*Phonetica (2008) **65**: 210-230.*

Encoding emotions in speech with the size code — A perceptual investigation

Suthathip Chuenwattanapranithi (Corresponding author)

Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

Email: chuenwattana@yahoo.com

Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, U.K.

Email: yi.xu@ucl.ac.uk

Bundit Thipakorn

Songrit Maneewongvatana

Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

Running Title: Encoding emotions in speech with the size code

Abstract

Our current understanding of how emotions are expressed in speech is still very limited. Part of the difficulty has been the lack of understanding of the underlying mechanisms. Here we report the findings of a somewhat unconventional investigation of emotional speech. Instead of looking for direct acoustic correlates of multiple emotions, we tested a specific theory, the *size code hypothesis of emotional speech*, about two emotions — anger and happiness. According to the hypothesis, anger and happiness are conveyed in speech by exaggerating or understating the body size of the speaker. In two studies consisting of six experiments, we synthesized vowels with a 3D articulatory synthesizer with parameter manipulations derived from the size code hypothesis, and asked Thai listeners to judge the body size and emotion of the speaker. Vowels synthesized with a longer vocal tract and lower F_0 were mostly heard as from a larger person if the length and F_0 differences were stationary, but from an angry person if the vocal tract was dynamically lengthened and F_0 was dynamically lowered. The opposite was true for the perception of small body size and happiness. These results provide preliminary support for the size code hypothesis. They also point to potential benefits of theory-driven investigations in emotion research.

Keywords: Emotional speech, Vocal emotion, Anger, Happiness, Size code, Articulatory synthesis

1. Introduction

Human speech conveys not only linguistic information, but also emotional messages. The emotional content of speech can be heard by listeners even when the word meanings are emotionally neutral. How humans are able to do so has been pondered throughout history, but has been especially actively researched in recent years [see Scherer, 2003 for a comprehensive review]. However, most of the research in this area has been data driven, and there is a general lack of theoretical investigation of the underlying mechanisms of emotional speech [Scherer, 2003]. The most common practice of the field is to examine as many acoustic parameters as possible and measure their correlations with multiple emotions [Murray and Arnott, 1993; Scherer, 2003; Shami and Verhelst, 2007]. The large amounts of data generated this way have not yet led to strong predictive models of emotional speech, however. This is in sharp contrast with the general finding that human listeners are rather accurate in decoding emotional meanings from vocal cues, which is actually somewhat better than facial emotion recognition [Scherer, 1981]. As pointed out by Scherer [1986:144] "[this] discrepancy is striking, particularly in comparison with studies on the recognition of facial expressions of emotion, for which there is abundant evidence of discrete patterning correlated with discriminability [Ekman, 1982]." The situation has not been really improved since then, as pointed out by Scherer [2003]. The present study is an attempt to explore a more theory-driven approach to vocal emotion, as advocated by Scherer [1979, 1986, 2003] and Ohala [1984]. Specifically, we test here what we call the *size code hypothesis of emotional speech*, according to which the emotions along the continuum of anger and happiness are encoded based on the principle of *body-size projection*.

The line of theorization explored here goes back as early as Darwin [1871, 1872] who suggested that human emotions are likely to have evolved from animal emotions, and they are likely to have been shaped by selection pressure for their survival value. Morton [1977] has observed a consistent tendency in the sounds of many birds and mammals, i.e., harsh, relatively low-frequency sounds are made when being hostile, and higher-frequency, more pure-tone-like sounds are made when being submissive, appeasing, or friendly. Morton explains this pattern in terms of what he termed "motivational-structural rules" which is based on a hypothetical selection pressure of evolution involving body size. Because a larger animal is likely to win a physical confrontation over a smaller one, there is a strong selection pressure on animals to adopt the strategy of appearing as large as possible to scare off the opponent. Thus an aggressor would erect its hair or feathers, elevate its tail or tail feathers, or arch its back or hunch its shoulder to appear larger [Davies and Halliday, 1978; Hauser, 1993, 1997; Ohala, 1984]. Some animals have even developed permanent size markers such as hunched back or thick facial hair [Ohala, 1984]. But as argued by Morton [1977], animals not only use visual signals to exaggerate their body size, but also use acoustic cues such as F_0 and quality of voice to help achieve the same effect, since both are related to body size, everything else being equal. The rough quality in animal vocal expression of aggression has been noted as early as Darwin [1872]. Following the same principle, a submissive animal does the opposite to express non-threat and appeasement. It flattens the ears, the tail, and the hair or feathers [Hauser, 1997; Morton, 1977; Ohala, 1984], and often produces a high-pitched, tone-like sound [Hauser, 1997; Morton, 1977]. In addition to indicating non-threat, the high-frequency and pure-tone like sounds also mimic the sounds of infants in order to elicit parental response of care and protection [Morton, 1977].

Morton's [1977] theorization was extended by Ohala [1980, 1984], who proposed that the body-size projection principle applies to humans too. For example, the human smile may actually originate from body-size projection. That is, lip spreading during smile shortens the

vocal tract, which modifies the vocalization in the direction of resembling the spectral patterns generated from a small body. Similar grin facial expressions are shared by many primate species [van Hooff, 1972], which are used to cue amiability, submissiveness, contentment, and non-threat. Likewise, many animals share a facial expression, which Ohala [1984] calls o-face, to expresses aggression and disapproval [Fitch, 1997; Ohala, 1984]. This facial expression is made by protruding the lips, which effectively lengthens the vocal tract, and would generate the impression of a larger body size during vocalization [Ohala, 1984]. Similar to that for the visual size projection, the selection pressure for acoustic size projection is strong enough for many animals to develop permanently or mobilely lengthened vocal tracts [red deer *Cervus elaphus*: Fitch and Reby, 2001, fallow bucks *Dama dama*: McElligott, Birrer and Vannoni, 2006, bird of paradise *Phonygammus*: Clench, 1978, and 60 bird species: Fitch, 1999]. Further evidence that such lengthening of the vocal tract is for the sake of exaggerating body size is the fact that it is often found only in male animals, which suggests its importance for mating competition [Fitch, 1994; Ohala, 1984]. Even in humans [as in Chimpanzees, Fitch, 1994], males have lower larynx than females, and the dimorphism occurs at puberty [Negus, 1949; Goldstein, 1980], i.e., just at a time when males have the need to compete with other males for mates and, later, as *pater familias*, compete with others for territory and other resources for the family [Feinberg et al., 2005].

Of course, body-size projection maneuvers would not work if their perceptual effects are negligible. There has been accumulating evidence that both animals and humans are perceptually sensitive to size-relevant acoustic cues. Charlton, Reby and McComb [2007a] find that female red deer can hear the difference in formant patterns due to changes of vocal tract length, and the same authors find that female deer prefer the roar of larger male deer [Charlton, Reby and McComb, 2007b]. Fitch and Kelley [2000] show that formant pattern differences related to vocal tract length can be heard by whooping cranes *Grus americana*. Fitch [1994] shows that human listeners use vocal tract length to gauge the relative body size of a speaker. Smith et al. [2005] find that human listeners can make fine judgments about the relative size of speakers from vowels re-synthesized with different vocal tract lengths and F_0. Feinberg et al. [2005] have shown that manipulation of formants and F_0 affected the attractiveness of human male voice. Furthermore, Ohala [1984] shows preliminary evidence that, other things being equal, lower F_0 makes a human voice sound more dominant to human listeners. Also, it has been shown that the smile during speech is audible to human listeners [Aubergé and Cathiard, 2003], and that listeners can perceive happiness or unhappiness from speech spoken with a smiling or frowning face [Tartter and Braun, 1994].

Given that the body-size projection principle is applicable to humans, a further question would be whether it can account for some of the human emotions. Emotion is very difficult to define, although it is usually considered to be different from mood, attitude and personal traits in that emotions are more intense and are of shorter duration compared to the other affective states [Scherer, 2003]. Intuitively, if a vocal expression sounds angry, it also feels aggressive and threatening; and if it sounds happy, it also feels friendly and sociable. This means that it is possible that vocal expression of anger is actually an acoustic display of agressiveness and vocal expression of happiness an acoustic display of sociability, and such displays are based on the body-size projection principle. We may call this the *size code hypothesis of emotional speech*, as it assumes that anger and happiness are actively coded expressions rather than unintentially revealed emotional states. This hypothesis is derived from the "motivational-structural rules" proposed by Morton [1977] for animal communication and the "frequency code" by Ohala [1984] for human communication. The term "size code" is first proposed by Gussenhoven [2002] as a possible alternative to "frequency code." We prefer it to frequency code because body size projection is the core of

the proposed mechanism, and frequency is only one of the coding dimensions. The size code hypothesis predicts that angry speech is produced with lengthened vocal tract, lowered fundamental frequency (F_0), and roughened voice, and that happy speech is produced with shortened vocal tract, raised F_0, and tone-like voice. It also predicts that speech sounds generated with these acoustic properties are heard as either angry or happy. There have been some reports of acoustic measurements that are consistent with this hypothesis. In an acoutic analysis of emotional speech, it is found that there is an overall tendency for F_0 to be lower and the falling slope to be steeper in angry than in happy speech, and for formants, especially F2 and F3, to be lower in angry than in happy speech [Chuenwattanapranithia et al., 2006a]. Those tendencies were not strong, however, due possibly to the limitation of the enacted emotions [Scherer, 2003]. As pointed out by Scherer and Bänziger [2004], when it comes to systematic testing of a hypothesis obtained by the more exploratory methods, emotion speech synthesis should be the method of choice. The goal of the present study is therefore to test the size code hypothesis by checking its predictions about the perception of emotions. Specifically, speech sounds made with a longer vocal tract and lower F_0 is heard as angry and spoken by a larger person, while speech sounds made with a shorter vocal tract and higher F_0 is heard as happy and spoken by a smaller person. We will test these predictions in a series of perception experiments using vowels generated by an articulatory speech synthesizer. Due to the limitation of the synthesizer we used, however, we will not test the predictions of the size code hypothesis regarding voice quality.

