The Vowel-Size Relationship Re-Examined Using Speeded Classification

Yuka Ohtake (4588208926@mail.ecc.u-tokyo.ac.jp)
Graduate school of Education, University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan

Etsuko Haryu (haryu@p.u-tokyo.ac.jp)
Graduate school of Education, University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan

Abstract

The vowel-size relationship has been repeatedly reported: the vowels /a/ and /i/ elicit bigger/smaller images respectively. Previous studies reporting this relationship have required participants to make explicit decisions about the meaning of the target words including these vowels. In the present study, we attempted to re-examine the vowel-size relationship in two experiments using speeded classification tasks. The results of Experiment 1 indicate that participants associated the vowels with a bigger/smaller image even when they were not motivated to pronounce the vowels during the task. The results of Experiment 2 indicate that the proprioception of the absolute size of the mouth may not contribute to the vowel-size relationship. The process underpinning the vowel-size relationship is discussed.

Keywords: sound symbolism; vowel-size relationship; speeded classification; kinesthetic experience

Introduction

The relationship between a word and its referent is said to be arbitrary, but many studies have reported relationships between them. This is referred to as “sound symbolism.” Among these studies, the vowel-size relationship has been repeatedly reported: the vowel /a/ is likely to make us imagine objects of bigger size whereas the vowel /i/ elicits images of smaller objects. Previous studies (e.g., Sapir, 1929; Newman, 1933; Tarte & Baritt, 1971) that reported this relationship have required participants to make explicit decisions about the meaning of target words including these vowels. For example, Sapir (1929) asked participants which table was bigger, /mal/ or /mil/, while Tarte & Baritt (1971) required participants to match CVC trigrams (e.g., /was/ or /wis/) with geometric figures of different sizes. However, these studies suffer from a weakness in the way the process underpinning the vowel-size relationship was investigated: participants had enough time to pronounce or simulate the target words during the task, which makes it difficult to determine which of the following two factors contributed to the vowel-size correspondence.

The first of these factors is the component formant frequencies of the vowels (Tarte, 1982). The second and third formants of the vowel /i/ are higher than those of the vowel /a/. Given that higher frequency sounds correspond to smaller images and lower frequency sounds correspond to bigger images (Gallace & Spence, 2006), frequencies of vowels may explain why the vowel /a/ is likely to elicit bigger images and /i/ to elicit smaller images.

The second is the contribution of the kinesthetic experience of pronunciation (e.g., Newman, 1933). The vowel /a/ is pronounced with the mouth wide open and the tongue positioned low in the mouth. In contrast, the vowel /i/ is pronounced with the mouth slightly open, and the tongue positioned high in the mouth. Since the oral cavity is larger when pronouncing /a/ than when pronouncing /i/, the vowel /a/ is likely to elicit bigger images than the vowel /i/.

In the present study, we attempted to investigate the vowel-size relationship in a way that distinguished between these two factors. More specifically, in Experiment 1, we examined whether participants would associate the vowels /a/ and /i/ with bigger and smaller images, respectively, even when they were not motivated to pronounce the vowels during the task. In Experiment 2, we examined whether kinesthetic experience around the mouth (i.e., proprioception of the size of the oral cavity when pronouncing the vowels) on its own and without auditory experience could elicit bigger/smaller images.

To examine these problems, we used speeded classification tasks, which have been widely used in studies of cross-modal perception (e.g., Gallace & Spence, 2006). In this kind of task, participants have to discriminate between stimuli in one dimension while trying to ignore an irrelevant dimension, which enables us to see whether their response to the relevant dimension is influenced by the variation of the irrelevant dimension.

In Experiment 1, to investigate whether the vowels /a/ and /i/ elicit bigger/smaller images without the kinesthetic experience of pronunciation, we asked participants to judge the relative size of the target disk, while an irrelevant sound (the vowel /a/ or /i/) was presented simultaneously. If reaction times for judging the size of the target disk were influenced by the variation of the vowels, this would allow us to conclude that the acoustical features of vowels /a/ and /i/ elicit bigger/smaller images without kinesthetic experience.

In Experiment 2, to investigate whether the proprioception of the size of the oral cavity when pronouncing /a/ and /i/ could elicit bigger/smaller images on its own without the subject actually hearing any vowel sounds, we asked participants to judge the relative size of the target disk, while ensuring that they opened their mouths in the same way they would if pronouncing each vowel. If reaction times for judging the size of the target disk were influenced by the variation in the way the participants opened their mouths, this would allow us to conclude that the kinesthetic experience alone elicits bigger/smaller images without auditory experience.

