expressing anger. And they hear speech that projects a small body size as spoken by a small person and expressing happiness and friendliness [10, 17, 18].

In Xu et al. [17] an extension to the body size projection principle is proposed to incorporate additional "bio-informational dimensions" that may also serve to influence listener behaviour. These include dynamicity, audibility and association. One of these dimensions, namely, dynamicity, has been tested and shown to be relevant to the perception of a number of emotions [17, 18]. So far, however, there has not been systematic testing of either size projection or any of the other bio-informational dimensions in production studies. Thus, despite the demonstrated listener sensitivity to some of them, it is not yet known whether and how consistently speakers use any of them in the production of emotional speech. Also, the bio-informational dimensions have been tested on anger, happiness, sadness and fear [17], but not yet on disgust, which is also considered as one of the basic emotions [4].

Identifying acoustic properties that are clearly emotion-relevant from production data has not been easy, however, partly because it is generally difficult to elicit genuine emotions in the laboratory, even from trained actors and actresses. Recently, however, an emotion portrayal method has been developed for inducing emotions, and it is argued that the method can induce reliable and natural emotions in laboratory conditions because it reflects people's daily strategy to control and display emotions [12,13].

The present paper reports the results of a production study to test the bio-informational dimensions of emotion, using the emotion portrayal as the induction method, and Mandarin Chinese as the target language. Three emotional expressions are examined—anger, disgust and happiness, with the aim to find out whether the predictions of the theory can be confirmed in speech production, and how effective emotion portrayal is as a method of emotion induction in the laboratory.

## 2. Methodology

The basic design of the study is to have native speakers of Mandarin produce sentences with anger, disgust, happiness and neutral emotion using emotion portrayal as the induction method.

The stimuli, as shown in Table 1, consist of four Mandarin sentences with the target words *mao* and *men* imbedded in sentence-medial and sentence-final positions. The selection of *men* and *mao* is for the ease of phonetic segmentation while minimizing consonantal perturbation of $F_0$. To test for interaction between tone and emotion, we assigned the key word *mao* with four lexical tones: High tone *mao1* for sentence 1, rising tone *mao2* for sentence 2, low tone *mao3*

for sentence 3 and falling tone *mao4* for sentence 4. The syllable *men* following *mao* is assigned with tone 5 (neutral tone) for semantic/pragmatic naturalness of the sentences constructed. Note that *xiaomaomen* in the sentences is a compound word with three syllables, denoting the name of a brand and their album, although *mao* and *men* when used separately have their own independent meanings.

Table 1. Stimuli of the experiment. The number in each syllable represents the lexical tone: 1 for H (High tone), 2 for R (Rising tone), 3 for L (Low tone), 4 for F (Falling tone), and 5 for N (Neutral tone).

| | | | | | | |
|---|---|---|---|---|---|---|
| *xiao3* little | *mao1* cat | *men5* particle | *yue4dui4 fa1xing2 le5* | *xiao3* little | *mao1* cat | *men5* particle |
| | *mao2* fur | | | | *mao2* fur | |
| | *mao3* mortise | | band release particle | | *mao3* mortise | |
| | *mao4* hat | | | | *mao4* hat | |
| The band *Xiaomaomen* has released the album *Xiaomaomen*. | | | | | | |

Ten native Mandarin speakers with drama training background were recruited as subjects. They reported no speech or hearing problems. Emotion portrayal method was used to induce emotion, i.e., having subjects imagine themselves being in an emotional state as vividly as possible when saying each sentence [12, 13]. The recording was conducted in a sound-controlled booth. All the sentences were repeated 3 times by 10 subjects in anger, disgust, happiness and neutral emotion, resulting in 4 (*mao1/2/3/4*) * 4 (emotions) *10 (subjects) *3 (repetitions) = 480 sentences.

