PERCEIVING ANGER AND JOY IN SPEECH THROUGH THE SIZE CODE

Yi Xu[1], Suthathip Chuenwattanapranithi[2]

[1]Department of Phonetics and Linguistics, University College London, London NW1 2HE, U.K.
[2]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand.
yi.xu@ucl.ac.uk, chuenwattana@yahoo.com

ABSTRACT

Human speech conveys emotions not only by words, but also by nonverbal acoustic cues. The hypothesis was tested that anger and joy can be conveyed in speech by displaying effort to sound larger or smaller, just as expressing dominance and submission in animal communication. Human listeners perceived vowels synthesized with a statically lengthened vocal tract and lowered pitch as from a large person, but from an angry person when the lengthening and lowering were dynamic. The opposite was true for perceiving small body size and joy. These results point to a "size code" shared by human and nonhuman communications.

Keywords: emotion, anger, joy, size code, target approximation

1.INTRODUCTION

There has been evidence that human listeners can correctly perceive anger and joy even from the speech of languages they do not know [2, 15]. However, the exact acoustic cues that distinguish anger and joy have been difficult to identify, because both emotions involve similarly amplified gross acoustic patterns due to highly activated affective states [2, 10]. Being important for social life, the ability to vocally express emotions probably predates the ability to express propositional meanings in the evolution of human language. This means that the expression of emotions in speech may be linked to nonhuman animal communication. For anger and joy the link could be a biological code known as the "size code" [8], which uses body-size related cues as a means of communication. An aggressor makes itself appear larger to intimidate its adversary by erecting the hair or feathers, elevating the tail or tail feathers [12]. It often accompanies these actions with the emission of low-pitched, rough quality sounds that also indicate a large body size [12]. A submissive animal does the opposite to express non-threat and appeasement. It flattens the ears, the tail, and the hair or feathers [13], and often produces a high-pitched, tone-like sound [11]. The acoustic indication of body size can be done not only by pitch, but also by dispersion of resonance peaks (formants) of the vocal tract, which is inversely related to vocal tract length [4]. The survival and mating advantage of sounding large has led to the drastically descended larynx in animals like red deer [3, 6] The size code is also known to be used by various species of primates including human. They share a grin or smile face to cue amiability, submissiveness, contentment, and non-threat, which is done by shortening the vocal tract to sound smaller [17]. Similarly, the o-face shared by many primates expresses aggression, disapproval and the desire for the viewer to leave by lengthening the vocal tract to sound larger [11]. The lower larynx in male humans as well as male chimpanzees than females is also suggested to be driven by the selection pressure to sound larger for reproductive advantages [4, 13]. The most recent evidence is that the pitch and formant patterns of speech spoken with joy or anger are consistent with the use of the size code [2].

2.PERCEPTUAL EXPERIMENTS

To test whether variations in vocal tract length and pitch are effective in conveying anger and joy, we synthesized the vowels /i/, /e/, /æ/ and /a/ with a 3D articulatory model that can change the vocal tract length by varying the height of the larynx and the protrusion of the lips [1]. For each vowel the shape of the vocal tract was first configured based on general knowledge about speech production [16] and then adjusted till it was appropriate for Thai as judged by the first author, so as to make the stimuli suitable for Thai subjects. The

perceptual tests were carried out in a quiet room. The listeners were 485 undergraduate students (384 males and 101 females) whose ages were between 19-22 years old (mean 20.22 years). The listeners were separated into 5 groups for the 5 experiments. The perceptual tests were conducted in small sessions of 5-10 listeners each.