2. Study 1 — Initial explorations

2.1 Stimuli

To test the predictions of the size code hypothesis, we need to generate speech sounds that are largely free of linguistic contents, but are characteristic of a longer or shorter vocal tract, and with a lower or higher fundamental frequency. To do that, we made use of a 3D articulatory synthesizer developed by Birkholz [Birkholz and Jackèl, 2003], with which vocal tract length could be changed by varying the height of the larynx and shape of the lips, and F_0 could be changed by directly controlling the pitch parameter. The speech sounds we synthesized were four Thai vowels: /æ/, /a/, /e/ and /i/, as the listening tests were conducted in Thailand. For each vowel the shape of the vocal tract was first configured based on general knowledge about speech production [Fant, 1960; Stevens, 1998] and then adjusted till it was appropriate for Thai as judged by the first author, so as to make the stimuli suitable for Thai listeners. The duration of all the vowels was fixed at 0.4 s.

In the Birkholz articulatory model, larynx height was controlled by the parameter *HY*, which specifies the vertical position of the larynx. Lip protrusion was controlled by the parameter *LP* (which is independent of *LH* that controls lip opening), and F_0 was controlled by the parameter F_0. The maximum variation of larynx height in the Birkholz model is 9 mm, which is much smaller than the 40 mm found in singing [Sonninen, 1956], 22 mm in non-singing pitch changes [Shipp, 1975] or 10-17 mm for different vowels [Demolin et al., 2000; Hoole and Kroos, 1998]. To test the effectiveness of changing the larynx height with the model, we conducted an exploratory perceptual test with a small group of subjects (15 Thai listeners). The results showed that most listeners could perceive the difference between two versions of each of the four vowels generated by the articulatory model whose larynx position differed by 7 mm or more, as shown in Figure 1 (Three of the listeners could not hear any difference between the synthetic vowels with different vocal tract lengths. One of them later reported having a hearing problem due to a car accident). We analyzed the effect of changing vocal tract length and vowel on the perceptual detection of the difference using a 2-way

ANOVA. The results showed that both vocal tract length and vowel had a significant effect ($F[9,27] = 49.25$, $p < 0.001$ and $F[3,27] = 4.78$, $p = 0.009$, respectively). By comparing the number of detections between pairs of vowels with different vocal tract lengths, we could see how much change in vocal tract length was perceptible. We performed Fisher's pairwise comparisons on all vowel pairs in terms of number of detections. Vocal tract length changes of 6 mm or more exceeded the significance level of 0.05, but the number of detections between 7-9 mm did not differ significantly.

Based on these results, the amount of larynx height variation in Experiments 1-4 and the amount of lip protrusion variation in Experiment 5 were set to 7 mm. For the higher and lower larynx positions, *HY* was set to -62.3 mm and -69.3 mm, respectively.

For the greater and smaller lip protrusion, *LP* was set to 15.6 mm and 8.6 mm, respectively. The manipulations of larynx height and lip protrusion resulted in clear differences in formant frequencies, as shown in Figure 3. We also calculated the spectral differences in terms of formant dispersion using the equation proposed by Fitch [1997]:

$$D_f = \sum_{i=1}^{N-1} \frac{F_{i+1} - F_i}{N-1} \qquad [1]$$

where D_f is the formant dispersion in Hz, N is the total number of formants measured, and F_i is the frequency of formant i. In addition, we calculated spectral suppression (which is useful in case one wants to directly manipulated spectral properties to simulate the effect of changing vocal tract length) with the following equation:

$$S_f = \sum_{i,j=1}^{N} F_i - F_j \Big/ \sum_{i=1}^{N} F_i \qquad [2]$$

where S_f is the spectral suppression in percentage, N is the total number of formants measured, and F_i and F_j are the frequencies of ith and jth formants of the vowel with higher or lower larynx position, respectively.

Table 1 shows the dispersion and spectral suppression values of the four vowels based on the first three formants. The dispersions of all the four vowels with the lower larynx are smaller than the same vowels with higher larynx, and the amount of spectral suppression is sizeable for all the vowels, indicating that the manipulation of larynx height effectively changed the spectral patterns. The dispersions of the vowels with protruded lips are smaller than the same vowels with spread lips, with the exception of /a/ for which the difference is reversed. This is probably because the calculation of dispersion takes into account only the distance between adjacent formants. The spectral suppression values in Table 1, in contrast, seem to better reflect the formant differences in Figure 2, because their calculation takes into account both inter-formant distances and absolute formant heights. But the spectral suppression values are also much smaller for the lip shape manipulation than for the larynx height manipulation, which may indicate that lip shape manipulation is not as effective as laryngeal height manipulation for changing vocal tract length.

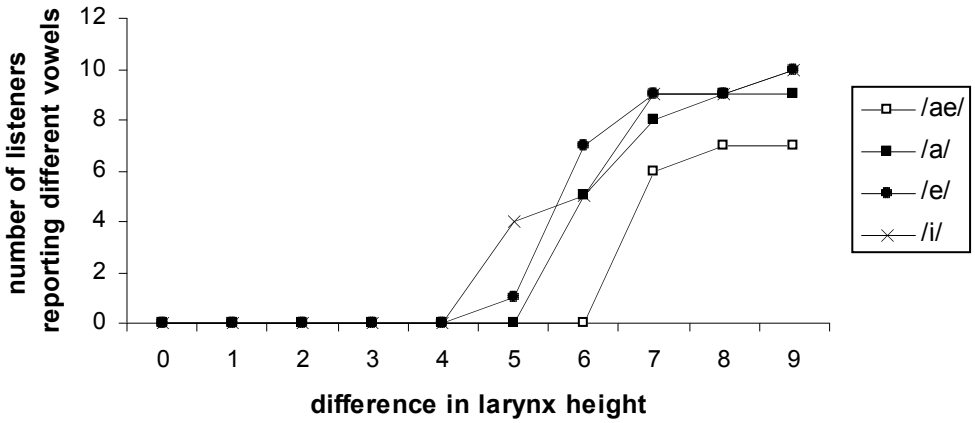

Fig. 1. Audibility of vowel differences due to variation in larynx height. The x-axis shows the difference in larynx height in millimeter when synthesizing 4 Thai vowels with the Birkholz 3D articulatory model. The y-axis shows average number of listeners who heard pairs of vowels as different.

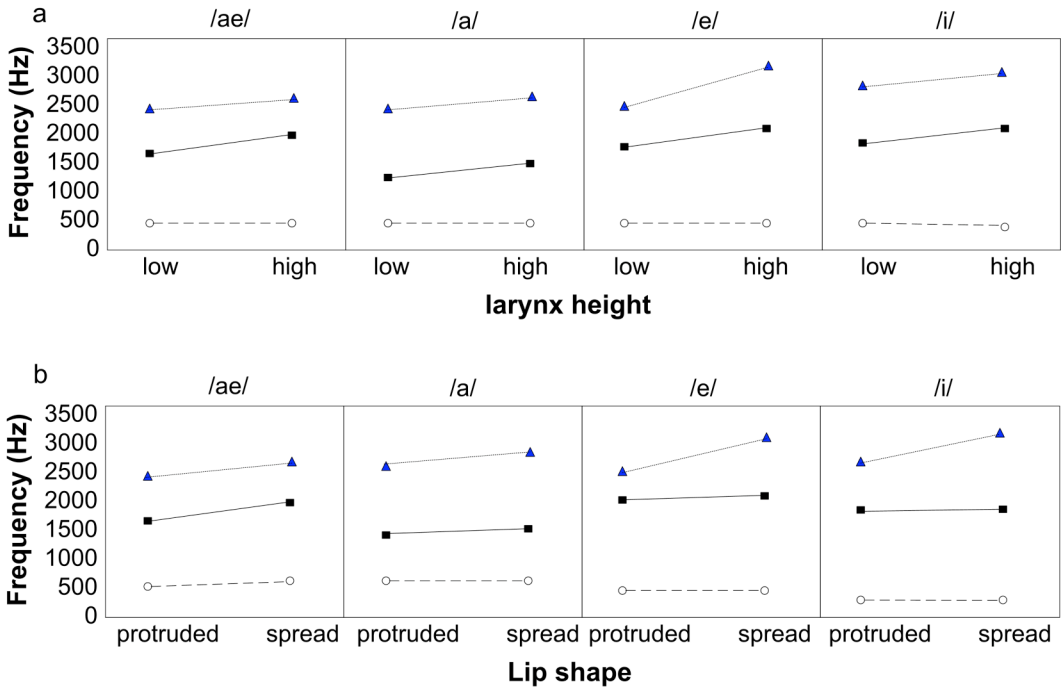

Fig. 2. Frequencies of the first three formants of the vowels used as stimuli in the perception tests. a) Vowels generated with low or high larynx position. b) Vowels generated with protruded or spread lips.

