Discrimination and Recognition of Scaled Word Sounds

Toshio Irino[1], *Yoshie Aoki*[1], *Yoshie Hayashi*[1], *Hideki Kawahara*[1], *and Roy D. Patterson*[2]

[1] Faculty of Systems Engineering, Wakayama University, Japan.
[2] CNBH, Dept. of Physiology, Development, and Neuroscience, Cambridge Univ., U.K.
{irino, kawahara}@sys.wakayama-u.ac.jp, rdp1@cam.ac.uk

Abstract

Smith et al. [2] and Ives et al. [3] demonstrated that humans could extract information about the size of a speaker's vocal tract from speech sounds (vowels and syllables, respectively). We have extended their discrimination and recognition experiments to naturally pronounced words. The Just Noticeable Difference (JND) for size discrimination was between 5.5% and 19% depending on the listener. The smallest JND is comparable to that of the syllable experiments; the average JND is comparable to that of the vowel experiments. The word recognition scores remain above 50% for speaker sizes beyond the normal range for humans. The fact that good performance extends over such a large range of acoustic scales supports Irino and Patterson's hypothesis [1] that the auditory system segregates size and shape information at an early stage in the processing.

1. Introduction

The sounds that convey words from speaker to listener contain information about the length of the speaker's vocal tract as well as its shape (the message). Humans can extract the message from the voices of men, women, and children without being confused by the size information, and they can extract the size information without being confused by the message. This suggests that the auditory system can extract and separate information about the size (or, strictly speaking, the acoustic scale) of the vocal-tract from information about its shape. It has been hypothesized that the auditory system applies a scale transform to sounds to accomplish the segregation [1]. Smith et al. [2] performed discrimination and recognition experiments using vowel sounds pronounced from vocal tracts which are equivalently dilated and contracted (or scaled) in the length. They showed that the ability to discriminate speaker size extends beyond the normal range of speaker sizes. Ives et al. [3] performed size discrimination experiments using a much larger set of speech sounds (180 syllables) and found similar results. Van Dinther and Patterson [4] reported discrimination and recognition experiments on acoustic scale and size perception of musical instruments. Together the experiments suggest that information about the size of the source is segregated from information about the shape and structure of the source, automatically, at an early stage in the processing.

This paper reports an extension of the size experiments using a set of Japanese, four-mora[1)] words which were pronounced naturally by a native speaker to determine whether the size perception results reported earlier are robust; that is, the phenomena are observed with natural speech sounds by untrained listeners. The experimental methods are essentially the same as the previous studies [2,3]. We describe the size discrimination experiment in session 2 and the word recognition experiment in section 3.

[1)] One morae corresponds to a consonant-vowel syllable or a single-vowel syllable; so, 'Ka-wa-ha-ra' is a four-mora word.

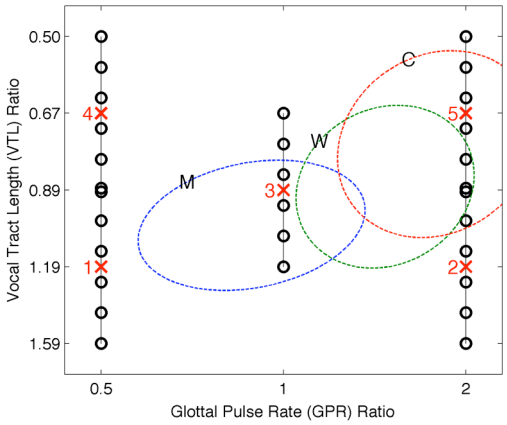

Figure 1. The GPR-VTL combinations for the discrimination experiments. The five reference speakers are shown by the numbered red crosses and the test speakers by the black circles. The ellipses show approximate distributions of GPR and VTL values in normal speech for men, women, and children.

2. Size discrimination experiment

2.1. Method

2.1.1. Synthesis of stimuli

The stimuli consisted of words that were analyzed, manipulated, and resynthesized by STRAIGHT, a high-quality vocoder [5,6] commonly used in auditory research for communication sounds [1-3,7]. STRAIGHT allows one to separate the vocal tract length (VTL) and glottal pulse rate (GPR) information in speech sounds and resynthesize the same utterance with different combinations of VTL and GPR values. VTL is varied simply by dilating or contracting the STRAIGHT spectral envelope of the original sound. Thus, the change in VTL is inversely proportional to the spectral envelope ratio (SER).