A customized version of Prosody Pro [16], running under Praat [1], was used to extract and analyse F0 contours, voice quality and formant dispersion (as indicator of vocal tract length [5]). Segmental boundaries were labelled by hands with visual inspection and listening validation. Given the difficulty in assessing voice quality, we obtained multiple measurements used in previous studies as follows:

$H_1$-$H_2$*, $H_1$-$A_1$*, and $H_1$-$A_3$* ($H_1$ and $H_2$ refer to the amplitude of the first and second harmonics of voiced segments; $A_1$ and $A_3$ refer to the amplitude of the first and third formants. The * symbol indicates that the measurements are estimates based on frequency bands without literally tracking harmonics and formants.);

Centre of spectral gravity—A measure for how high the frequencies in a spectrum are on average [1];

Energy of voiced segments below 500Hz and 1000Hz [14];

Hammarberg index—Maximum energy difference between the range of 0-2000 Hz and the range of 2000Hz to 5000Hz in the voiced section of the speech under examination [7];

Skewness—A measure for how much the shape of the spectrum below the centre of gravity is different from the shape above the mean frequency [1].

The above five measurements are indicators of spectral slope which is directly related to voice quality [6].

Jitter—Mean absolute difference between consecutive periods, divided by the mean period [6];

Shimmer—Mean absolute difference between the amplitudes of consecutive periods, divided by the mean amplitude [6];

Harmonicity—Harmonics-to-noise ratio measuring "the extent of acoustic periodicity expressed in dB" [6].

The above three measurements are indicators of voice roughness. Each of them indicates in one way or another the amount of aperiodicity in the voice.

Formant dispersion—Mean frequency differences between adjacent formants, which is an indicator of vocal tract length [5, 11].

The perception test by Xu et al. [18] shows that anger is associated with smaller formant dispersion than happiness. Therefore, this is a vocal dimension worth examination as well in this study.

F0 values were measured in semitones to reduce the bias towards higher pitch range over lower pitch range.

## 3. Results

The measurements used in statistical analyses were taken from the target syllables *mao* and *men* (both sentence medial and final across the four emotions). A series of three-way repeated measures ANOVAs were performed on the measurements of voice quality, formant dispersion and F0 of these syllables. The independent variables were emotion (anger, disgust, happiness and neutral), tone of *mao* (high, rising, low and falling), and sentence position (medial and final). A series of post-hoc Tukey HSD tests were also conducted to examine which pair of the emotions was significantly different.

The ANOVA results on tone, sentence positions and the interaction between tone, sentence positions and emotion are non-significant (hence not displayed), suggesting that tone and sentence position of the target syllable (*mao* and *men*) do not affect the acoustic characteristics of the emotional speech. In contrast, emotion is found to affect all features except formant dispersion (Table 2a-2b).

Table 2a. Means and *p* values of ANOVAs for $H_1$-$H_2$*, $H_1$-$A_1$*, $H_1$-$A_3$*, centre of gravity (COG), formant dispersion (FD), jitter (JI), shimmer (SH) and harmonicity (HA) of the four types of emotional speech (A=anger, D=disgust, H=happiness, N=neutral).

| | H1-H2* | H1-A1* | H1-A3* | COG | JI | SH | HA |
|---|---|---|---|---|---|---|---|
| A | -3.1 | -3.9 | 21.9 | 882.7 | 0.06 | 0.22 | 8.9 |
| D | -2.6 | -3.8 | 22.8 | 876.6 | 0.05 | 0.21 | 9.11 |
| H | -0.1 | -1.8 | 26.1 | 811.6 | 0.05 | 0.2 | 10.24 |
| N | -0.3 | 0.7 | 31.1 | 608.4 | 0.04 | 0.19 | 11.58 |
| *p* | <.05 | <.05 | <.05 | <.05 | <.05 | <.05 | <.05 |
| F | 3.84 | 3.92 | 4.01 | 4.06 | 4.35 | 3.65 | 3.98 |
| df | 3,27 | 3,27 | 3,27 | 3,27 | 3,27 | 3,27 | 3,27 |

With regard to $H_1$-$H_2$*, $H_1$-$A_1$*, and $H_1$-$A_3$*, it has been reported [18] that the decrease in the values of the three parameters results in not only an increase in perceived body size but also a change of perceived emotion from happiness to anger. This is in the same direction as the results of this study. As can be observed from Table 2a, the smallest values all correspond to anger, indicating an exaggerated

body size. As the values go up, the emotions produced change gradually from anger to disgust, happiness and neutral emotion.