Table 1. Dispersion (in Hz) of the first three formants in vowels synthesized with two larynx heights (up 2 rows) and two lip protrusion values (lower rows).

Vowel	æ	a	e	i
Higher larynx	993.3	1056.9	1301.9	1328.48
Lower larynx	928.5	969.2	1013.3	1217.6

Spectral suppression	7.83%	7.25%	14.26%	6.10%
Spread lips	964.74	1058.04	1163.52	1387.9
Protruded lips	963.0	1060.0	1124.1	1376.0
Spectral suppression	0.26%	0.37%	1.81%	0.28%

The base F_0 of all the vowels was set to 108 Hz, which is appropriate for the average male voice. The amount of F_0 variation in the experiments was 10 Hz, which is above the pitch perception threshold [Klatt, 1973], but smaller than the differences between the tone categories of Thai [Abramson, 1978; Gandour, 1983]. We did not adjust base F_0 for intrinsic F_0., because we expect its perceptual effect to be rather small [Fowler and Brown, 1997].

The parameter manipulation in the experiments was done in two ways: static shift and dynamic variation over time. The motivation for the dynamic variation is that the inherent dynamic nature of speech may affect emotion encoding, as noted by Scherer [2003]. For the dynamic variations, the movement trajectories were quasi-linear, with a brief initial acceleration and final deceleration, which is the built-in transition function of the 3D synthesizer.

2.2 Subjects

The subjects were 485 undergraduate students at Udonthani Rajabhat University and King Mongkut's University of Technology Thonburi, Thailand (384 males and 101 females), aged between 19-22 years (mean = 20.22 years). Of these subjects, 393 participated in Experiments 1-4 and another 92 participated in Experiment 5.

2.3 Experiments and results

The five experiments in this study were performed in a quiet room. The subjects each sat at a computer terminal wearing headphones. They were presented with pairs of vowels (with a non-fixed inter-stimulus interval of 2-3 s) in which one was synthesized with lower larynx and/or lower F_0, and the other with higher larynx and/or higher F_0. Subjects were allowed to ask for repetition before making their decisions (after listen through all stimuli in the first round). In experiments 1-4, 196 of the subjects performed the task of judging the body size of the speaker. They were asked to determine which vowel in each pair was spoken by a larger person. Another set of 197 subjects performed the task of judging the emotion of the speaker. They were asked to determine which vowel in each pair was spoken by an angry person. The four sets of stimuli for experiments 1-4 were mixed together in a randomized order to reduce the possibility of memorizing individual stimuli by rote learning and to minimize the effect of order. For data analysis, the responses of the four experiments were first analyzed with separate paired t-tests to examine the effects of each spectral and F_0 manipulations, and then together in two pooled analyses to examine the gross effects of manner of acoustic manipulations, as will be described subsequently. Subjects were tested in small groups of 5-10 instead of individually so as to reduce the total amount of time needed for testing such a large number of people. A single randomization of the stimuli was used for all the groups. In experiment 5, 44 and 48 subjects made body size and emotion judgments, respectively. This experiment was also run in small groups of 5-10 subjects, all with the same

randomization of the stimuli. The average time to complete the test for one session was 5.53 minutes for Experiments 1-4 and 2.40 minutes for Experiment 5.

The instructions to the subjects were given in Thai. The word / *grōd* / was used to represent the state of anger and irritation, and the word / *dee jaI* / was used to represent the state of happiness, satisfaction, and pleasure. In previous work [Chuenwattanapranithi et al, 2006b], we performed listening tests by asking Thai subjects to specify emotional content in various Western languages including English. The accuracy of Thai subjects in recognizing anger, happiness and sadness from those languages was over 70%. Those results suggest that the meanings of / *grōd* / and / *dee jaI* / to the Thai listeners parallel the meanings of "angry" and "happy" to the listeners of the Western languages.

2.3.1 Experiment 1

In experiment 1, we tested the effect of larynx position on the body size and emotion judgments. The stimuli used in this experiment were the four vowels generated with two larynx positions, high and low (*HY* = -62.3 mm and *HY* = -69.3 mm, respectively). The F_0 of the vowel was fixed at 108 Hz. Figure 3a, b display the listening results in terms of percentage of subjects who made larger size (a) or anger (b) judgments for each of the vowels. Binomial tests showed that all the vowel pairs were heard as different above the chance level of $\alpha = 0.05$ in both body size (N = 196) and emotion (N = 197) judgments. However, as can be seen in Figure 3a, the consistency of judgment across the vowels differed in the two kinds of tasks. The body size judgments were quite consistent, and were in the same direction as found in previous studies, i.e., lower larynx was heard as from a larger person [Feinberg et al., 2005; Fitch, 1994; Smith et al., 2005]. The emotion judgments (proportion of subjects answering anger), on the other hand, were inconsistent. These results were further assessed by paired t-tests at 0.05 level of significance with 3 degrees of freedom. For the body size judgments, the difference between the two larynx positions was significant ($p < 0.001$). For the emotion judgments, there was no significant difference ($p = 0.614$).

2.3.2 Experiment 2

Experiment 2 was to test the role of F_0 in the perception of emotion and body size. The stimuli were similar to those of experiment 1 but with added differences in the overall F_0 of the vowels. F_0 was raised by 5 Hz for vowels with higher larynx and lowered by 5 Hz for vowels with lower larynx. Therefore, the F_0 of the vowels generated with lower larynx was 103 Hz and F_0 of those generated with higher larynx was 113 Hz. The subjects were the same groups of students as in Experiment 1 and the listening procedure was also the same. Their judgments in terms of percent larger body and percent of angry emotion are shown in Figure 3c, d. With the exception of /e/ in the emotion task, all the vowel pairs were heard as different according to Binomial tests at the chance level of $\alpha = 0.05$ in both body size (N = 196) and emotion (N = 197) judgments. As shown in Figure 3c, the added F_0 difference enhanced the difference in body size judgment ($p < 0.001$, paired t-test, df = 3). The perception of anger and happiness, however, remained ambiguous, as seen in Figure 3d ($p = 0.284$, paired t-test, df = 3). The results of the first two experiments thus show that larynx height and F_0 both provide perceptual cues for judging the body size of the speaker, but not for judging anger versus happiness.

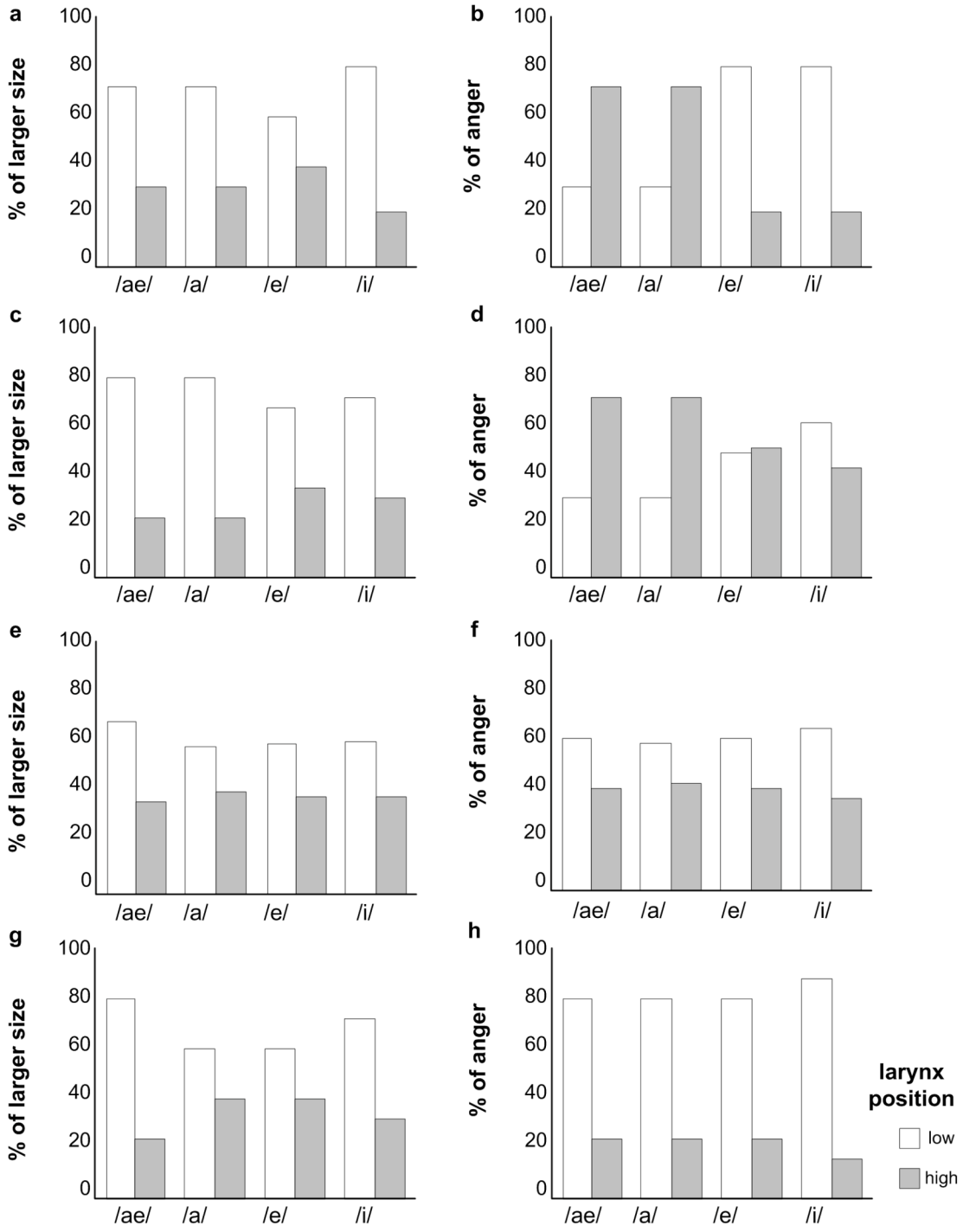

Fig. 3. Perceptual results of experiment 1-4. Left: percentage of larger size judgments; right: percentage of anger judgments. a, b — Experiment 1: Different but static larynx positions, with no F_0 difference. c, d — Experiment 2: Different but static larynx positions, with different but static F_0. e, f — Experiment 3: Different and dynamic larynx positions, with no F_0 difference. g, h — Experiment 4: Different and dynamic larynx positions, with different and dynamic F_0.