We used Japanese four-mora words selected from a database (FW03) controlled with respect to both word familiarity and phonetic balance [8]. All of the words were spoken by one male speaker (mya). The average fundamental frequency was about 150 Hz in voiced segments over all words. The height of the speaker is unknown. For scaling

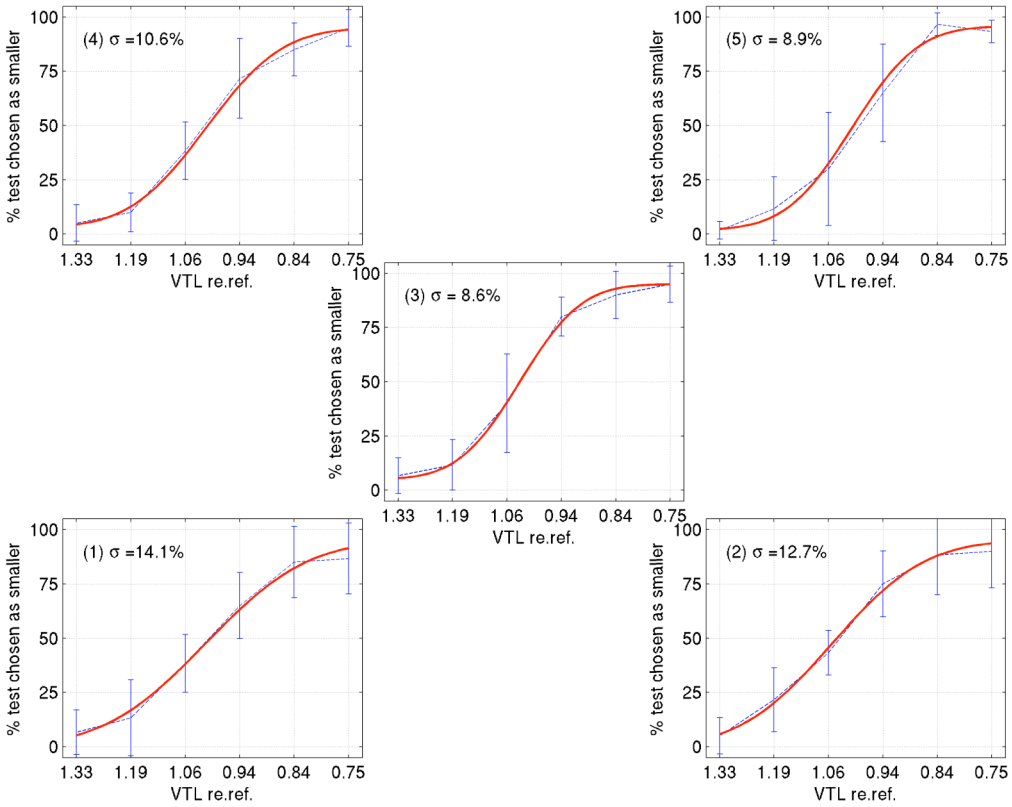

Figure 2. Psychometric functions for the triple-word condition for the five reference speakers (shown in Fig. 1) averaged over all listeners. The error bars represent ±1 standard deviations over all listeners.

purposes, we assumed he was 170 cm which is about average for a Japanese middle-aged man. The data of Fitch and Giedd [9] suggest that his VTL would be about 15 cm. So, GPR and VTL ratios are normalized to F0 and VTL values of 150 Hz and 15 cm, respectively.

2.1.2. Experimental design

Figure 1 shows the five combinations of GPR ratio and VTL ratio (=1/SER) that were used in the experiment; they were chosen to be characteristic of five speaker types. The red crosses show the ratio for standard sounds and the black circles show the ratio for test sounds for each standard; The GPR-VTL ratios for the standards were (1) [0.5, 1.19], (2) [2, 1.19], (3) [1, 0.89], (4) [0.5, 0.67], and (5) [2, 0.67]; The VTL ratios of the test sounds are $2^{-5/12}$, $2^{-3/12}$, $2^{-1/12}$, $2^{1/12}$, $2^{3/12}$, and $2^{5/12}$ (0.75, 0.84, 0.94, 1.06, 1.19, and 1.33) relative to the standard without changing the GPR ratio. The ellipses (dashed lines) show the approximate distribution of GPR and VTL ratios in normal speech for men, women and children [2,3]. The standards cover a range of VTL and GPR values that extends beyond that normally encountered during everyday experience.

