Influences of vowel duration on speaker-size estimation and discrimination

Chihiro Takeshima¹, Minoru Tsuzaki¹, and Toshio Irino²

¹Graduate School of Music, Kyoto City University of Arts, Japan  
²Faculty of Systems Engineering, Wakayama University, Japan

ctakeshima@kcua.ac.jp, minoru.tsuzaki@kcua.ac.jp, irino@sys.wakayama-u.ac.jp

Abstract

Several experimental studies have shown that the human auditory system has a mechanism for extracting speaker-size information, using sufficiently long sounds. This paper investigated influence of vowel duration on the processing for size extraction using short vowels. In a size estimation experiment, listeners subjectively estimated the size (height) of the speaker for isolated vowels. The results showed that listeners’ perception of speaker size was highly correlated with the factor of vocal-tract length in all the tested durations (from 16 ms to 256 ms). In a size discrimination experiment, listeners were presented with two vowels scaled the vocal-tract length and were asked which vowel was perceived to be spoken by a smaller speaker. The results showed that the just-noticeable differences (JNDs) in speaker size were almost the same for the durations longer than 32 ms. However, the JNDs rose considerably for 16-ms duration. These observations of the experiments suggest that the auditory system can extract speaker-size information even for 16-ms vowels although the precision of size extraction would deteriorate when the duration becomes less than 32 ms.

Index Terms: auditory processing for speaker-size extraction, duration, size discrimination, subjective size estimation

1. Introduction

There is information about speaker size in speech sounds. For example, as humans grow in height, the vocal-tract length (VTL) increases [1], and as a result, the formant frequencies of the vowels decrease [1, 2, 3]. Irino and Patterson [4] have hypothesized that the auditory system automatically extracts and normalizes the size information from sounds, and thereby enhance both perception of speaker size and speech recognition. Several psychophysical studies demonstrated that the human auditory system has a mechanism for the size extraction and normalization by size discrimination and recognition experiments using vowels [5], syllables [6], words [7] and musical instrument sounds [8]. In these studies, the sound duration was longer than 260 ms to make listener’s size judgment stable.

The previous studies have not focused on the minimum duration of sounds required for stable size extraction. Recently, Takeshima et al. [9] performed experiments of recognition of vowel sequence in which the VTLs were alternated per 85 ms and they showed a possibility that the auditory system segregates the vowel sequence into two perceptual streams based on speaker-size information. They suggested the occurrence of stream segregation reflects the fact that the auditory processing for size extraction operates appropriately for 85-ms vowels.

This paper more systematically investigated influences of vowel duration on the precision of size extraction. We performed two psychophysical experiments of absolute judgment (size estimation) and relative judgment (size discrimination) using single vowels with shorter duration than that of the previous studies of size estimation [10] and size discrimination [5] (500 ms).

2. Experiment of speaker-size estimation

This experiment investigated influences of vowel duration on absolute judgment of speaker size.

2.1. Method

2.1.1. Stimuli

We recorded five vowels (/a/, /e/, /i/, /o/, /u/) spoken by a Japanese male speaker, using a microphone (Schoeps CMC5-U) with a 16-bit quantization and 44.1-kHz sampling rate. The height of the speaker was 160 cm. According to the regression data of speaker height versus VTL in the study of Fitch and Giedd [1], the VTL value of the speaker can be estimated approximately 14 cm.

The vowels were manipulated and resynthesized using STRAIGHT [11]. It is a sophisticated vocoder which analyzes an utterance and it extracts spectral envelope information (i.e., information about vocal-tract shape and length) independent of glottal pulse rate (GPR). VTL of the speaker can be scaled by compressing or dilating the spectral envelope of the original vowels.
speech sound along a linear frequency axis. Thus, the change in VTL is inversely proportional to the spectral envelope ratio (SER). GPR can be scaled by compressing or dilating the time axis of the sequence of glottal events.

GPR and VTL values in this experiment are shown in Fig. 1 (open circles). The GPR value of the stimuli was set to 125 or 250 Hz. The VTL ratio (=1/SER) was set to $2^{-13/12}, 2^{-9/12}, 2^{-5/12}, 2^{-1/12}, 2^{3/12}, 2^{7/12},$ or $2^{11/12}$ (0.47, 0.59, 0.75, 0.94, 1.19, 1.50, or 1.89).

