Perception of anger and happiness from resynthesized speech with size-related manipulations

Yi Xu and Andrew Kelly

University College London
yi.xu@ucl.ac.uk, andy.kelly@ucl.ac.uk

ABSTRACT

Recent research has shown that listeners can hear anger and happiness from articulatorily synthesized vowels with body-size-related manipulations. In the present study we explore the possibility that direct manipulation of spectrum and F_0 of naturally produced speech along the size dimension can also lead to perception of certain emotions. Ten English digits spoken in a neutral emotion by a native speaker of British English were resynthesized with spectral and F_0 manipulations to simulate changes in auditory impression of body size. Seven native listeners judged the size and emotion of the speaker. Results show that they heard digits with lower F_0 and/or smaller spectral dispersion as said by a large or angry speaker, and digits with higher F_0 and/or larger spectral dispersion as by a small or happy speaker. These results are consistent with a previous finding based on synthetic speech. This is further evidence that size projection is a basic encoding mechanism for anger and happiness in the vocal expression of emotions.

1. INTRODUCTION

A predominant view about emotional expressions is that they are overt displays of one's internal neuro-physiological state as a reaction to an emotion evoking event [3, 10]. As such the characteristics of the emotional expressions should be understood on the basis of their links to the person's internal states. Findings about various emotions based on this understanding have not been highly consistent, however [15]. An alternative view, as proposed by Ohala in 1996, is that emotional expressions are evolutionarily designed to elicit behaviors from the receiver that are favorable to the signaler [9]. One of the mechanisms proposed to make this work is based on what is known as the size code [2, 4], or frequency code [8]. That is, due to the advantage of a larger body size over a smaller size in cases of physical confrontation, animals of many species have developed strategies to appear as large as possible to scare off the opponent or to win over a potential female mate. They erect their body hair or feathers, stand erect or spread out their wings. In addition to the visual strategies, animals have also developed means to *sound* as large as possible. They lower the voice pitch, roughen the voice quality and lengthen the vocal tract [7, 8]. The size code is also known to be exploited by animals in the opposite direction, i.e., to appear as small as possible to show non-threat, submission and sociability by mimicking infants [9]. Visually, they flatten the ears, body hair or feathers, and crouch down or cringe. Vocally, they raise the pitch, and make the voice quality tone-like, and shorten the vocal track. It has been further proposed that the size projection strategies seen in animals are also used by humans in their emotional expressions [9]. In particular, the human smile, which is homologous to the fear grimace of other primates [13], is for the sake of shortening the vocal track during vocalization [8].

The relevance of the size code to emotional expressions in humans has been demonstrated recently by testing the hypothesis that anger and happiness are vocally encoded by projecting body size along a large-small continuum [2]. In that study, human listeners were asked to judge the size and emotion of the speaker from vowels synthesized with different vocal tract lengths (VTL) and F_0. The results showed that vowels with longer VTL and lower F_0 were heard both as produced by a larger person and as by a person who is angry, and those with a shorter VTL and higher F_0 were heard as produced by a smaller and a happier person.

The goal of the present study was to replicate the findings of [2] with a different method. Instead of generating vowels with an articulaory synthesizer, we resynthesized real human speech while manipulating F_0 and spectral dispersion (inverse of spectral density) along the size-projection dimension. Altering spectral dispersion had the equivalent effect of altering the length of the vocal tract. The use of real speech, if proven effective, would make it easier for future studies to further test the size code hypothesis. More importantly, the method would test the possibility that emotional coding is done in parallel with the coding of other, more linguistic information in speech [14], since all the non-manipulated aspects of the speech signal can be kept as constant as possible.

2. METHOD

2.1. Stimuli

The stimuli were the English digits 1, 2, 3 … 10, spoken by a male speaker of South British English, age 20, recorded in an anechoic chamber at University College London, in an emotionally 'neutral' voice. The spoken digits were then modified in terms of F_0 and spectral dispersion using the program Speech Filing System [5].

Three factors were controlled in modifying the digits: *acoustic parameter* (F_0, spectral dispersion or both), *direction of modification* (up or down) and *manner of modification* (static or dynamic). Thus the total number of stimuli were 3 parameters x 2 directions x 2 manners x 10 digits = 120. Such a design is to avoid combinations of parameter changes that are ambiguous in terms of size projection, e.g., increasing F_0 but decreasing spectral dispersion.