2.3.3 Experiment 3

In experiment 3, we tested the effect of dynamic movement of larynx on the body size and emotion judgments. In this experiment, larynx height was dynamically changed over the time course of the vowel. That is, two versions of each vowel were synthesized. In the

descending version, larynx height was dynamically lowered by 6.90-8.00 mm from vowel onset (t = 0 s) to vowel offset (t = 0.4 s), starting from HY = -62.3 mm. In the ascending version, larynx height was dynamically raised by the same amount from vowel onset to vowel offset, starting from HY = -69.3 mm. To preserve the phonetic identity of the vowels, some adjustments were made to the 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, and there was no significant difference in the size or emotion judgments by a group of 15 listeners for the vowels generated with small *TCA* changes (p = 0.8756, paired t-test). F_0 was fixed at 108 Hz for all the stimuli. The subjects and the listening procedure were the same as in Experiments 1 and 2.

Binomial tests (chance level: $\alpha = 0.05$) showed that all the vowel pairs were heard as different in both body size (N = 196) and emotion (N = 197) judgments. Similar to the previous two experiments, vowels with a dynamically lowered larynx were mostly heard as from a larger person (Figure 3e) ($p < 0.001$, paired t-test, df = 3). Unlike in the first two experiments, however, as shown in Figure 3f, more subjects selected vowels with dynamically lowered larynx as produced by an angry speaker ($p < 0.001$, paired t-test, df = 3).

2.3.4 Experiment 4

In this experiment, dynamic F_0 movement was added to accompany the movement of the larynx: F_0 was dynamically raised by 5 Hz from 108 Hz for vowels with ascending larynx and lowered by 5 Hz for vowels with descending larynx. Again, the subjects and the listening procedure were the same as in Experiments 1-3. The perceptual results are shown in Figure 3g and 3h. Binomial tests (chance level: $\alpha = 0.05$) showed that all the vowel pairs were heard as different in both body size (N = 196) and emotion (N = 197) judgments. As can be seen in Figure 3h, more subjects selected vowels with descending larynx and falling F_0 as spoken by an angry person, and more subjects selected vowels with ascending larynx and rising F_0 as produced by a happy person ($p < 0.001$, paired t-test, df = 3). The perception of body size, while still significantly different in the same direction as in previous experiments, became less robust, as seen in Figure 3g, with the significant level p = 0.002 (paired t-test, df = 3).

2.4 Pooled analysis of data from Experiments 1-4

To further examine the separate contributions of formant movements and F_0 to the perception of anger/happiness and body size, we conducted additional analyses by pooling all the data from experiments 1-4 to perform two two-way ANOVAs, with percentage of "correct" judgments (i.e., as predicted by the size code hypothesis) of body size and emotion as dependent variables and F_0 variability (same, different) and manner of larynx height manipulation (static, dynamic) as independent variables. The interaction plots are shown in Figure 4. For body size perception (Figure 4a), the static larynx height differences led to greater number of "correct" body size judgments than dynamic laryngeal movements, and the effect is significant, $F(1,12) = 7.73$, p = 0.017. F_0 variability, however, had no significant effect on body size judgment despite the higher average percentage values when F_0 was different in each pair than when F_0 was the same. There was no interaction between F_0 variability and manner of larynx height manipulation.

For anger/happiness perception (Figure 4b), dynamic laryngeal movements led to greater number of "correct" emotion judgments than static larynx height differences, and the effect was significant, $F(1,12) = 7.79$, p = 0.016. F_0 variability did not lead to any significant difference in the number of "correct" emotion judgments. And the interaction between F_0 variability and manner of larynx height manipulation was not significant.

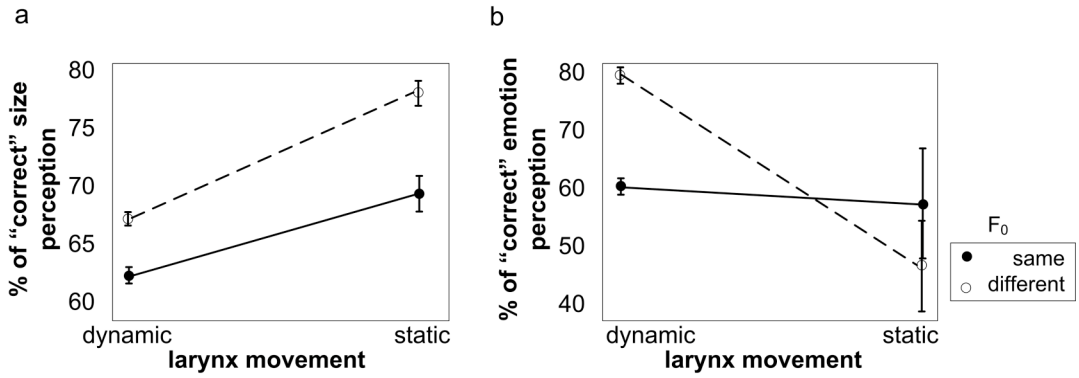

Fig. 4. Interaction of the affects of F_0 variation and manner of larynx height manipulation on the percentage of (a) "correct" body size judgments and (b) "correct" emotion judgments.

2.3.5 Experiment 5

This experiment is to investigate whether lip protrusion and spreading, which also changes the length of the vocal tract, also has an effect on the perception of body size and emotion. Based on the results of Experiments 1-4, only dynamic changes of larynx positions and F_0 were used, as they would generate the strongest effects. To prepare the stimuli, larynx height was dynamically changed by manipulating the parameter *LP* in the articulatory model for lip protrusion. Two stimulus sets were generated. The first was synthesized with protruding lips by dynamically changing *LP* from 8.6 mm to 15.6 mm between the onset and offset of each vowel, and the second set with spreading lips by changing *LP* from 15.6 mm to 8.6 mm. Similar to the larynx height manipulation, the values of other parameters, especially *LH* which controls the opening of the lips, were adjusted, but only within a very small range, to preserve the phonetic identity of the vowels. The effects of these manipulations on the formant frequencies are shown in the lower part of Table 1. F_0 was also dynamically varied throughout each vowel as in experiment 4, raised by 5 Hz in vowels with spreading lips and lowered by 5 Hz in vowels with protruding lips. 92 Thai students (70 males and 22 females), were split into two groups, 48 and 44 in each. The first group judged whether the two versions of each vowel were spoken by a larger or smaller person. The second group judged whether the vowels were spoken by an angry or happy person. The results are shown in Figure 5. Vowels synthesized with dynamically protruding lips and falling F_0 were more frequently heard as spoken by an angry person, and those with dynamically spreading lips and rising F_0 more frequently heard as spoken by a happy person ($p < 0.001$, paired t-test, df = 3) (Figure 5b). At the same time, vowels with protruding lips and falling F_0 were more frequently heard as from a larger person, and those with spreading lips and rising F_0 more frequently as from a smaller person ($p < 0.001$, paired t-test, df = 3) (Figure 5a).

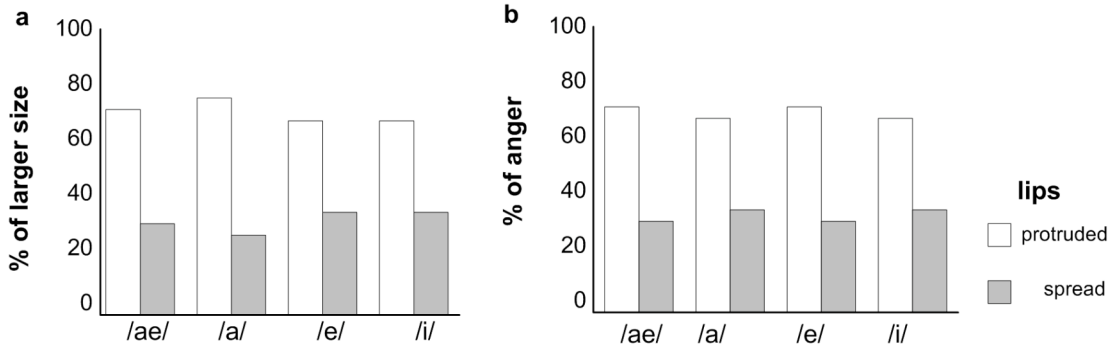

Fig. 5. Perceptual results of experiment 5 in terms of percentage of bigger size (a), and anger (b) judgment.