In the following, we will discuss the possible process underpinning the vowel-size relationship, taking both the above factors into account.

Experiment 1

In Experiment 1, we attempted to investigate whether the vowels /a/ and /i/ elicit bigger/smaller images without the kinesthetic experience of pronunciation. In the experiment, participants were asked to judge whether a target disk was bigger or smaller than a standard disk. The target disk was presented following the standard disk. It was 10% or 20% shorter or longer in diameter compared to the standard disk. A task-irrelevant sound (/a/ or /i/) was sometimes presented simultaneously along with the presentation of the target disk. If it is the case that the vowel sounds (/a/ and /i/) elicit bigger/smaller images without the kinesthetic experience of pronunciation, the reaction times should have been shorter when the vowel-size relation is congruent (/a/ being presented when the target disk was bigger and /i/ being presented when it was smaller) than when it was incongruent.

Method

Participants Thirty Japanese-speaking undergraduate students (14 males, 16 females; mean age, 22.2 years; range 20-36 years) took part in the experiment.

Apparatus The visual stimuli were presented on a laptop computer (Dell Inspiron 1526) with a 15.4-inch screen, or on a desktop computer (VAIO VGC-RA72P) with a 17-inch screen. Auditory stimuli were presented through headphones (Audio-Technica ATH-ANC7 or Sehnheiser HDA200). The presentation of the stimuli and the recording of the participants' responses were controlled using Cedrus Superlab 4.0 software.

Materials The visual stimuli were the standard disk, the target disks, and the mask. The standard disk was gray and 3 cm in diameter, and the target disks were ±10% and ±20% of the diameter of the standard disk. The visual mask was a light-gray screen with dark-gray spray. Four different auditory stimuli were used for presentation of the vowels /a/ and /i/, respectively. The auditory stimuli were a recording of a Japanese female who had been asked to pronounce Japanese vowels. Her speech was recorded on a Roland R-09. The duration of each vowel was 300 ms. For the vowel /a/, the mean fundamental frequency was 240.3 Hz (*SD* =2.16 Hz), the mean first formant frequency was 780.3Hz (*SD* =46.1 Hz), the mean second formant frequency was 1374.8Hz (*SD* =57.6 Hz), the mean third formant frequency was 3077.5 Hz (*SD* =241.7 Hz), and the mean intensity was 48.79 dB (*SD* =2.48 dB). For the vowel /i/, the mean fundamental frequency was 240.6 Hz (*SD* =3.59 Hz), the mean first formant frequency was 416.5 Hz (*SD* =20.9 Hz), the mean second formant frequency was 2712.3Hz (*SD* =67.6 Hz), the mean third formant frequency was 3494.0 Hz (*SD* =75.3 Hz), and the mean intensity was 49.94 dB (*SD* =2.43 dB).

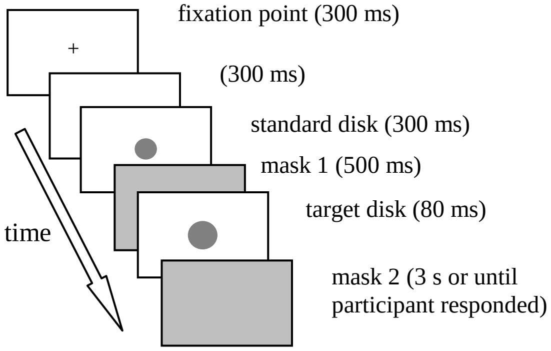

Figure 1: Illustration of the sequence of visual stimuli presented in each trial in Experiment 1 and 2.

Procedure The participants sat at a desk, 45 cm from the computer. It took about 10 minutes to complete the entire experiment.

Figure 1 illustrates the sequence of events in each trial. At the start of each trial, the word "Ready?" appeared at the center of the screen, and the participants could choose when to start by pressing the space key. At first, a fixation point was presented in the middle of the screen for 300 ms, followed by a blank white screen. After a 300-ms presentation of the blank screen, the standard disk was presented at the center for 300 ms, followed by the mask screen. The mask screen was presented for 500 ms and was followed by the target disk. The position of the target disk varied randomly (by up to ±0.3 cm vertically and horizontally from the center of the screen) to prevent the participants from using superimposition cues to judge the relative size of the target disk. At the same time the target disk was presented, a vowel (/a/ or /i/) was presented in 20 trials for each vowel, and no sound was presented in the remaining 20 trials. The target disk was presented for 80 ms, followed by the mask screen. The mask screen stayed on the screen until the participant responded or until 3 seconds had elapsed, at which point the screen displaying the word "Ready?" appeared and the next trial was ready to begin.