To manipulate fundamental frequency, the median F_0 of all the spoken digits was first set to 106 Hz and then for each digit the F_0 is either raised or lowered by 10 Hz. Also the change is applied either statically, i.e., by the same amount throughout a digit, or dynamically, i.e., gradually increasing the amount of change from 0 to 10 Hz from the onset to the offset of the digit. The manipulation of spectral dispersion was done by either compressing or expanding the entire spectrum by 10%. Like F_0 modification, the spectral changes were applied either statically or dynamically throughout each digit.

2.2. Subjects and Procedure

Seven native speakers of British English participated as subjects. They were university students aged 20-22, 4 males and 3 females, with no self-reported hearing problems.

The perceptual tests were carried out in a quiet room. The tests were run by the ExperimentMFC module of the Praat program [1] on a laptop computer. Subjects listened to the stimuli through a set of BOSE Quiet Comfort 2 Acoustic Noise Cancelling headphones and performed two forced choice tasks in two separate sessions. The first was to determine whether the speaker was large or small in body size, and the second was to determine whether the speaker was angry or happy. During each trial, a resynthesized digit was played once, and the subject indicated his/her decision by pressing a button on the screen.

The tokens were presented in random order and repeated in three blocks for each task. Thus each subject made 360 judgments in a task. The subjects carried out the experiment individually and were given a practice round to customize themselves to the voice and to the procedure of the experiment. They were instructed to make judgments instinctively without thinking too hard.

2.3. RESULTS

2.3.1. *Size perception*

Each of the subjects' responses was coded as 1 for judging the speaker as small or happy, and 0 for judging the speaker as large or angry, and the average of the three repetitions for each combination of parameter changes was used as the response score. Figure 1a displays response scores for body size as a function of acoustic parameter and direction of manipulation. Digits with increased F_0, increased spectral dispersion or both lead to higher scores for smaller body size judgment, while those with decreased F_0, decreased spectral dispersion or both lead to lower scores. A three-way repeated measures ANOVA shows that the effect of manipulation direction is highly significant (F[1,6] = 166.21, $p < 0.001$). Figure 1a also shows that the scores differed across the three parameter conditions, and the differences are significant (F[1,6] = 3.99, $p < 0.05$). However, a Bonferroni/Dunn post hoc test shows that only the difference between F_0 and both F_0 and spectral dispersion is significant. Also the effect of direction becomes larger as the acoustic parameter changes from F_0 to spectrum to both F_0 and spectrum, as is shown by the significant interaction between direction and parameter of manipulation (F[2,12] = 18.25, $p < 0.001$).

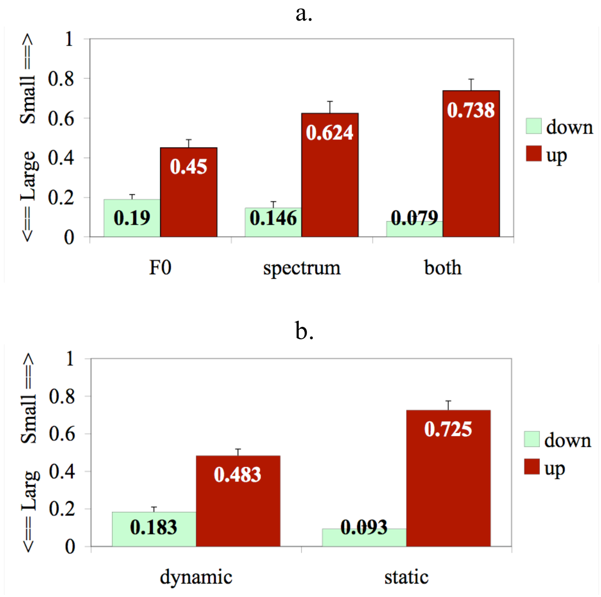

Figure 1. *a) Response scores for body size as a function of acoustic parameter and direction of manipulation. b) Response scores for body size as a function of manner and direction of parameter manipulation.*

Figure 1b shows that size judgment scores are also affected by manner of parameter manipulation (F[1,6] = 15.71, $p < 0.01$). The scores are more extreme when the parameter change is static than when it is dynamic, as indicated by the significant interaction between manner and direction of parameter change (F[2,12] = 128.99, $p < 0.0001$).

