EXPRESSING ANGER AND JOY WITH THE SIZE CODE

Suthathip Chuenwattanapranithi[1,2], Yi Xu[2,3], Bundit Thipakorn[1], and Songrit Maneewongvatana[1]

King Mongkut's University of Technology Thonburi, Bangkok, Thailand[1]
University College London, London, United Kingdom[2]
Haskins Laboratories, New Haven, CT, USA[3]

ABSTRACT

This paper reports our finding of the use of a proposed biological code – the size code in anger and joy speech. In searching for explanations for an F_0 peak delay phenomenon related to angry speech that cannot be accounted for by known articulatory constraints, we hypothesized that the delay was due to the lowering of the larynx to exaggerate body size, a biological code known to be used by animals. Our analysis of the formant frequencies in existing emotional speech databases revealed that anger speech had lowered formants and joy speech had raised formants. The results confirm our hypothesis and suggest that the size code is being actively used by humans to express emotions.

1. INTRODUCTION

Human speech conveys multiple layers of information [1]. In addition to the "linguistic" messages, speech signal also carries information about the identity, age, geographical origin, attitude, and emotional state of the speaker. Here we are especially interested in how emotions are encoded in the speech signal. Experimental results from previous work suggest that some acoustic features are associated with the general characteristics of emotional state rather than a specific emotion [2]. For example, high-activation emotions such as anger and joy have similar characteristics, such as greater loudness, higher pitch, and faster speed than low-activation emotions such as sadness. It has therefore been difficult to separate anger and joy with simple acoustical parameters. One possibility that has not yet been seriously explored is the use of "biological codes" proposed by Ohala [3]. For example, F_0 of animal voice is inversely related to their body size, and this relation is shown to be used actively to exaggerate body size to other animals [4, 5]. It has been suggested that humans also exploit this relation and use low F_0 to sound assertive and authoritative and high F_0 to sound unthreatening or deferential [3]. Further along this line, the descent of the human larynx is proposed to have been originally driven by natural selection to exaggerate body size [6], a proposal that has found support in the finding of drastically descended larynx in other animals like the red deer [7]. Similarly, the smile face in human has been proposed to be related to the shortening of the vocal tract so as to sound submissive [3]. Such exaggeration or understatement of body size in communication may be referred to as the "size code" [8].

In this paper, we report evidence that the size code is actively used in expressing anger and joy in human speech. Anger is expressed by exaggerating the body size to sound authoritative and threatening, whereas joy is expressed by understating the body size to sound unimposing and sociable. The evidence is found in two acoustic cues, F_0 and formant frequency.

2. F_0 CONTOURS

In this section, we investigate the F_0 contour patterns for anger and joy speech. Six parameters that specify detailed pitch contours, maximum pitch, minimum pitch, pitch range, rising strength, falling strength, and peak alignment, were taken from accented and unaccented syllables in anger and joy speech samples. According to the work of Banziger et al. [9], accented syllables are the ones which have local maximum pitch value located between two local minimum pitch values. Figure 1 shows the pitch contours of accented (/grænd/) and unaccented (/tʃɪl/) syllables of the word "grandchildren". The following measurements were taken using Praat [10].

- *Minimum and maximum F_0*: For accented syllables, there are two local minimum values (min1, min2).
- Pitch range: Measured as the distance between maximum and minimum pitch values in a syllable (also known as excursion size). It is measured in semitone in order to make the data from individual speakers more comparable.

$$rF0(\,semi\,tone\,)=(\frac{F0_{max}}{F0_{min}})\times(\frac{1}{2^{1/12}}) \qquad (1)$$

where *rF0* is F_0 range in semi tone unit $F0_{max}$,and $F0_{min}$ are the maximum and local minimum of F_0 in the syllable, respectively.

- *Rising and falling strengths*: First, velocity of F_0 curves were computed by taking the first order derivative of each F_0 curve (Eq. 2). After that, linear regressions were performed to obtain linear equations with excursion size as the predictor (x) and velocity as the dependent variable (Eq. 3). Then, the values of velocity (v) for two F_0 curves at a given excursion size (x) were compared.

$$v(t) = \frac{d}{dt} f_0(t) \quad (2)$$

$$v(x) \approx ax + b \quad (3)$$

where $v(t)$ is the velocity curve, $f_0(t)$ is the F_0 contour, and $v(x)$ is the linearly approximated velocity when excursion size is equal to x.