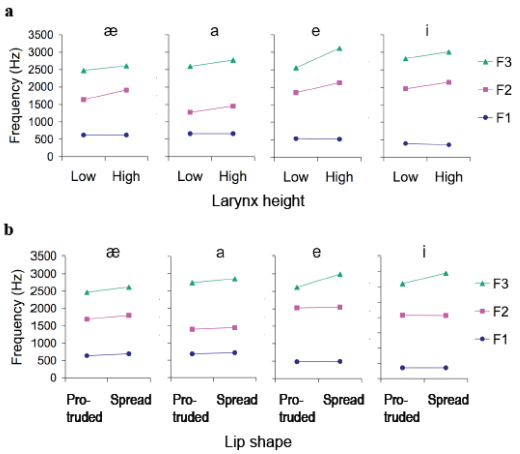
2.1.Experiment 1

In this experiment, the vowels were generated with two larynx positions, high and low. To synthesize vowels with the two static laryngeal positions, the value of the parameter *HY* in the articulatory model, which specifies the vertical position of larynx, was adjusted. For the higher laryngeal position, *HY* = -6.23 cm. For the lower position, *HY* = -6.93 cm. The duration of all the vowels was fixed at 0.4 s., and F_0 fixed at 108 Hz. The distance between the two positions is 6.90 mm to 8.00 mm. The laryngeal height manipulations resulted in clear differences in the spectral patterns in terms of formant dispersion based on the equation proposed in [5]: 928.48-1217.57 Hz for the lower larynx, and 993.25-1328.48 Hz for the higher larynx, as shown in Fig. 1. The mean F_0 and duration for both synthesized sounds are about the same, which are 106.6-108.6 Hz and 0.4 s, respectively. 196 students at Udonthani Rajabhat University and King Mongkut's University of Technology Thonburi, compared the two sounds of each vowel and judged whether they were spoken by a larger or smaller person. Another 197 undergraduate Thai students also at the same institutes judged whether the same vowels were spoken by someone who was happy or angry. The effect of vertical position of larynx on the perception of size and emotion of speaker was assessed by the paired t-test at the 0.05 level of significant with 3 degrees of freedom. In the body size judgment, subjects heard the vowels generated with the lower larynx as produced by a larger person than those generated by the higher larynx (Fig. 2b). The difference between the two laryngeal heights was significant ($p < 0.001$). There was no significant difference in anger perception between the higher and lower positions of the larynx ($p = 0.614$) (Fig. 2a).

2.2.Experiment 2

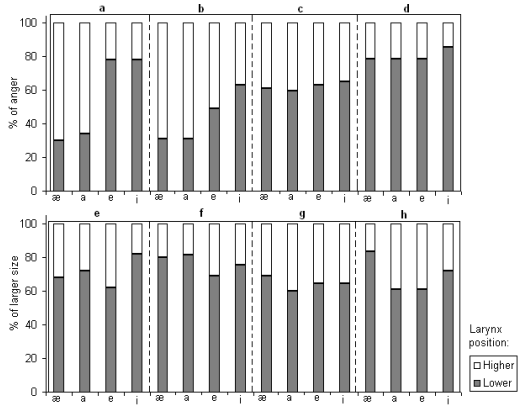
In experiment 2 we tested the role of F_0 on the perception of body size and emotion. The stimuli were similar to those of experiment 1 but with added variations in the overall F_0 of the vowels.

Figure 1: Frequencies of the lowest three formants of the vowels used as stimuli in the perception tests. **a**, Vowels generated with low and high larynx positions. **b**, Vowels generated with protruded and spread lips.



The F_0 was raised by 5 Hz for vowels with higher larynx and lowered by 5 Hz for vowels with lower larynx from the values in experiment 1. Therefore, the F_0 of the vowels generated with the lower larynx was 10 Hz lower than those generated with the higher larynx. This amount of F_0 difference would be audible [9] but very small as compared to more linguistically significant F_0 variations [20]. The subjects were the same groups of students as in the first experiment. The added F_0 difference enhanced the difference in body size judgment between the lower and higher larynx positions (Fig. 2d) ($p < 0.001$). The perception of anger and joy, however, remained ambiguous, as seen in Fig. 2c ($p = 0.284$). The results of the first two experiments thus show that vocal tract length and F_0 both provide perceptual cues for judging the body size of the speaker. Interestingly, such consistency in judgment is despite the fact that the correlation of actual body size is low with either F_0 or vocal tract length [4].