Table 2b. Means and *p* values of ANOVAs for energy below 500/1000Hz (E<500Hz, E<1000Hz), Hammarberg index (HI), skewness (SK), formant dispersion (FD) and F0 (semitone) of the four types of emotional speech (A=anger, D=disgust, H=happiness, N=neutral).

| | E <500 | E <1000 | HI | SK | FD | F0 |
|---|---|---|---|---|---|---|
| A | 0.22 | 0.65 | 21.09 | 1.1 | 773.8 | 96.07 |
| D | 0.44 | 0.82 | 22.74 | 1.3 | 762.7 | 89.45 |
| H | 0.33 | 0.7 | 23.46 | 2.31 | 756.6 | 93.93 |
| N | 0.55 | 0.86 | 27.31 | 2.04 | 760.9 | 92.11 |
| *p* | <.05 | <.05 | <.05 | <.05 | >.05 | <.05 |
| F | 4.09 | 4.16 | 3.52 | 3.67 | 2.79 | 3.86 |
| df | 3,27 | 3,27 | 3,27 | 3,27 | 3,27 | 3,27 |

Table 3. Results of Post hoc Tukey HSD tests, showing pairs of emotions that are significantly different (labelled by ✓) in terms of each of the parameters (A=anger, D=disgust, H=happiness, N=neutral emotion).

| | A vs. D | A vs. H | A vs. N | D vs. H | D vs. N | H vs. N |
|---|---|---|---|---|---|---|
| H1-H2* | | ✓ | ✓ | ✓ | ✓ | ✓ |
| H1-A1* | | ✓ | ✓ | ✓ | ✓ | ✓ |
| H1-A3* | | ✓ | ✓ | ✓ | ✓ | ✓ |
| COG | | ✓ | ✓ | ✓ | ✓ | ✓ |
| E<500Hz | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| E<1000HZ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| HI | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| JI | | ✓ | ✓ | ✓ | ✓ | ✓ |
| SH | | ✓ | ✓ | ✓ | ✓ | ✓ |
| HA | | ✓ | ✓ | ✓ | ✓ | ✓ |
| FD | | | | | | |
| F0 | ✓ | ✓ | ✓ | | ✓ | ✓ |

In terms of centre of spectral gravity, the prediction from perception findings [10, 18] is that it should show a decrease in values from anger to happiness. This is confirmed in Table 2a: Anger and disgust both have higher values than happiness, again indicating a larger projected body size while happiness has a low value, indicating a smaller projected body size.

Jitter and shimmer show higher values for anger than happiness (Table 2a), which, though not specifically predicted before, is in line with Morton's [8] hypothesis that aggressiveness is associated with rough voice.

Harmonicity (Table 2a) is the highest for neutral emotion and lowest for anger, with happiness in the middle having slightly higher values than disgust. Table 2b shows that energy below 500Hz and 1000Hz of happy speech is higher than that of angry speech, and that the tendency is also true with regard to Hammarberg index. As for skewness, happiness has the highest degree of skewness, followed by neutral and disgust emotion, with anger having the smallest skewness (Table 2b). These measurements are, therefore, consistent with the results of $H_1$-$H_2$*, $H_1$-$A_1$* and $H_1$-$A_3$* and hence with the body size projection theory.

The difference between anger and disgust is non-significant in most of the parameters mentioned above (Table 3), indicating a tendency of anger and disgust being positioned towards the same end of the body size dimension.

The differences in formant dispersion, however, are non-significant between different emotions (Table 2b and 3). Also, Table 2b shows that anger has a higher pitch than happiness and neutral emotion, with disgust having the lowest pitch. This is not consistent with the findings of previous perception studies.

## 4. Discussions

The results of voice quality are consistent with the body size hypothesis in that anger is found to project a large body size with pressed (as indicated by all the spectral tilt measurements) and rough (as indicated by higher jitter and shimmer but lower harmonicity) voice. Happiness, in contrast, is found to project a small body size with breathier and more harmonious voice. As argued in [18], with its greater spectral tilt, a breathy voice approximates a "pure tone" voice which is observed by Morton in animal calls associated with appeasement and sociability [8]. So this study is the first to show clear speech production support for the body size projection theory with voice quality data. Somewhat surprisingly, however, neutral emotion shows greater breathiness than happiness, which has not previously been predicted. It is possible that voice quality is affected not only by body size projection, but also by vocal effort, because anger, happiness and disgust are all more *activated* than neutral voice.