These results show that listeners are highly sensitive to the parameter manipulations performed on the spoken digits when judging the body size of the speaker. They judged digits with higher F_0, greater spectral dispersion or both as spoken by a smaller person, and digits with lower F_0, smaller spectral dispersion or both as spoken by a larger person. Also they were more sensitive to the static than the dynamic parameter manipulations.

2.3.2. Emotion perception

Figure 2a displays response scores for emotion as a function of acoustic parameter and direction of manipulation. Digits with increased F_0, increased spectral dispersion or both lead to higher happiness scores, while those with decreased F_0, decreased spectral dispersion or both lead to lower happiness scores. A three-way repeated measures ANOVA shows that the effect of manipulation direction is highly significant (F[1,6] = 79.17, $p < 0.001$). The effect of acoustic parameter is not significant, despite the differences in the means. A Bonferroni/Dunn post hoc test also did not find significant difference between any pair of parameters. There is, however, a significant interaction between direction and parameter of manipulation (F[2,12] = 32.64, $p < 0.001$). This is because the direction effect becomes larger as the acoustic parameter changes from F_0 to spectrum to both F_0 and spectrum.

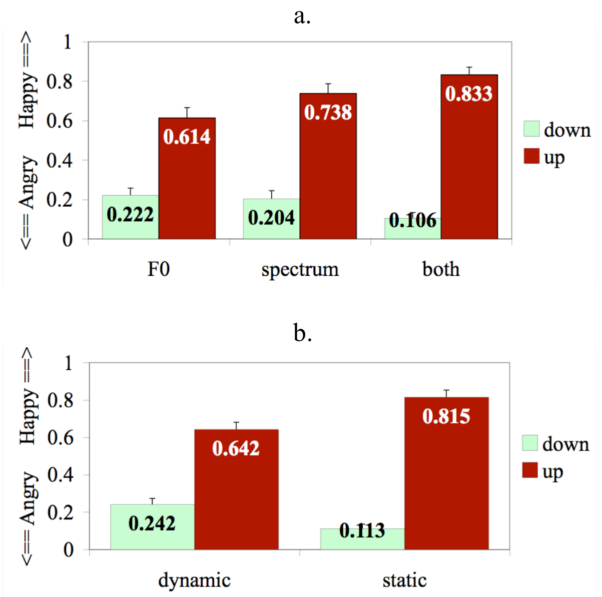

Figure 2. *a) Response scores for emotions as a function of acoustic parameter and direction of manipulation. b) Response scores for emotions as a function of manner and direction of parameter manipulation.*

There is no main effect of manner of manipulation, but there is a significant interaction between manner and direction of parameter change (F[2,12] = 142.64, $p < 0.0001$). This is because the scores become more extreme when the parameter change is static than when it is dynamic.

These results show that listeners are highly sensitive to the parameter manipulations performed on the spoken digits when judging the emotion of the speaker. They judged digits with higher F_0, greater spectral dispersion or both as spoken by a happy person, and digits with lower F_0, smaller spectral dispersion or both as by an angry person. Also they were more sensitive to the static than the dynamic parameter manipulations.

Overall, there is a bias toward hearing a large body size and angry voice, as can be seen in Tables 1 and 2, in which the scores for the down stimuli have been transformed by applying the following equation:

$$S' = 1 - S \qquad (1)$$

where S' is the new score and S the original score. Such a bias could be due to the fact that the speaker who produced the original neutral-emotion digits is a tall male.

Table 1. *Mean size judgment scores computed with equation (1). Standard errors are shown in parentheses.*

Parameter Direction	F_0	Spectrum	Both
down	0.81 (0.024)	0.854 (0.032)	0.921 (0.025)
up	0.45 (0.041)	0.624 (0.06)	0.738 (0.058)
Manner	**dynamic**	**static**	
down	0.817 (0.026)	0.907 (0.016)	
up	0.483 (0.035)	0.725 (0.05)	

Table 2. *Mean emotion judgment scores computed with equation (1). Standard errors are shown in parentheses.*

Parameter Direction	F_0	Spectrum	Both
down	0.778 (0.036)	0.796 (0.041)	0.894 (0.028)
up	0.614 (0.053)	0.738 (0.05)	0.833 (0.039)
Manner	**dynamic**	**static**	
down	0.758 (0.031)	0.887 (0.023)	
up	0.642 (0.039)	0.815 (0.038)	