After scaling of GPR and VTL by STRAIGHT, a stationary part of the vowel was selected with duration of 16, 32, 64, 128 or 256 ms. Then, the scaled vowels were gated on and off over 5 ms using a cosine-squared amplitude function. The stimuli had at least a single pulse information in all duration.

### 2.1.2. Listeners and procedures

Four listeners, aged from 20 to 30, participated in this experiment. They reported no history of hearing impairment.

They were seated in a sound-attenuating booth. Stimuli were presented by a DSP (Digital Signal Processing) system controlled by a workstation (SimbolicSound Capybara 350 controlled by Apple iMac G5) through headphones (Sennheiser HD 600 amplified with Luxman P1) at a level of about 60 dB SPL. The level was roved in intensity within ±3 dB range in each trial.

The listeners were presented with a scaled vowel and had to estimate the speaker size (1:very short, 2:short, 3:rather short, 4:average, 5:rather tall, 6:tall, or 7:very tall) and answer the confidence level of the size estimation (1:very low, 2:low, 3:medium, 4:high, or 5:very high). The responses were made by selecting the corresponding button on a response box displayed on a graphical interface. Listeners were required to estimate subjectively with no feedback. Each experimental session consisted of 350 trials (5 vowel types × 7 VTL ratios × 5 durations × 2 GPRs). The order of the trials was randomized. There were six replicated sessions.

### 2.2. Results and discussion

For each listener, the first session was regarded as a practice and its data was omitted from the analysis. Therefore, the following analysis was performed for the other data (350 trials × 5 sessions) except for data of the first session.

Figure 2 shows perceived speaker-size rating for each duration. High values indicate a large speaker, and vice versa. The data show that there is an abrupt decrease and vice versa. The data show that there is an abrupt decrease in confidence level for both GPRs when the stimulus duration is 16 ms. To examine difference between 16 ms and the other durations, a planned contrast test for each GPR was performed between data for 16 ms and those for the other durations, after performing one-way analysis of variance (ANOVA) with a factor of duration on the confidence data pooled over the other factors (i.e., VTL ratio and vowel conditions). The results revealed that the confidence level between them were significantly different for both of 125 Hz and 250 Hz [F(1, 95) = 19.67, p < 0.0001, F(1, 95) = 28.75, p < 0.0001, respectively].

This decrease in confidence level for 16 ms suggest that the precision of the size extraction decreases for the duration; that is, it might become more difficult to estimate the speaker size for 16 ms than for the other longer durations. To clarify the degree of the precision of the size extraction, we performed an experiment of speaker-size discrimination (relative judgment task) in the next section.

Table 1: Mean listener’s confidence levels for each duration in the size estimation experiment.

<table>
<thead>
<tr>
<th>GPR</th>
<th>16 ms</th>
<th>32 ms</th>
<th>64 ms</th>
<th>128 ms</th>
<th>256 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>125 Hz</td>
<td>3.07</td>
<td>3.67</td>
<td>3.80</td>
<td>3.80</td>
<td>3.82</td>
</tr>
<tr>
<td>250 Hz</td>
<td>3.23</td>
<td>3.55</td>
<td>3.66</td>
<td>3.72</td>
<td>3.70</td>
</tr>
</tbody>
</table>

Figure 2: Mean perceived speaker-size ratings and the regression lines as a function of VTL ratio of the stimulus. Each symbol represents data for 16 ms (○), 32 ms (○), 64 ms (●), 128 ms (△), and 256 ms (×). Panel (a) and (b) represent data for 125 and 250 Hz, respectively, where the error bars are ±1 standard deviation across listeners.
3. Experiment of speaker-size discrimination

This experiment investigated influences of vowel duration on discrimination of a slight change in speaker size.

3.1. Method

3.1.1. Stimuli

This experiment used same vowel sounds uttered by a male speaker as in the size estimation experiment. The VTL ratio and GPR were manipulated using STRAIGHT. The GPR values were set to 125 and 250 Hz as in the size estimation experiment.

3.1.2. Listeners and procedures

Four listeners, aged from 20 to 23, participated in this experiment; they reported no history of hearing impairment. Three listeners were who participated in the estimation experiment and another one was new listener.