2.5 Discussion

In study 1, we conducted 5 experiments to test step by step the size code hypothesis for encoding anger and happiness. The first two experiments established that listeners were sensitive to the manipulation of larynx height and overall F_0 height, and perceived vowels with lower larynx and lower F_0 as uttered by a person with a larger body size than those with higher larynx and higher F_0. These two experiments also showed that, even when given a two-way forced choice, listeners could not consistently hear the two versions of synthetic vowels as conveying either anger or happiness. Experiments 3 and 4 showed that when larynx height and F_0 are dynamically changed over the course of a vowel, listeners could hear the vowels as conveying either anger or happiness, and that both larynx height and F_0 contribute to the perception of the two emotions, although the effect of the former is more robust. The pooled analyses for Experiments 1-4 further show that a) dynamic changes of the larynx height have a more robust effect than static larynx variations on the judgment of anger and happiness, and b) laryngeal variability is more effective than F_0 variability for both body size and emotion judgments. Experiment 5 further demonstrates that dynamically changing the vocal tract length by protruding the lips, when accompanied by dynamic F_0 changes, had the same effect on emotion perception as changing the larynx height. This is despite the fact that lip shape manipulation had a much smaller effect than larynx height manipulation on the formant frequencies as shown in Table 1.

3. Study 2 — An integrated test

While the results of Study 1 are consistent with the size code hypothesis, the unusual nature of the hypothesis calls for an even more stringent test. In Study 2 we conducted an integrated experiment with all the factors manipulated at the same time. Also, to allow subjects more flexibility in their judgment, we used a 3-level rating scale instead of the two-way forced choice used in Study 1.

3.1 Stimuli

The stimuli were similar to those used in Study 1 in most respects. Four Thai vowels, /æ/, /a/, /e/ and /i/, were generated using the Birkholz 3D articulatory synthesizer, with the following parameter manipulations.

1) Larynx position — high: *HY* = -62.3 mm; low: *HY* = -69.3 mm;

2) F_0 — high: 111 Hz, low: 101 Hz;

3) Dynamic patterning — static: high or low larynx height and high or low F_0 throughout a vowel; dynamic: larynx height and F_0 start either from the high or

low level at vowel onset and end at the opposite level at vowel offset;

4) Vowel — /æ/, /a/, /e/, /i/

The duration of all the vowels was fixed at 0.4 s.

3.2 Subjects and procedure

The listeners, who were a different group from that of Study 1, were 68 undergraduate students from the same universities as in Study 1 (33 males and 35 females). Their age varied between 18-22 years (mean = 21.2 years). With the same organization of Study 1, subjects were divided into two sets (34 in each). Those in the first set were asked to rate the speaker's body size and those in the second set were asked to rate the speaker's emotion. The perceptual tests were carried out in a quiet room, and in small groups of 5-10 subjects. They heard all the synthetic vowels mixed in random order through the headphones and rated the speaker's body size and emotion using a three-level scale: 1 = happy or small, 2 = neutral, and 3 = angry or larger. The average time to complete the test for one session was 6.34 minutes.

3.3 Results

Figures 6 and 7 display the mean scores of body size judgment. Figure 6 shows that vowels with lower F_0 and lower larynx were perceived as produced by a larger person and those with higher F_0 and higher larynx were perceived as produced by a smaller person. Both effects are significant. For F_0, $F(1, 3) = 11.52$, $p = 0.005$; for larynx height, $F(1,13) = 7.06$, $p = 0.020$. The effect of dynamic patterning is not significant, $F(1, 13) = 3.94$, $p = 0.069$. There is also no significant effect of vowel.

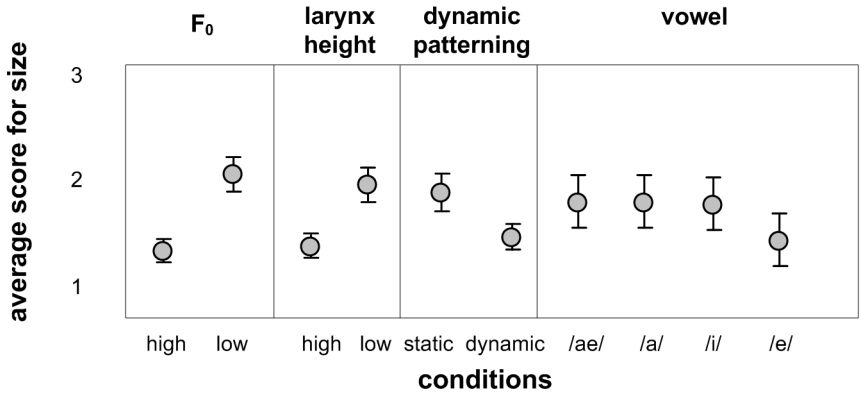

Fig. 6. Effects of F_0 level, larynx height, dynamic patterning, and vowel on the perceptual rating of body size. A larger number represents a larger size.

Figure 7 shows plots of two-way interactions among the four factors of synthetic vowels on the perception of body size. Here the most interesting trend is seen in the third column, where we can see that F_0 and larynx height provide better perceptual cues to body size when the larynx is static than when it is dynamic. The interaction between larynx height and dynamic patterning is significant, $F(1, 13) = 7.06$, $p = 0.020$; but the interaction between F_0 and dynamic patterning is only marginal, $F(1, 13) = 4.20$, $p = 0.061$. None of the other interactions is significant.

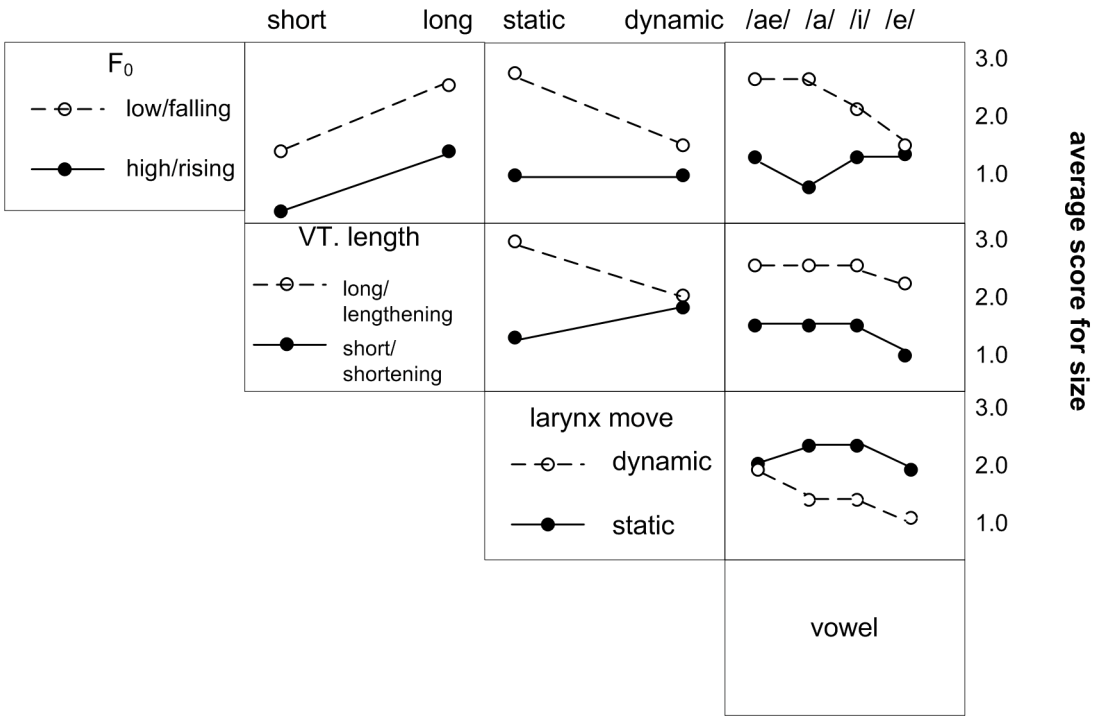

Fig. 7. Plots of two-way interactions across the four factors of synthetic vowels for the perception of body size. Row 1: Interaction of F_0 with larynx height, dynamic patterning and vowel. Row 2: Interaction of larynx height with dynamic patterning and vowel. Row 3: Interaction of dynamic patterning with vowel.

Figures 8 and 9 show the mean scores of anger/happiness judgment. Figure 8 shows that lower F_0 and lower larynx were perceived as angry, and higher F_0 and higher larynx were perceived as happy. Both effects are significant. For F_0, $F(1, 13) = 4.71$, $p = 0.05$. For larynx height, $F(1,13) = 11.23$, $p = 0.005$. The effect of dynamic patterning is also significant, $F(1, 13) = 6.99$, $p = 0.020$. There is no significant effect of vowel.