The non-significant difference in formant dispersion across the emotions is somewhat puzzling. It could be that the prediction of the body size projection theory is wrong, but it could also be due to the possibility that emotion portrayal method implemented in this study is not powerful enough to make speakers alter their vocal tract length when trying to produce emotional prosody. If so, it may suggest that voice quality is the most easily elicited emotion-relevant acoustic changes in speakers. But this needs confirmation from future research.

The higher pitch in anger than in happiness, as shown in Table 2b, may present a problem for the body size projection theory. But many previous studies have found that anger is associated with a higher pitch than neutral emotion [9, 12, 15]. This suggests that anger also projects high dynamicity through vocalization [17]. As speculated in [17], high dynamicity may serve to convince the listeners of the high energy the signaller possesses, thus helping to frighten them away. On the other hand, however, there are also different types of anger. As the participants in this study were instructed to portray "hot anger", nearly all of them managed to achieve it. In contrast, the happiness portrayed by the participants was more similar to calm happiness/pleasantness than to elation. This may further explain why the average pitch is higher for anger than for happiness.

The current results also show that disgust is similar to anger in terms of voice quality. This, together with the fact that the $F_0$ of disgust is the lowest among the four emotions, suggests that disgust involves the projection of a large body size, as has been speculated in [17].

## 5. Conclusions

In this study, we have used speech production experiments on Mandarin to test the body size projection theory of emotional speech. The results suggest that, firstly, tone and sentence position do not contribute significantly to the acoustic characteristics of emotional speech; secondly, disgust is similar to anger in terms of body size projection; thirdly and most importantly, among the features of emotional speech, only voice quality is fully consistent with the predictions of the body size theory: Pressed and rough voice, both projecting a large body size, is associated with anger; breathy voice, which projects a small body size, is associated with happiness. Pitch and formant dispersion did not generate results either in support or opposition to the body size projection theory. This could be due to insufficient authenticity of emotional speech under laboratory conditions, for which solutions need to be sought in future research. Overall, nevertheless, it is interesting, and in fact slightly surprising that voice quality turns out to be the vocal dimension that is the most easily elicited by the emotion portrayal method in this study. Naturally, further explorations in future research are needed.

## 6. References:


[1] Boersma, P. and Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program]. Version 5.3.59, retrieved 3rd Janurary 2013 from http://www.praat.org/

[2] Chen, Y. and Xu, Y. (2006). Production of weak elements: Evidence from neutral tone in Standard Chinese. *Phonetica*, 63, 47–75.

[3] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code—A perceptual investigation. *Phonetica*, 65, 210-230.

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

[5] Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America*, 102, 1213-1222.

[6] Goudbeek M. and Scherer K. R. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *Journal of the Acoustical Society of America*, 128, 1322–1336.

[7] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities, *Acta Otolaryngologica*, 90, 441-451.

[8] Morton, E. W. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, 111, 855-869.

[9] Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93, 1097-1108.

[10] Noble, L. and Xu, Y. (2011). Friendly Speech and Happy Speech – Are they the same? In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 1502-1505.

[11] Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41, 1-16.

[12] Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40, 227–256.

[13] Scherer, K. R. (2013). Vocal markers of emotion: comparing induction and acting elicitation. *Computer Speech and Language*, 27, 40–58.

[14] Van Bezooijen, R. (1984). *The characteristics and recognizability of vocal expressions of emotion. Dordrecht*. The Netherlands: Foris.

[15] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48, 1162-1181.

[16] Xu, Y. (2012). ProsodyPro.praat. University College London, London, UK.

[17] Xu, Y., Kelly, A. and Smillie, C. (2013a). Emotional expressions as communicative signals. In S. Hancil and D. Hirst (eds.) Prosody and Iconicity, John Benjamins Publishing Co, pp. 33-60.

[18] Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013b). Human vocal attractiveness as signaled by body size projection. *PLoS ONE*, 8(4), e62397.


.