The Just-Noticeable Difference (JND) in speaker size was measured in a size discrimination experiment. The experiment employed a two interval, two-alternative, forced-choice paradigm (2AFC) with the method of constant stimuli. One interval contained the standard stimulus, the other contained the test stimulus in which the simulated VTL of the speaker was either larger or smaller. The inter-stimulus interval was 500 ms. The asterisks and crosses in Fig. 1 show VTL ratio of the standard and test stimulus, respectively. The VTL ratio of the standard stimulus was set to 0.94. The VTL ratio of the test stimulus was set to $2^{-5/12}$, $2^{-3/12}$, $2^{-1/12}$, $2^{1/12}$, $2^{3/12}$, and $2^{5/12}$ (0.75, 0.84, 0.94, 1.06, 1.19, and 1.33) relative to that of the standard stimuli. The two temporal intervals of each trial contained a single vowel either of the same category (e.g., /a/ was paired with /a/) or different category (e.g., /a/ was paired with /ɛ/); the former condition is referred to as same vowel combination and the latter as different vowel combination.

The listeners’ task was to choose the interval in which the vowel was perceived to be spoken by the smaller speaker by selecting the corresponding button on a graphical interface. No feedback was provided after the response. The total number of trials for each listener was 3000 [25 vowel combinations (5 same vowel types and 20 different vowel types) × 6 VTL ratios of test stimulus × 5 durations × 2 GPRs × 2 orders of the standard and test stimulus]; they were divided into 6 experimental sessions.

To familiarize the listeners with the task, each listener performed a training session with 100 trials before the main experiment. During this training session, feedback was provided after the response. To minimize the training effect in the main experiment, listeners were presented with scaled vowels with only large VTL difference and the speaker of the vowels was different (i.e., a female speaker) from that of the main experiment.

3.2. Results and discussion

The results from the six VTL ratios of the test stimuli were combined for each stimulus condition, and they were fitted to an S-shaped function (logistic sigmoid function) to produce psychometric functions. The examples of the psychometric functions are shown in Fig. 4. Each data point is percentage of trials on which the test stimulus was judged to be spoken by the smaller speaker. The smooth curve is best-fitting sigmoid function for same vowel combination (solid curve) and different combination (dashed curve).

Figure 3: Regression slope values as a function of duration.

Figure 4: Examples of psychometric functions for speaker-size discrimination, as a function of VTL ratio of the test stimulus relative to VTL of the standard stimulus. The examples are one listener’s data for GPR of 125 Hz and duration of 256 ms. Each symbol represent percentage of times the test stimulus was judged to be spoken by the smaller speaker. The smooth curve is best-fitting sigmoid function for same vowel combination (solid curve) and different combination (dashed curve).
not a simple function of VTL change when the original formant frequencies differ between different vowel combinations. Therefore, we focus on data only for the different combination in the following discussions, assuming that listeners perform the task based on only a change in perceived speaker size for this condition.

For 32 ms or longer durations, the data show low JND values; the mean JNDs ranged from 7.4% to 13%. There was no large change in JND value depending on the duration. For 16-ms duration, all the JNDs were higher than JNDs for the other longer duration. The JND values for 16 ms had considerable variability across listeners; the JND values ranged from 10.5% to 87%.

To examine whether there was any systematic deviation of JND between durations, one-way ANOVA was performed on the JNDs for different vowel combination, where the main factor was duration. The ANOVA revealed a highly significant effect of duration \( F(4, 35) = 5.1185, p < 0.0024 \). Furthermore, we calculated a post hoc, paired comparison using Tukey-Kramer honestly significant difference (HSD) test. The result revealed there was significant difference between 16 ms and the other longer durations \( p < 0.05 \). There is no significant difference between durations longer than 16 ms. These results indicate that it is more difficult to discriminate the size change for 16-ms vowels than for the other longer durations. This is consistent with observation in the estimation experiment, suggesting that the precision of size extraction decreases remarkably for 16 ms.

4. Summary and conclusion

We performed two experiments of speaker-size estimation and discrimination using single vowels with various duration to investigate influences of duration on the auditory processing for speaker-size extraction. For the estimation task, listeners’ perception of speaker size was highly correlated with VTL factor of vowels even for 16-ms duration. For the discrimination task, there was no large difference in the value of speaker-size JND when the vowel duration was 32 ms or longer; the JND values were less than 13%. When the duration was 16 ms, the JND values rose considerably for both GPRs. These results of the two experiments suggest that the processing for speaker-size extraction operates appropriately even for short duration, although it would need longer duration than 16 ms for discrimination of a slight change in VTL.

5. Acknowledgement

This work was supported in part by the Research Fellowships for Young Scientists and Grant-in-Aid for Scientific Research (B), No.18300060, Japan Society for the Promotion of Science.

6. References