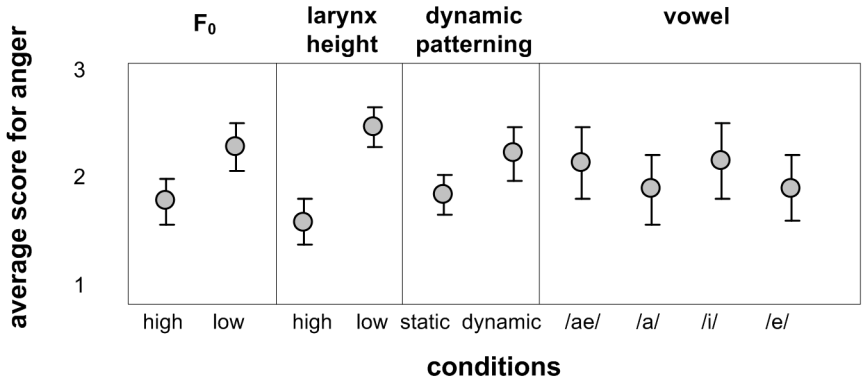

Fig. 8. Effects of F_0, larynx height, dynamic patterning, and vowel on anger perception. A larger number represents an angrier judgment.

Figure 9 shows plots of two-way interactions among the four factors of synthetic vowels on the perception of anger versus happiness. Here we can see that it is the lower larynx with lower F_0 that sounded the most angry. Also we can see the trend that F_0 and larynx height provided better perceptual cues to anger vs. happiness when the larynx was dynamically changed than when it was static. The interaction between larynx height and dynamic patterning is significant, $F(1, 13) = 5.00$, $p = 0.044$. But the interaction between F_0 and dynamic patterning is not significant. None of the other interactions is significant either.

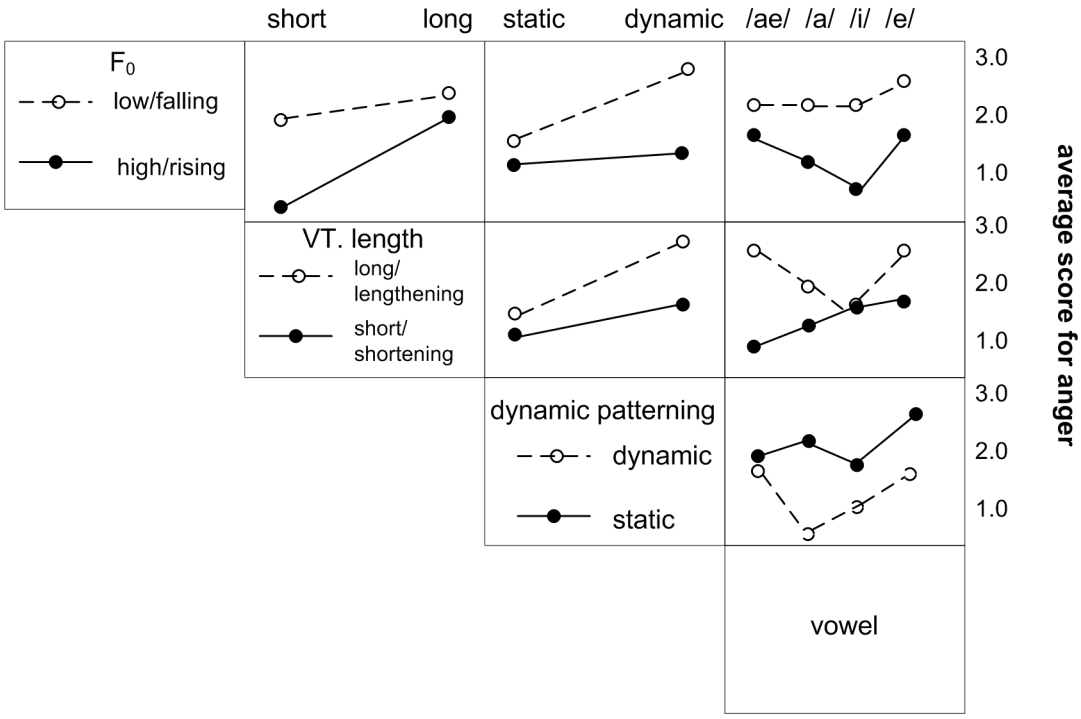

Fig. 9. Plots of two-way interactions across the four factors of synthetic vowels for the perception of anger/happiness. Row 1: Interaction of F_0 height with larynx height, dynamic patterning and vowel. Row 2: Interaction of larynx height with dynamic patterning and vowel. Row 3: Interaction of dynamic patterning with vowel.

4. General Discussion

The purpose of this research is to test the predictions of the size code hypothesis of emotional speech, namely, anger and happiness are encoded based on the principle of body-size projection, which has been suggested to underlie animal expressions of threat and appeasement [Morton, 1977, 1982]. Specifically, the hypothesis predicts that human listeners will hear speech sounds produced with lengthened vocal tract, lowered F_0, and roughened voice quality as spoken by an angry person, and those produced with shortened vocal tract, raised F_0, and tone-like voice quality as spoken by a happy person. We tested the first two of these predictions in two studies. In Study 1 we conducted 5 perceptual experiments using vowels synthesized with manipulated vocal tract length and F_0 as stimuli. Results showed that vowels synthesized with statically longer vocal tract and lower F_0 were heard as spoken by a larger person than those synthesized with statically shorter vocal tract and higher F_0. But vowels synthesized with dynamically lengthened vocal tract and lowered F_0 were heard as spoken by an angry person, while those with dynamically shortened vocal tract and raised F_0 were heard as spoken by a happy person. In Study 2 we conducted an experiment in which

vocal tract length, F_0 and manner of larynx/F_0 variation were simultaneously controlled. Results showed that all three manipulations significantly affected both size and emotion perception as predicted by the size code hypothesis. Due to the limitation of the synthesizer, however, we did not test the effect of voice quality on emotion perception as also predicted by the size code hypothesis.

Thus we have shown preliminary evidence that the size code is involved in the perception of anger and happiness in human speech. Our finding is consistent with the previous finding that the smile during speech is audible to human listeners [Aubergé and Cathiard, 2003], and that speech produced with a smiling or frowning face is perceived as happiness or unhappiness [Tartter and Braun, 1994]. We are the first to show, however, that listeners can hear anger from speech sounds produced with a lowered larynx.

The significance of the findings needs to be put in the general context of emotional speech research. Contrary to what one might expect based on intuition, anger and happiness have been quite difficult to characterize based on the most obvious acoustic parameters. They both involve highly activated affective states which generate similarly amplified acoustic patterns [Chuenwattanapranithi et al., 2006a; Nwe et al., 2003; Scherer, 2003]. As summarized by Nwe et al. [2003], angry and happy speech both has ascending pitch contours, increased average pitch, wide pitch range, raised intensity, increased speech rate and tense voice quality. And, according to Murray and Arnott [1993], both angry and happy speech have faster speech rate, higher average pitch, wider pitch range and higher intensity than neutral speech. These common properties make anger and happiness easily separated from low-activation emotions such as sadness and boredom, but not from each other. Kwon et al. [2003] used acoustic parameters such as F_0, formant, energy and mel-frequency cepstral coefficients to recognize 5 emotions, including anger, happiness, sadness, boredom and neutral emotion. The results showed that angry samples were frequently categorized as happy, and happy samples frequently categorized as angry. These confusions reduced the overall recognition rate to only 42.3%. In a recent study by Shami and Verhelst [2007] which presumably reflects the state of the art in automatic emotion recognition, 20% of the angry speech was classified as happy, and 35.6% of the happy speech classified as angry. The severity of the situation is further highlighted by the fact that anger and happiness are actually the two most frequently encountered emotions after neutral emotion. As found by Morrison et al. [2007], anger and happiness are by far the largest emotion categories (3.1% and 1.8%, respectively) communicated through a call center system following neutral speech (93.3%), while the other emotional classes such as sadness, fear, surprise, and disgust have much lower percentages of occurrence (the highest being sadness, at only 0.1%). Therefore, the current findings may have interesting implications not only for theoretical understanding of emotions but also for practical applications that process emotions.

Although the results of the present study are quite robust, they actually raise more questions. But answering these questions may help us better understand emotions in speech. First, it could be argued that the current results are artifacts of the two-way or three-way forced choice tasks, and the outcome could have been quite different had listeners been given more emotions to choose from, or even allowed to make open choices. We note, however, that the inclusion of many emotions in a recognition test may actually hide the overall difficulty in processing emotions. For example, when given 5-7 alternative emotions to choose from, the chance level is only 20%-14%, which is why the typical emotion recognition rate is as high as five times above chance [Mozziconacci, 2002; Scherer, 2003]. Had the more difficult emotions been pitted against each other, the performance would have been much worse. For example, anger and happiness are among the most difficult to distinguish from each other, as just mentioned, but their discrimination has rarely been tested

without other emotions. The usual practice of multiple choices in emotion research actually tends to exaggerate the recognition rate relative to chance. With two choices the chance level is raised to 50% and three choices to 33%. As a result, any token that could have been perceived as a third or fourth choice would now have to be randomly assigned by subjects to either of the two forced choices in Study 1 and to the neutral choice in Study 2. If a cue is consistently used by subjects for a particular emotion, it can only mean that it is relevant to it in some way. A case in point is the finding of Experiments 1 and 2 that static larynx height and F_0 differences cannot be heard as carrying emotional information despite their consistency in signaling body size. Having said that, we recognize that the present findings are only about how anger and happiness could be distinguished from each other. They do not tell us about how they are distinguished from other emotions, which has to be examined in experiments designed for that purpose. Most importantly, the size code is just one of the possible mechanisms relevant for some of the emotions. Many other emotions may well be based on other mechanisms that are unrelated to anger and happiness.