Figure 2: Perceptual results of experiment 1-4



2.3.Experiment 3

Apparently, static vocal tract length and F_0, while useful as cues for body size, are not perceived as expressing anger or joy. Assuming that emotions are deliberately conveyed in speech, speakers should employ the same strategy to encode emotion as they encode linguistic information carried by vowels, consonants and lexical tones. A series of studies have shown that lexical tones are produced in continuous speech as unidirectional F_0 movements toward the underlying tonal targets [18, 19, 22]. More recently, evidence of such unidirectional movements toward vocalic and consonantal targets has also been demonstrated for segmental production [21]. Thus it is possible that emotion perception is also sensitive to dynamic changes of vocal tract length and F_0. In experiment 3, we synthesized the stimuli by changing the larynx height dynamically over the duration of each vowel. For the vowels with dynamic laryngeal movements, two versions of each vowel were synthesized. In the ascending version, larynx height started from *HY* = -6.23 cm at $t = 0$ and was raised over time by 6.90 to 8.00 mm till $t = 0.4$ s (vowel offset). In the descending version, larynx height started from *HY* = -6.93 cm at the initial position, and was lowered by the same amount till $t = 0.4$ s. The rising and falling trajectories were quasi-linear, as stipulated by the 3D synthesizer, with a brief initial acceleration and final deceleration. To preserve the phonetic identity of the vowels, sometimes it was necessary to adjust the values of other parameters of the model such as *TCA* which controls the position and configuration of the tongue. The adjustments were made as small as possible. The F_0 was fixed at 108 Hz. The results (Fig. 2e) show that significantly more subjects heard anger from the vowels with dynamically lowered larynx than from those with dynamically raised larynx; and more subjects heard joy from vowels with raised larynx than from those with lowered larynx ($p < 0.001$). At the same time, similar to the previous two experiments, the lowered larynx was mostly heard as from a larger body size and raised larynx from a smaller body size (Fig. 2f) ($p < 0.001$).

2.4.Experiment 4

In this experiment, we added F_0 movements to accompany the movement of the larynx: falling F_0 with the descending larynx and rising F_0 with the ascending larynx. The F_0 was raised dynamically by 5 Hz from 106 Hz for vowels with higher larynx and lowered dynamically by 5 Hz for vowels with lower larynx. The raising and lowering was quasi-linear from the onset to the offset of the vowel, just as the dynamic changes of the larynx height. The results were clear: More subjects heard anger from vowels with descending larynx and falling F_0, and joy from vowels with ascending larynx and rising F_0 (Fig. 2g) ($p < 0.001$). The perception of body size, while still significantly different in the same direction as in previous experiment, became less robust, with the significant level $p = 0.002$ (Fig. 2h).

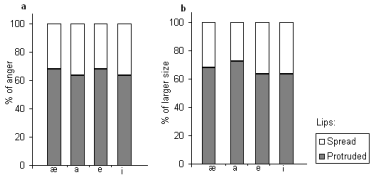
2.5.Experiment 5

Changing vocal tract length can be achieved also by protruding or spreading the lips. To synthesize the vowels with dynamic lip shape changes, the value of *LP* in the articulatory model was varied over time, while the mean F_0 and duration of both synthesized sounds were similar to the values of the vowels synthesized with dynamic laryngeal movements. Two stimulus sets were generated. The first set was synthesized with spreading lips by changing *LP* from 1.56 cm to 0.86 cm between the onset and offset of each vowel. The second set was synthesized with protruding lips by changing *LP* from 0.86 cm to 1.56 cm. Each synthesized vowel, again, had the duration of 0.4 s. The F_0 was varied dynamically throughout each vowel as in experiment 4, raised by 5 Hz for vowels with spreading lips and lowered by 5 Hz for vowels with protruding lips. 44 Thai students were split into two groups, 22 in each. The first group judged whether the two versions of each vowel were spoken by someone who was happy or angry. The second group judged whether the vowels were spoken by a larger or smaller person. The result is shown in Fig. 3. Vowels synthesized with dynamically protruding lips were more frequently heard as angry or spoken by a larger person; those with dynamically spreading lips were more frequently heard as happy ($p < 0.001$) or spoken by a smaller person ($p < 0.001$).

3.DISCUSSION AND CONCLUSION

Taken together, the results show an apparent link between the judgment of emotional state and that of body size. Both judgments are sensitive to the length of the vocal tract and F_0, which point to the size code — a communication strategy used by many non-human species: A larger body size is

Figure 3: Perceptual results of experiment 5.



more threatening and dominant, and a smaller size more appeasing and submissive [3, 5, 12]. But the two kinds of judgments are found to be sensitive to vocal tract length and F_0 in different ways. Static length and F_0 are perceived as directly reflecting body size, while dynamic variations are perceived as expressing joy or anger. It seems as if encoding anger and joy is not to sound convincingly larger or smaller, but to display an effort to do so. But this is exactly the kind of strategy speakers use when trying to encode lexical contrasts conveyed by tonal and segmental phonemes, as demonstrated by recent research [18, 19, 21, 22]. To say a Rising tone in Mandarin in connected speech, for example, speakers produce a unidirectional F_0 movement that continuously approaches an ideal rising slope, regardless of the starting F_0 due to the preceding tone [18, 19]. When such *target approximation* [22] is simulated in speech synthesis, listeners can correctly perceive the intended tones [14]. Similar target approximation strategy has been demonstrated for segmental production [21]. The similarity between the encoding strategy of anger and joy and that of phonemes as found in the present study suggest that these emotions are deliberately expressed rather than unintentionally revealed.