Informal listening suggested that it was difficult to judge speaker size from a single word because of the variation in pitch. So, we prepared three types of stimuli: each word sequence contains either triples of words, pairs of words, or a single word. All of the words were selected at random from a list of 1000, highly familiar words. The number of words for sessions with single words, pairs of words, and triplets of words was 600, 1200, and 1800, respectively. Although there was some overlap, the combinations of words in the pairs and triplets, they were chosen randomly. Then, they were resynthesized with specific combinations of GPR and VTL, and concatenated to form the required sequences.

Table I. Jnd values for triple-word, double-word, and single-word combinations, with the jnd's for the vowel [2], and syllable[3] experiments. The numbers under 'GPR-VTL ratio condition' correspond to the five reference speakers shown in Fig. 1.

		GPR-VTL ratio condition					
		1	2	3	4	5	mean
word	triple	14.1	12.7	8.6	10.6	8.9	11.0
	double	13.4	8.4	10.8	8.8	6.2	9.5
	single	10.7	12.1	9.9	10.5	9.8	10.6
Vowel		10.5	17.2	6.6	10.6	9.3	10.8
Syllable		4.3	4.8	4.1	6.6	5.7	5.1

2.1.3. Listeners, task, and settings

Six Japanese subjects aged between 21 and 33 participated in the experiments; they all had normal hearing thresholds between 250 and 8000 Hz.

On each trial, the listener was presented with a standard-word sequence and a test-word sequence; triple words in the first session, double words in the second session, and single words in the final session. The only consistent difference between the two intervals of a trial was the VTL ratio; the listener's task was to identify the interval with the smaller speaker. The total number of trials per subject was 900 (5

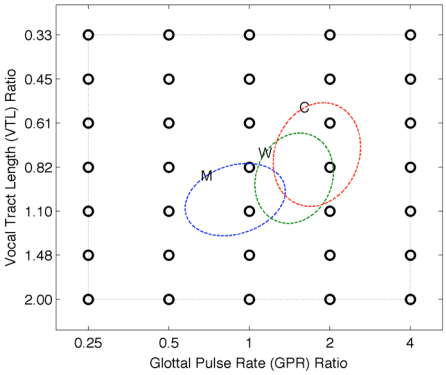

Figure 3. The GPR-VTL combinations for the recognition experiments. The ellipses are the same as in Fig. 1. A GPR ratio of 1 corresponds to an F0 of 150 Hz.

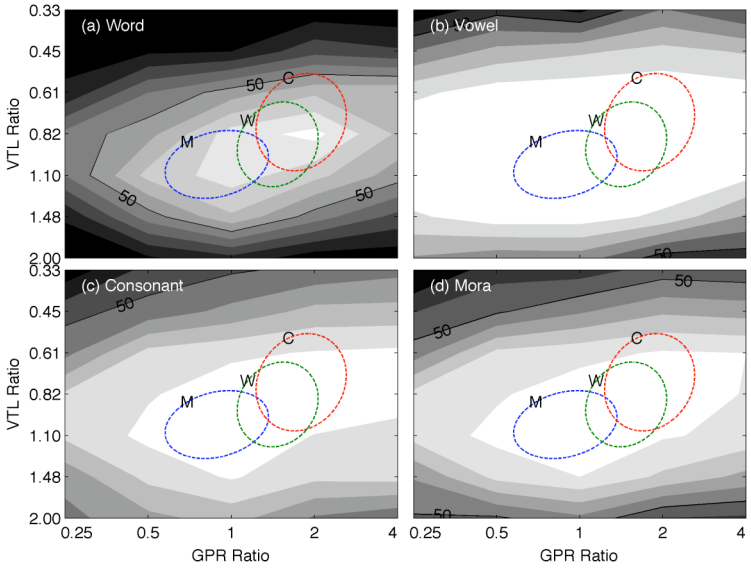

Figure 4. Recognition performance. Gray tones show mean percent correct, in 10% steps, calculated from 10 trials for each of 6 listeners. The white region shows where the recognition rate exceeds 90 % correct.

references x 6 pairs x 10 trials x 3 sessions). Feedback was provided during an initial, brief practice session, but not in the test sessions.