Second, the current findings do not necessarily suggest that there must exist a strong correlation of body size with *F_0* or spectral density in the real world. In fact, past research has repeatedly found no correlation between F_0 and body size in humans when sex and age are controlled [Lass and Brown, 1978; Künzel, 1989; Van Dommelen, 1993; Hollien, Green and Massey, 1994; González, 2004], or even in red deer [Reby and McComb, 2003] and some amphibian species [Asquith and Altig, 1990]. However, Evans, Neave and Wakelin [2006] did find significant negative correlations between F_0 of male humans and measures of their body shape including shoulder and chest circumferences, and shoulder–hip ratio. What is interesting is that these measures are the ones that are determined at puberty by the action of testosterone, which is also related to changes in F_0 and vocal tract length of males at puberty [Dabbs and Mallinger, 1999 and Pfefferle and Fischer, 2006 for F_0; Fitch and Giedd, 1999 for vocal tract length]. Thus one interpretation of the lack of correlation between body height and F_0 could be that the selection pressure for using voice characteristics for generating perceptual impressions of body size has actually weakened the actual correlation. What is more interesting is that despite the reported lack of correlation of body size and *F_0*, our subjects still used F_0 as a cue for body size judgment, which parallels the finding of Collins [2000] that females consistently, but wrongly, judged males with lower F_0 and smaller formant dispersion as being heavier, older, more likely to have a hairy chest and a muscular body type. It is conceivable that for any perception-based selection pressure to work, the actuating perceptual preferences need to be consistent and unambiguous, whereas the individual adaptations to those preferences would naturally result in anatomical and behavioral changes in the direction of weakening the source correlations that had been strong at the start of the process.

Third, the finding that anger and happiness perception was more robust when vocal tract length and *F_0* were dynamically changed than when they were only statically different was somewhat unexpected. Equally unexpected was that body size judgment was more consistent when vocal tract length was static than when it was dynamic. It seems as if listeners can "tell" when the acoustic cues indicate "true" body size and when they are used as a code. From the perspective of encoding, it seems that conveying anger and happiness is not to sound convincingly larger or smaller, but to show an effort to do so. But this is exactly the kind of strategy speakers use when encoding lexical contrasts conveyed by tonal and segmental sounds [Xu, 1997, 1999; Xu and Liu, 2007], as depicted by the Target Approximation (TA) model [Xu and Wang, 2001]. To produce a phonetic unit in connected speech, the speaker produces a unidirectional articulatory movement that asymptotically approaches an ideal target, starting from the initial states of the articulators [for tone: Xu,

1997, 1999; for segments: Xu and Liu, 2007]. The similarity between the encoding strategy of phonetic units and that of anger and happiness suggest that the expression of these emotions are not mere passive reflections of the inner psycho-physiological state of the speaker, but "designed to induce in receivers behavior that benefits the signaler" [Ohala, 1996:1812].

Fourth, the size code hypothesis also has predictions about the relation of voice quality to anger and happiness. That is, angry voice would be rougher while happy voice would be more tone-like. This prediction was not directly tested in the current experiments due to the limitation of the synthesizer we used. However, Gobl and Chasaide [2003] have shown, also with synthetic speech, that voice quality does provide strong cues for some emotions, especially anger. Thus it would be interesting to further examine the effectiveness of voice quality as an emotion cue when used in conjunction with vocal tract length and F_0.

Fifth, questions may be raised as to whether the stimuli used here are true prototypes of angry and happy speech, or even whether they truly resemble angry and happy emotions in speech. Admittedly, with the vocal tract length difference of only 7 mm, and F_0 difference of only 10 Hz, these stimuli are probably far from representing the true emotional extremes. Instead, they are more likely to be points not too far from the center along a continuum between hot anger and exuberant elation. On the other hand, their perceptual effectiveness probably does reflect listeners' high sensitivity to the size code. As for the true emotional relevance of the acoustic manipulations in the present study, a deeper question is actually the exact nature of emotions in speech. Are they merely reflections of the internal physiological state of the individual, or are they designed to have an effect on the receiver of the signal? In our view, while the current results lend strong support to the latter at least for anger and happiness, they do not reject the former. This is because the principle of body-size projection is about the design of the emotional code rather than about how it is neurally controlled. It is highly likely that only the right internal physiological/hormonal state can lead to *simultaneous* control of all the dimensions of a particular emotional code, e.g., vocal tract length (which is determined by both larynx position and lip shape), F_0 and voice quality in the case of anger versus happiness. This is probably why a good actor would make himself genuinely emotional before acting out an emotive scene, as doing so probably would generate the right hormonal state necessary for controlling all the encoding dimensions of an emotional code.

Finally, the current findings do not necessarily suggest that the size code is used exclusively for expressing anger and happiness or vise versa. In Morton's [1977] theory, the body-size projection principle underlies not only friendliness, but also fear, in animal calls. It may be a theoretical stretch to suggest that happiness and fear are equivalent, but the mechanism of the size code implies that there is a link between the two emotions. Research on primate facial expressions has suggested that the human smile is homologous to the bared-teeth display, or grimace, of many mammalian species, from canids to primates [Parr and Waller, 2006]. Among primates, this expression can mean either subordination or appeasement, depending on the social structure of the species [van Hooff, 1967]. It has been suggested that the human smile is functionally more similar to the appeasement meaning of the bared-teeth display [Parr and Waller, 2006; Preuschoft and van Hooff, 1997; Waller and Dunbar, 2005]. On the other hand, the human happiness may not be a single emotion. Intuitively, a hearty laughter elicited by a comedy show feels very different from a charming social smile. Research on primate facial expression has suggested that the human laughter is homologous to the primate 'play face' (with relaxed lips and covered teeth) which expresses playful, affiliative, friendly emotions rather than to the bared-teeth display [Preuschoft and van Hooff, 1967; van Hooff, 1967, 1972; Waller and Dunbar, 2005]. It is further suggested

that laughter and smile have become increasingly similar in humans because both are related to social cohesion and therefore have similar ultimate (evolutionary) functions [Parr and Waller, 2006]. But the source mechanism of the play face, which is likely different from the size code, remain unclear, although there is also a possible link to vocalization [Waller and Dunbar, 2005]. Thus for laughter, as well as for many other emotions, there may exist codes based on biological principles other than body-size projection, some of which have been suggested by Gussenhoven [2002]. The present research has demonstrated that it is worthwhile to search for such codes, and that testing them in controlled experiments may be more fruitful than directly tallying acoustic characteristics of conventional emotion categories.

5. Conclusion

In this work, we have found that human listeners heard synthetic speech generated with statically lengthened or shortened vocal tract as indication of larger or smaller body size, but heard synthetic speech with dynamically lengthened or shortened vocal tract as expression of anger or happiness when given a two-way or three-way choice. The perception of anger and happiness was further enhanced by concurrently lowered or raised F_0. We believe that these findings constitute preliminary evidence for the size code hypothesis of emotional speech, i.e., anger and happiness are conveyed by exaggerating or understating the body size of the speaker, just as nonhuman animals are argued to exaggerate or understate their body size to communicate threat or appeasement [Morton, 1977, 1982]. In particular, the two major perceptual cues for body size, namely, vocal tract length and voice pitch, as reported by previous studies [Fitch, 1994; Smith et al., 2005], both seem to contribute to the perception of anger and happiness, so long as their variations are dynamic. The contribution of the third cue — voice quality, as proposed by Morton [1977], however, needs to be assessed in future research. The dynamic encoding strategy found here is consistent with models of speech in which communicative information is encoded by articulatory movements toward meaning-associated phonetic targets [Xu, 2005]. Although the current results are quite preliminary and their full validity awaits replication with different methodologies and more realistic speech materials, the work reported here opens the door to future investigations of speech emotions in terms of their underlying mechanisms.

Acknowledgment

We would like to thank Peter Birkholz for making the synthesizer publicly available, and two anonymous reviewers for their valuable comments. Part of the results of Study 1 was reported at ICPhS 2007. The research is supported in part by NIH grant No. 1R01DC006243 to the second author.

References

Abramson, A. S.: Static and dynamic acoustic cues in distinctive tones. Lang. Speech *21*: 319-325 (1978).

Asquith, A.; Altig, R.: Male call frequency as a criterion for female choice in Hyla cinerea. Journal of Herpetology *24*: 198-201 (1990).

Aubergé, V.; Cathiard, M.: Can we hear the prosody of smile. Speech Communication *40*: 87-97 (2003).

Birkholz, P.; Jackèl, D.: A three-dimensional model of the vocal tract for speech synthesis. Proc. The 15th International Congress of Phonetic Sciences, pp. 2597-2600 (Barcelona, Spain 2003).

Charlton, B.; Reby, D.; McComb, K.: Female perception of size-related formant shifts in a nonhuman mammal. Animal Behaviour *74*: 707-714 (2007a).

Charlton, B.; Reby, D.; McComb, K.: Female red deer prefer the roars of larger males. Biology Letters *3*: 382-385 (2007b).

Chuenwattanapranithi, S.; Xu, Y.; Thipakorn, B.; Maneewongvatana, S.: Expressing anger and joy with the size code. Proc. Speech Prosody 2006, pp. OS4-1_0090 (Dresden, Germany 2006a).

Chuenwattanapranithi, S.; Xu, Y.; Thipakorn, B.; Maneewongvatana, S.: The roles of pitch contour in differentiating anger and joy in speech. International journal of signal processing *3*: 129-134 (2006b).