- *Peak alignment:* The proportion of time taken to reach maximum pitch value relative to syllable duration. This value is calculated only for the accented syllables.

$$Pk = \frac{(t_{max} - t_0)}{(t_{end} - t_0)} \quad (4)$$

where Pk is the peak alignment, t_{max} is the time of the maximum F_0, t_0 is the time of the first local minimum F_0, and t_{end} is the time of the second local minimum F_0.

Speech samples used in this work are obtained from English, German, French, Spanish, and Slovenian emotional speech databases which are publicly available [11, 12] and have been validated by human listeners. Each database consists of words or sentences of male and female speakers classified into four classes of emotional states: anger, joy, sadness, and neutral. To assess the perceptibility of the emotions in these databases, we performed an experiment to see how speakers would judge the emotional class of each sample word or sentences. The criterion for choosing the sample words or sentences is that their meanings would not lead human subjects to guess their emotional class. The perception experiment was performed by 20 Thai listeners who were not familiar with the languages of the databases. Each subject was asked to identify the emotional state of each speech sample by hearing the words and sentences one at a time. They were able to distinguish anger from joy at the rate of 74.58%. This contrast sharply with the performance of previous recognition algorithms which is no more than 55.45% [13]

After measuring the parameters from all anger and joy samples, we found that the same emotional state may have alternative acoustic manifestations [13]. These manifestations can be described as different strategies as shown in Figure 2. Paired t tests were performed to examine if there are differences between pairs of anger and joy syllables uttered by the same speaker. The first strategy we have found involves pitch range. With this strategy, pitch range is wider for joy than for anger ($p < 0.0001$) and also with higher strength for both F_0 rises and falls ($p < 0.0001$ and $p = 0.0002$, respectively). The second strategy involves F_0 peak alignment. With this strategy the maximum pitch is reached earlier for joy than for anger ($p = 0.0016$) and the falling strangth for anger is higher than joy ($p = 0.0232$). The third strategy is for unaccented syllables. Speakers produced low pitch for both anger and joy but the strength of falling pitch is higher in anger than in joy ($p = 0.0374$).

In strategy 2 and 3, the values of falling strength seem to suggest that there is a greater pitch falling tendency in anger speech than in joy speech and the former has greater F_0 peak delay than the latter, as shown in Table 1. This peak delay cannot be explained by known articulatory constraints. This is because lowering F_0 even by 1 octave takes only 170 ms, which can be easily achieved within the duration of a syllable [14]. Some extra mechanisms may be involved. We hypothesized that the delay was related to the lowering of the larynx to exaggerate body size. This hypothesis will be tested in the next section.

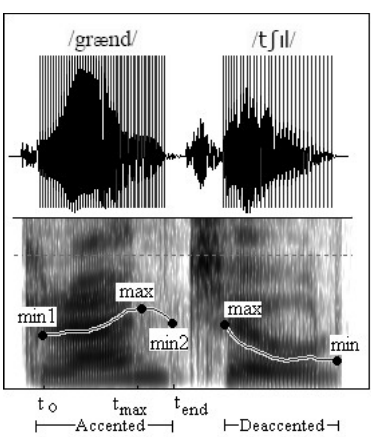


Figure 1: *Pitch contours of accented and unaccented syllables.*

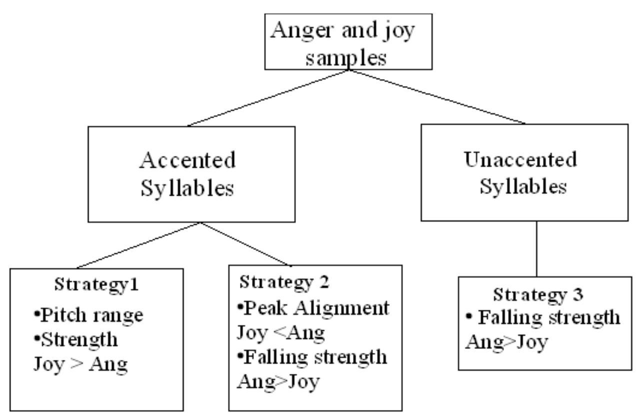


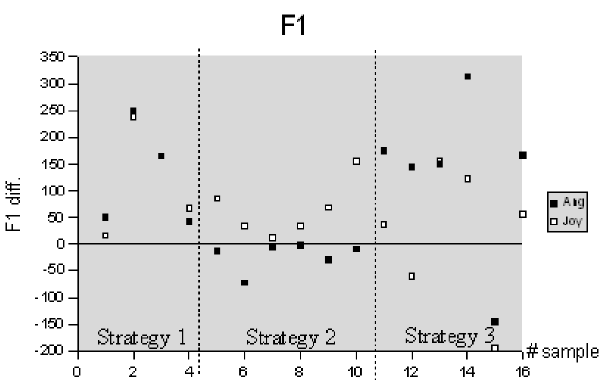
Figure 2: *Multi-strategy classification method for differentiating anger and joy in speech.*