3. DISCUSSION

The results show that listeners are highly sensitive to variations in F_0 and spectral dispersion in judging both body size and emotion even when the manipulations are performed on naturally spoken words. Increased F_0 and spectral dispersion lead to perception of smaller body size and happiness, and decreased F_0 and spectral dispersion lead to perception of larger body size and anger. The perceptual sensitivity in the case of body size judgment agrees well with previous findings on size perception [6, 11, 12]. The sensitivity of emotion judgment is consistent with the findings of [2]. Thus further support is seen in the present results for the size code hypothesis.

One finding of Chuenwattanapranithi et al. [2] not replicated here is that temporally dynamic parameter manipulations did not lead to more consistent emotion judgment. Rather, it is the stimuli with fixed parameter shifts that elicited more consistent judgments. A likely explanation is that the acoustic parameters in question — F_0 and spectral properties — were already dynamic in the spoken digits, whereas in [2] the parameters in the steady-state vowels were genuinely static. It is possible that the sensitivity of emotional perception is subject to the presence/absence rather than the magnitude of dynamic movements. Another possibility is that the dynamic manipulations performed in the present experiment generated smaller overall differences in F_0 and spectral property than in [2], judging from the fact that in Figure 1, the difference in score is much smaller in the dynamic condition than in the static condition. This is rather different from the very similar size judgment difference between static and dynamic conditions in [2].

It is unlikely that the current results are the artifacts of a two-way forced choice task. Anger and happiness are among the most difficult to distinguish from each other in terms of acoustic cues [10], but their perceptual discrimination has rarely been tested without other emotions. The usual practice of including many emotions in a recognition test actually often helps to hide the difficulty in processing certain emotions. That listeners consistently use the size-related cues as found in the present study as well as in [2] means that these cues are highly relevant to the distinction between anger and happiness. This is similar to the findings about many other aspects of speech which are made by studies that also use two-way forced choice to test the effectiveness of acoustic cues for various phonetic contrasts.

4. CONCLUSIONS

The present results have replicated the findings of [2] that manipulating F_0 and spectral property of speech along the dimension defined by the size code leads to consistent perceptual judgment of both body size and emotion in terms of anger versus happiness. This is further indication that size projection is a highly effective mechanism for encoding the emotional contrast between anger and happiness. It also provides further evidence in support of the view that emotional expressions are evolutionarily designed to elicit behaviors beneficial to the emotion bearers rather than to just reflect their internal neurophysical states [9, 15]. Finally, the acoustic manipulations in the present study were done directly on short spoken utterances through resynthesis. The effectiveness of this method suggests that it can be used in future research to manipulate more complex speech utterances and to study expressions of emotions other than happiness and anger.

5. ACKNOWLEDGEMENT

We thank Mark Huckvale for helping us with the generation of the stimuli using SFS.

6. REFERENCES

[1] Boersma, P., 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10: 341-345.

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

[3] Ekman, P., 1997. Expression or communication about emotion. In *Uniting Psychology and Biology: Integrative Perspectives on Human Development.* N. Segal, G. E. Weisfeld and C. C. Wiesfeld. (eds.) Washington, DC: APA: 315-338.

[4] Gussenhoven, C., 2002. Intonation and interpretation: Phonetics and Phonology. In Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France. pp. 47-57.

[5] Huckvale, M., 2008. SFS Speech Filing System 4.7, http://www.phon.ucl.ac.uk/resource/sfs/, University College London.

[6] Ives, D. T.; Smith, D. R. R.; Patterson, R. D., 2005. Discrimination of speaker size from syllable phrases. *Journal of the Acoustical Society of America* 118: 3816-3822.

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

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

[9] Ohala, J. J., 1996. Ethological theory and the expression of emotion in the voice. In *Proceedings of ICSLP96*. pp. 1812-1815.

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

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

[12] Turner, R. E.; Patterson, R. D., 2003. An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited. *Journal of the Acoustical Society of Japan* 33: 585–589.

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

[14] Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46: 220-251.

[15] Xu, Y.; Kelly, A.; Smillie, C., forthcoming. Emotional expressions as communicative signals. To appear in *Prosody and Iconicity*. S. Hancil and D. Hirst (eds).