Clench, M. H.: Tracheal elongation in birds-of-paradise. Condor *80*: 423-430 (1978).

Collins, S. A.: Men's voices and women's choices. Animal Behaviour *60*: 773–780 (2000).

Dabbs, J. M.; Mallinger, A.: High testosterone levels predict low voice pitch among men. Personality and Individual Differences *27*: 801–804 (1999).

Darwin, C.: The Descent of Man and Selection in Relation to Sex (John Murray, London 1871).

Darwin, C.: The Expression of the Emotions in Man and Animals (John Murray, London, England 1872).

Davies, N. B.; Halliday, T. R.: Deep croaks and fighting assessment in toads. Nature *274*: 683- 685 (1978).

Demolin, D.; Metens, T.; Soquet, A.: Real time MRI and articulatory coordinations in vowels. Proc. 5th Speech Production Seminar, pp. 86-93 (München, Germany 2000).

Ekman, P.: Methods of measuring facial action; in K. R. Scherer and P. Ekman, Handbook of methods in nonverbal behavior research, pp. 45-90 (Cambridge University Press, Cambridge, England 1982).

Evans, S.; Neave, N.; Wakelin, D.: Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. Biological Psychology *72*: 160–163 (2006).

Fant, G.: Acoustic Theory of Speech Production (Mouton, The Hague 1960).

Feinberg, D. R.; Jones, B. C.; Little, A. C.; Burt, D. M.; Perrett, D. I.: Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. Animal Behavior *69*: 561-568 (2005).

Fitch, W. T.: *Vocal tract length perception and the evolution of language*. Ph. D. Dissertation (Brown University, 1994).

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

Fitch, W. T.: Acoustic exaggeration of size in birds by tracheal elongation: Comparative and theoretical analyses. Journal of Zoology (London) *248*: 31-49 (1999).

Fitch, W. T., and Giedd, J.: Morphology and development of the human vocal tract: a study using magnetic resonance imaging, Journal of the Acoustical Society of America *106*: 1511-1522 (1999).

Fitch, W. T.; Kelley, J. P.: Perception of vocal tract resonances by whooping cranes Grus

americana. Ethology *106*: 448-463 (2000).

Fitch, W. T.; Reby, D.: The Descended Larynx Is Not Uniquely Human. Proceedings of the Royal Society, Biological Sciences *268*: 1669-1675 (2001).

Fowler, C. A.; Brown, J. M.: Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. Perception & Psychophysics *59*: 729-738 (1997).

Gandour, J.: Tone perception in Far Eastern languages. Journal of Phonetics *11*: 149-175 (1983).

Gobl, C.; Chasaide, A. N.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication *40*: 189-212 (2003).

Goldstein, U.: *An articulatory model for the vocal tracts of growing children* Ph.D. dissertation (Massachusetts Institute of Technology, 1980).

González, J.: Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of Phonetics *32*: 277–287 (2004).

Gussenhoven, C.: Intonation and interpretation: Phonetics and Phonology. Proc. The 1st International Conference on Speech Prosody, pp. 47-57 (Aix-en-Provence, France 2002).

Hauser, M. D.: The evolution of nonhuman primate vocalizations: effects of phylogeny, body weight and social context. American Naturalist *142*: 528-542 (1993).

Hauser, M. D.: Artifactual kinds and functional design features: what a primate understands without language. Cognition *64*: 285–308 (1997).

Hollien, H.; Green, R.; Massey, K.: Longitudinal research on adolescent voice change in males. Journal of the Acoustical Society of America *96*: 2646-2653 (1994).

Hoole, P.; Kroos, C.: Control of larynx height in vowel production. Proc. The 5th International Conference on Spoken Language Processing, pp. 531-534 (Sydney, Australia 1998).

Klatt, D. H.: Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. Journal of the Acoustical Society of America *53*: 8-16 (1973).

Künzel, H. J.: How well does average fundamental frequency correlate with speaker height and weight? Phonetica *46*: 117-125 (1989).

Kwon, O. W.; Chan, K.; Hao, J.; Lee, T. W., . Emotion Recognition by Speech Signals. Proc. Eurospeech, pp. 125-128 (Geneva, Switzerland 2003).

Lass, N. J.; Brown, W. S.: Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. journal of the Acoustical Society of America *63*: 1218-1220 (1978).

McElligott, A. G.; Birrer, M.; E., V.: Retraction of the mobile descended larynx during groaning enables fallow bucks (Dama dama) to lower their formant frequencies. Journal of zoology *270*: 340-345 (2006).

Morrison, D.; Wang, R.; De Silva, L. C.: Ensemble methods for spoken emotion recognition in call-centres. Speech Communication *49*: 98-112 (2007).

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

Morton, E. S.: Grading, discreteness, redundancy, and motivational-structural rules; in D. Kroodsma and E. H. Miller, Acoustic Communication in Birds, pp. 183-212 (Academic Press, New York 1982).

Mozziconacci, S.: Prosody and Emotions. Proc. The 1st International Conference on Speech Prosody, pp. 1-9 (Aix-en-Provence, France 2002).

Murray, I. R.; Arnott, J. L.: 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 (1993).

Negus, V. E.: The comparative anatomy and physiology of the larynx (Hafner, New York 1949).

Nwe, T. L.; Foo, S. W.; Silva, L. C. D.: Speech emotion recognition using hidden Markov models. Speech Communication *41*: 603-623 (2003).

Ohala, J. J.: The acoustic origin of the smile. Journal of the Acoustical Society of America *68*: S33 (1980).

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

Ohala, J. J.: Ethological theory and the expression of emotion in the voice. Proc. ICSLP96, pp. 1812-1815 (1996).

Parr, L. A.; Waller, B. M.: Understanding chimpanzee facial expression: insights into the evolution of communication. Soc Cogn Affect Neurosci *1*: 221-228 (2006).

Pfefferle, D.; Fischer, J.: Sounds and size: identification of acoustic variables that reflect body size in hamadryas baboons, Papio hamadryas. Animal Behavior *72*: 43-51 (2006).

Preuschoft, S.; van Hooff, J. A. R. A. M.: The social function of "smile" and "laughter": variations across primate species and societies; in U. Segerstrale and P. Mobias, Nonverbal Communication: Where Nature Meets Culture, pp. 171-190 (Lawrence Erlbaum Associates, New Jersey 1997).

Reby, D.; McComb, K.: Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. Animal Behaviour *65*: 519-530 (2003).

Scherer, K. R.: Nonlinguistic vocal indicators of emotion and psychopathology; in C. E. Izard, Emotions in personality and psychopathology, pp. 493-529 (Plenum Press, New York 1979).

Scherer, K. R.: Speech and emotional states; in J. Darby, Speech evaluation in psychiatry, pp. 189-220 (Grune & Stratton, New York 1981).

Scherer, K. R.: Vocal affect expression: a review and a model for future research. Psychological Bulletin *99*: 143-165 (1986).

Scherer, K. R.: Vocal communication of emotion: A review of research paradigms. Speech Communication *40*: 227-256 (2003).

Scherer, K. R.; Bänziger, T.: Emotional expression in prosody: a review and an agenda for future research. Proc. Speech Prosody 2004, pp. 359-366 (2004).

Shami, M.; Verhelst, W.: An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication *49*: 201-212 (2007).

Shipp, T.: Vertical Laryngeal Position During Continuous and Discrete Vocal Frequency Change. J Speech Hear Res *18*: 707-718 (1975).

Smith, D. R. R.; Patterson, R. D.; Turner, R.; Kawahara, H.; Irino, T.: The processing and perception of size information in speech sounds. Journal of the Acoustical Society of America *117*: 305-318 (2005).

Sonninen, A.: The role of the external laryngeal muscles in length-adjustment of the vocal cords in singing. Acta Oto-laryngologica *suppl. 130* (1956).

Stevens, K. N.: Acoustic Phonetics (The MIT Press, Cambridge, MA 1998).

Tartter, V. C.; Braun, D.: Hearing smiles and frowns in normal and whisper registers. Journal of the Acoustical Society of America *96*: 2101-2107 (1994).

van Dommelen, W. A.: Does dynamic F0 increase perceived duration? New light on an old issue. Journal of Phonetics *21*: 367-386 (1993).

van Hooff, J. A. R. A. M.: The facial displays of the Catarrhine monkeys and apes; in D. Morris, Primate Ethology, pp. 7-68 (Weidenfeld and Nicolson, London 1967).

van Hooff, J. A. R. A. M.: A comparative approach to the phylogeny of laughter and smiling; in R. Hinde, Nonverbal Communication, pp. (Cambridge University Press, New York 1972).

Waller, B. M.; Dunbar, R. I. M.: Differential Behavioural Effects of Silent Bared Teeth Display and Relaxed Open Mouth Display in Chimpanzees (Pan troglodytes). Ethology *111*: 129-142 (2005).

Xu, Y.: Contextual tonal variations in Mandarin. Journal of Phonetics *25*: 61-83 (1997).

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

Xu, Y.: Speech melody as articulatorily implemented communicative functions. Speech Communication *46*: 220-251 (2005).

Xu, Y.; Liu, F.: Determining the temporal interval of segments with the help of F0 contours. Journal of Phonetics *35*: 398-420 (2007).

Xu, Y.; Wang, Q. E.: Pitch targets and their realization: Evidence from Mandarin Chinese. Speech Communication *33*: 319-337 (2001).