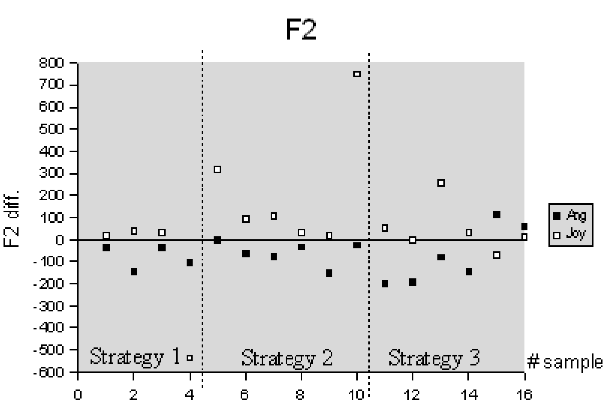
Table 1: *The comparison of mean values of falling range and peak alignment between anger and joy speech in strategy 1, 2, and 3.*

Strategy	Emotion	F_0 falling range (Hz.)	Pk
1	Anger	199.59-155.27	0.3091
	Joy	292.06-197.88	0.3687
2	Anger	219.12-149.05	0.4113
	Joy	288.36-196.68	0.2813
3	Anger	240.68-169.45	-
	Joy	332.58-263.12	-

3. FORMANT FREQUENCIES

To investigate the variation of formant related to anger and joy in speech, the frequencies of the first three formants were taken from accented and unaccented syllables using Praat. The differences of the first three formants from neutral, anger and joy speech are shown in Figure 3. As can be seen, in anger speech, F_2 and F_3 are lower in all three strategies, while F_1 is lower in strategy 2 but higher in strategy 1 and 3. In joy speech, all formants are higher than neutral emotion in every strategy.





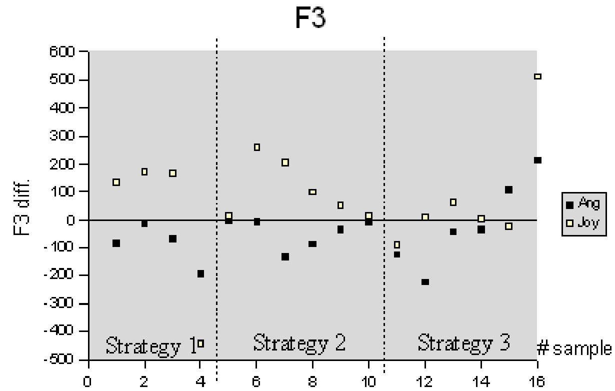


Figure 3: *The differences of the first three formants from neutral for anger and joy speech.*

The uniformly higher formants in joy speech may reflect vocal tract shortening due to smile [3]. While the higher F_1 in strategy 1 and 3 of anger speech could be due to wider opening of the jaw [15]. The lowering of F_2 and F_3 in all strategies of anger speech and the lowering of F_1 in strategy 2 could be related to vocal tract elongation due to lowering of the larynx. In strategy 2 of anger speech, in particular, as hypothesized earlier, the F_0 peak delay could be due to a substantial lowering of the larynx. Such lowering would cause the rotation of the cricothyroid joint in the direction that would shorten the vocal folds and thus lower F_0, as can be see in Figure 4 [16]. Presumably, because the larynx is pushed against the rigid structure of the trachea in laryngeal lowering, the movement is slower than the laryngeal movement to simply lower F_0, leading to F_0 peak delay.

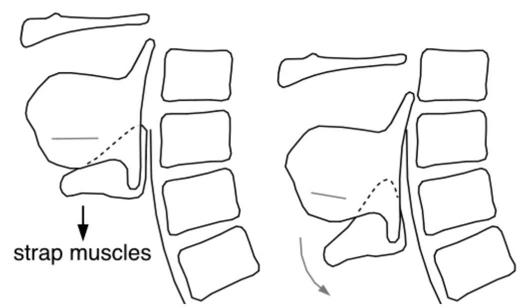


Figure 4: *Vertical components of the extralaryngeal* F_0 *control mechanism.Adapted from Figure 3 of [14] (courtesy of Kiyoshi Honda).*