The subject was seated in a soundproof room. The words were played by M-audio FireWire 410 with a 48-kHz sampling rate over headphones (Sennheiser HD-580) at an rms level of 70 dB SPL on average. The sound level was roved between words over a 6-dB range.

2.2. Results

Figure 2 shows the results of the triple-word experiment for each of the five reference speakers (referred to as 'standard speakers') as a set of psychometric functions. The data have been averaged across listeners. The abscissa for the psychometric functions is VTL normalized to the reference VTL; the ordinate is the percentage of trials on which the test interval was identified as having the smaller speaker. A cumulative Gaussian function has been fitted to the data for each psychometric function [10] and used to calculate the JND, defined as the difference in VTL for a 26% increase in performance from 50% to 76% correct (d'=1) [2,3]; the JND is shown in each panel at the top left corner. Figure 2 shows that size discrimination is possible for all five speaker types.

The results for the double-word and single-word experiments showed very similar patterns of performance and levels of performance. The JND values for these word experiments are summarized in Table I and compared with the results for vowels [2] and syllables [3] from previous experiments. There is no consistent trend of the JND values across the GPR-VTL conditions. The mean JNDs are almost the same for conditions with three, two or one word per interval; they are all about 10% which is about the same as in the vowel experiment and about twice the values in the syllable experiment.

There was considerable variability across listeners: The individual JNDs averaged over all conditions were 5.5%, 5.9%, 7.8%, 10.2%, 18.7%, and 19.2%. There was no consistent tendency for one GPR-VTL combination to be best or worst across listeners. The listeners with small JND values perform the task as well as the listeners in the syllable experiment [3]. It is possible that the listeners with large JND values were confused by changes in pitch between words.

The number of the words in the sequence had little effect on performance; listeners can judge relative size given a single word spoken with natural, unconstrained, pronunciation. Fundamental frequency (F0) varies within a word and the F0 contours are different between the words in the two intervals. So the pitch information is not consistently useful for size discrimination. Moreover, it is difficult to perform the task simply by comparing the spectral centroids of the word since they were randomly selected for both intervals. A more sophisticated process is required to extract and compare the size information contained in the sequence of words.

3. Word recognition experiment

3.1. Method

Figure 3 shows the 35 combinations of VTL ratio and GPR ratio in the word recognition experiment. The range of VTL and GPR ratios extends far beyond the normal range indicated by the three ellipses which are the same as in Fig. 1. The methods of sound synthesis are essentially the same as in section 2.1 except that the sound level was fixed at 70 dB SPL.

We selected words at random from the word list with the lowest familiarity. These words are rarely used in everyday life and many would not be recognized as words by many people even though they are listed in the dictionary. This list was chosen to restrict the use of the mental lexicon to assist guessing. The listeners were presented with one word and asked to write what they perceived in Japanese 'kana' - characters which uniquely identify the 'morae' of Japanese, which are like the consonant-vowel (or simple-vowel) syllables of other languages.

Six Japanese subjects aged between 21 and 23 with normal hearing thresholds in the range 250 to 8000 Hz participated in the experiments. Four of them also participated in the size discrimination experiment.

The total number of trials was 350 (7 VTLs x 5 GPRs x 10 trials); they were spread over 5 sessions. There was no practice session.

3.2. Results

Recognition performance averaged over the six listeners is shown in Fig. 4 for four different recognition measures, namely, the percent correct for the complete word (a), the vowels in the word (b), the consonants in the word (c), and the morae in the word (d). The abscissa is GPR ratio and the ordinate is VTL ratio. The percent correct is given by the tone of gray. The points where performance was measured are shown by the small, black circles in Fig. 3. The gray-scale density and the contours were created by interpolation between the data points. The ellipses show the approximate distribution of GPR and VTL ratios in normal speech for men (M), women (W), and children (C). Figure 4a shows that word recognition performance is greater than 50% throughout the normal range and beyond.