One could go a step further in interpreting the present data. That is, given that the size code is more directly about aggression versus submission [11], what our listeners heard were actually those emotional states rather than anger and joy as mandated by the forced choice paradigm we used. This could be true, but then a further question is whether anger and joy as human emotions have evolved to be sufficiently different from those primitive emotions. The answer could be found only in future research. What the present study has offered is a demonstration of not only the use of the size code in expressing emotions in speech, but also the use of dynamic target approximation as an encoding mechanism for these emotions. These findings thus open the door to future comprehensive investigations of speech emotions in terms of their underlying mechanisms.

4.REFERENCES

[1] Birkholz, P., Jackèl, D.A. 2003. Three-dimensional model of the vocal tract for speech synthesis, In Proc. of Internat. Congress of Phonetic Sciences (ICPhS), Barcelona, Spain, 2597-2600.

[2] Chuenwattanapranithi, S., Xu, Y., Tipakorn, B., Maneewongvatana, S. 2006. Expressing anger and joy with size code," In Proc. 3rd Internat. Conf. on Speech prosody, Dresden, 487-490.

[3] Clutton-Brock, T.H., Albon, S.D. 1979. The roaring of red deer and the evolution of honest advertising, *Behaviour* 69, 145-170.

[4] Fitch, W.T., *Vocal tract length perception and the evolution of language*. 1994. PhD. Dissertation. Brown University.

[5] Fitch, W.T. 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques, *J. Acoust. Soc. Am.* 102, 1213-1222.

[6] Fitch, W.T., Reby, D. 2001. The descended larynx is not uniquely human, In *Proc. of the Royal Society,* Biological Sciences 268, 1669-1675.

[7] Gauthier, B., Shi, R., and Xu, Y. Learning phonetic categories by tracking movements, Cognition (in press).

[8] Gussenhoven, C. 2002. Intonation and interpretation: Phonetics and Phonology, In *Proc. 1st Internat. Conf. on Speech Prosody*, 47-57.

[9] Klatt, D.H. 1973. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception, J. Acoust. Soc. of Am. 53: 8-16.

[10] Nwe, T.L., Foo, S.W., and De Silva, L.C. 2003. Speech emotion recognition using hidden Markov models, *Speech Comm*. 41, 603-623.

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

[12] Ohala, J.J. 1996. Ethological theory and the expression of emotion in the voice, In: *Proc. Internat. Conf. on Spoken Language Processing* (ICSLP 96), Philadelphia, vol. 3, 1812-1815.

[13] Ohala, J.J. 1997. Sound symbolism, In *Proc. Internat. Conf. on Linguistics* SICOL. , 98-103.

[14] Prom-on, S., Xu, Y. and Thipakorn, B. (forthcoming). Modeling tone and intonation in Mandarin and English as a process of target approximation.

[15] Scherer, K.R., Banse, R., and Wallbott, H.G. 2001. Emotion inferences from vocal expression correlate across languages and cultures, J. Crosscult. Psycho.32, 76-92.

[16] Stevens, K.N. 1998. *Acoustic Phonetics* (The MIT Press, Cambridge, MA)

[17] van Hooff, J. A. R. A. M. 1972. A comparative approach to the phylogeny of laughter and smiling, In R. Hinde (Ed.), Nonverbal Communication New York: Cambridge University Press.

[18] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* **25**: 61-83.

[19] Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* **27**: 55-105.

[20] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions, Speech Comm. 46, 220-251.

[21] Xu, Y. and Liu, F. (in press). Determining the temporal interval of segments with the help of F0 contours. To appear in *Journal of Phonetics*.

[22] Xu, Y., Wang, Q. E. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese, Speech Comm. 33, 319-337.