4. DISCUSSION

The experimental results in section 2 show that in both accented and unaccented syllables there is a greater pitch falling tendency in anger speech than in joy speech. These trends are interpretable in terms of the theory of biological code proposed by Ohala [3]. From this theory, we may speculate that humans make their speech sound assertive and authoritative by lowering F_0 to exaggerate their body size. Conversely, happiness may be conveyed by raising F_0 to understate body size so as to sound non-threatening and sociable. The experimental results in section 3 show that this size code may also be used by manipulating the size of the vocal tract, resulting in the alteration of formant frequencies. For anger, the raising of the first formant indicates that the jaw is opened wider as can be seen in joy. This may be related to the expression of high activated emotion. The lowering of all formants in strategy 2 supports our hypothesis that the low F_0 and delayed F_0 peak is due to the lowering of the larynx to lengthen the vocal tract.

All these results suggest that the size code is more actively used by humans than has been recognized. Since the code is implemented along a scale on which anger and joy probably occupy the two extremes, the related F_0 and formant frequency cues could be used as effective cues for distinguishing the two emotions in speech. It is therefore possible to use F_0 and formant frequency to improve the performance of emotional speech recognition and synthesis.

5. CONCLUSION

In this work, we examined the possible use of a proposed biological code – the size code, for expressing anger and joy in speech. Our experimental results show that anger speech has the tendency to have low F_0 and delayed F_0 peak. Analysis of formant frequencies confirmed our hypothesis that the F_0 peak delay was related to the lowering of the larynx. Additional analysis found lowered formant frequencies in anger speech even when there was no F_0 peak delay. Furthermore, in joy speech we found a tendency to raise F_0 as well as formant frequencies. These results suggest that the size code is still being actively used by humans: exaggerating body size in anger speech to sound authoritative and threatening, and understating body size in joy speech to sound unimposing and sociable. These findings not only are relevant to theory of emotions in speech, but also may have impact on improving the performance of emotional speech recognition and synthesis.

6. REFERENCES

[1] Xu, Y., 2005. Speech Melody as Articulatorily Implemented Communicative Functions. Speech Communication 46: 220-251.
[2] T.L. Nwe, S.W. Foo, and L.C. De Silva, Speech emotion recognition using hidden Markov models. Speech Communication, Vol. 41, Issue 4, pp. 603-623, 2003.
[3] Ohala, J., Ethological Theory and the expression of emotion in the voice. Proceedings of ICSLP 96., 1996.
[4] Hauser, M.D., The evolution of nohuman primate vocalizations: effects of phylogeny, body weight and social context, American Naturalist, 142(3), pp. 149-158, 1993.
[4] Hauser, M.D., The evolution of nohuman primate vocalizations: effects of phylogeny, body weight and social context, American Naturalist, 142(3), pp. 149-158, 1993.
[5] Davies, N.B., Halliday, T.R., Deep croaks and fighting assessment in toads, Nature, 274, pp.683-685, 1978.
[6] Ohala, J. J., Cross-language uses of pitch. *Phonetica* 40: 1-18, 1983.
[7] Fitch W.T., Reby D. The descended larynx is not uniquely human. Proceedings of the Royal Society, Biological Sciences 268:1669-1675, 2001.
[8] Gussenhoven, C., Intonation and interpretation: Phonetics and Phonology. In Proceedings of The 1st International Conference on Speech Prosody, pp. 47-57, 2002.
[9] T. Bänziger and K.R. Scherer, The role of intonation in emotional expressions. Speech Communication, Vol. 46, Issues 3-4, pp.252-267, 2005.
[10] P. Boersma, D.J.M. Weenink, Praat, a System for Doing Phonetics by Computer, Version 3.4 (132). Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, 1996.
[11] R.Banse, K.R.Scherer, Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, pp. 614–636, 1996.
[12] Emotional Speech Group, DSPLAB, University of Maribor web page: http://wwwbox.uni-mb.si/eSpeech/
[13] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B.,and Maneewongvatana, S., The roles of pitch contour in differentiating anger and joy in speech, In: Proc. Internat. Conf. On Computer Science, Prauge, 2006.
[14] Xu, Y., Articulatory constraints and tonal alignment. In Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France. pp. 91-100, 2002.
[15] Fant, G., Acoustic Theory of Speech Production. The Hague: Mouton., 1960.
[16] Honda K, Hirai H, Masaki S, Shimada Y., Role of vertical larynx movement and cervical lordosis in F0 Control. Lang Speech 42(4), pp.401-411, 1999.