In order to compare the results to those of previous studies [2], we calculated the recognition scores separately for the vowels, consonants, and morae within the words. The vowel recognition performance in Fig. 4b is very similar in pattern and overall level to that reported previously with isolated vowels [2]. The thick black contour marks 50 % recognition rate which is drawn for comparison with the recognition threshold (~50%, d'=1.0) for the 5AFC experiment in [2]. Performance drops to 50% only when the VTL ratio is more than 1.8 or less than 0.4 and this is largely independent of GPR ratio in the range 0.25 to 4 (average F0 values from 38 Hz to 600 Hz). Figures 4c and 4d show that recognition performance for consonants and morae was also greater than 50% in throughout the normal range and well beyond.

The results show that the perception of vowels and consonants is very robust to changes in GPR and VTL, even when their values are well beyond the normal range. Performance is probably assisted by co-articulation in these four-morae words and implicit knowledge concerning the statistical distribution of syllable sequences in Japanese words. Results from other experiments using the same database [11] suggest that performance would still be reasonably high even if these effects were taken into account.

4. Summary and Conclusions

We performed discrimination and recognition experiments with words whose acoustic scale and glottal-pulse rate was varied over a large range, to determine how results with scaled vowels and syllables [2,3] would be affected by word context. The original word sounds with natural pronunciation were selected from a well-controlled database and they were scaled over the full range of normal VTL-GPR combinations observed in human speech and well beyond. The JND values for the size discrimination task ranged from 5.5% to 19% depending on the listener. The smallest JND is about the same as those in previous syllable experiments and the average JND is about the same as those in previous vowel experiments. Word recognition performance was greater than 50% in the normal range and beyond Indicating that performance is robust to changes in VTL and GPR.

In everyday life, we encounter very few instances of speech with VTL and GPR combinations beyond the normal region, so there is very little data available in the normal world to assist a learning mechanism to generalize from normal speech to the extreme combinations in the current experiments. The robustness of speech perception suggests that the auditory system has an automatic mechanism to segregate size and shape information prior to the application of discrimination and recognition processing [1].

Acknowledgments:

This work was supported in part by the Japan Society for the Promotion of Science under Grant-in-Aid for Scientific Research (B), 18300060, and the UK Medical Research Council (G0500221).

References

[1] Irino, T. and Patterson, R. D., "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised Wavelet Mellin Transform, " *Speech Commun.*, **36** (3–4), pp. 181–203, 2002.

[2] Smith, D.R.R., Patterson, R. D., Turner, R., Kawahara, H. and Irino, T., "The processing and perception of size information in speech sounds," *J. Acoust. Soc. Am.* , **117**(1), pp. 305-318, 2005.

[3] Ives, D. T., Smith, D.R.R., and Patterson, R. D., "Discrimination of speaker size from syllable phrases," *J. Acoust. Soc. Am.*, **118** (6), pp.3816-3822, 2005.

[4] van Dinther, R. and Patterson, R. D., "Perception of acoustic scale and size in musical instrument sounds," *J. Acoust. Soc. Am.* **120**, pp.2158-2176, 2006.

[5] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, **27** (3-4), pp.187-207, 1999.

[6] Kawahara, H., "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Tech.*, **27**(6), pp. 349–353, 2006.

[7] Liu, C., and Kewley-Port, D., "STRAIGHT: a new speech synthesizer for vowel formant discrimination," *ARLO* **5**, pp.31–36, 2003.

[8] Sakamoto, S. , Iwaoka, N. , Suzuki, Y., Amano, S., and Kondo, T. "Compliment relationship between familiarity and SNR in word intelligibility test," *Acoust. Sci. Tech.*, **25**(4), pp. 290–292, 2004.

[9] Fitch, W. T., and Giedd, J., "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Am.* **106**, 1511–1522, 1999.

[10] Wichmann, F. A. and Hill, N. J., "The psychometric function: I. Fitting, sampling and goodness-of-fit," *Perception and Psychophysics*, **63**(8), pp.1293-1313, 2001.

[11] Irino, T., Satou, S., Nomura, S., Banno, H., and Kawahara, H., "Speech intelligibility derived from time-frequency and source smearing," *Proc. Interspeech 2005*, pp.1737-1740, 2005.