The participants were asked to judge whether the target disk was bigger or smaller than the standard disk as rapidly as possible. The participants were instructed to indicate the relative size of the target disk by pushing "/" with the index finger of the right hand, or "\" with the middle finger of the right hand. Which key corresponded to "big" or "small" was counterbalanced across the participants.

Table 1: The means and standard errors of reaction times (in milliseconds) as a function of condition and size of the target disk in Experiment 1 and 2.

	Condition					
	congruent		incongruent		control	
	RT		RT		RT	
size	*M*	*SE*	*M*	*SE*	*M*	*SE*
	Experiment 1					
+10%	408.7	29.6	423.9	25.8	397.9	23.7
+20%	375.6	18.2	410.2	20.2	381.3	21.9
-10%	469.8	29.4	492.8	41.0	482.3	36.6
-20%	387.5	18.6	423.1	26.7	397.5	21.6
mean	410.4	21.8	437.5	26.0	414.7	23.4
	Experiment 2					
+10%	378.5	18.0	387.3	18.5	385.1	15.0
+20%	362.6	12.2	369.8	19.8	367.0	12.2
-10%	427.4	20.8	411.8	16.8	421.2	22.0
-20%	376.7	18.5	367.7	16.0	370.5	15.1
mean	386.3	15.5	384.1	15.2	385.9	14.6

The participants were informed that a task-irrelevant sound would sometimes be presented, but they were instructed to ignore it. The response times were calculated from the beginning of the second mask screen to the time of the decision. The participants completed 12 practice trials before the experiment to ensure that they clearly understood the task.

The experiment was composed of 60 trials, 15 trials for each size of the target disk (±10%, ±20%). The order of trials was randomized for each participant. For each size of the target disk, five trials were presented with the vowel /a/, five trials were presented with /i/, and the remaining five trials were presented with no sound. Each of the trials was classified into three conditions, with 20 trials each: congruent condition (i.e., /a/ being presented when the target disk is bigger and /i/ being presented when it is smaller), incongruent condition (i.e., the opposite combination to the congruent condition), and control condition (i.e., no sound being presented along with the target disk).

Results

The means and standard errors of reaction times as a function of the condition and size of the target disk are shown in Table 1. Because the error rate was quite low (*M* = 3.1%, *SD* = 2.9%), subsequent analysis was only performed on the reaction times.

Reaction times The reaction times for the wrong decision (3.1%) and above +3SD from the mean reaction times of each participant (1.2%) were excluded from the analysis.

We performed repeated measures of variance on the reaction times as a function of condition (3) and size of the target disk (4). The analysis revealed a significant main effect of condition (*F*(2,58) = 7.08, *p* = .002), and a significant main effect of size of the target disk (*F*(3,87) = 10.89, *p* = .001),[1] but no significant interaction between condition and size (*F* (6,174) =.484, *p* >.10).[1]

A post hoc Bonferroni test of condition revealed significant differences between the congruent condition and the incongruent condition and between the control condition and the incongruent condition (all *p*s <.05), with the slowest responses occurring in the incongruent condition. There was no significant difference between the congruent condition and the control condition. A post hoc Bonferroni test of size revealed significant differences between -10% and other sizes (all *p*s < .05), with the slowest responses occurring in -10%. There were no significant differences between any pair of the remaining sizes (+10%, +20%, -20%).

In sum, reaction times were longer in the incongruent condition than in the congruent condition or control condition, and in -10% than in the other sizes.

Discussion

In Experiment 1, we attempted to investigate whether participants would associate the vowels /a/ and /i/ with bigger and smaller images, respectively, without the kinesthetic experience of pronunciation, using a speeded classification task. The results indicate that they did. The participants responded more slowly in the incongruent condition than in the congruent condition or control condition. The vowel /a/ elicited bigger images and the vowel /i/ elicited smaller images without kinesthetic experience, which could interfere with the response of "big"

[1] A Greenhouse-Geisser adjustment was used to correct for violations of sphericity.

while hearing /i/ and with the response of “small” while hearing /a/. As Tarte (1982) pointed out, the component formant frequencies of vowels can explain the vowel-size relationship.

Experiment 2

In Experiment 2, we attempted to investigate whether bigger/smaller images would be elicited only with the kinesthetic experience around the mouth when pronouncing the vowels /a/ and /i/ (i.e., the proprioception of the size of the oral cavity), without the subject actually hearing any vowel sounds, using the same task as Experiment 1. In the experiment, the participants completed the same speeded classification task with the kinesthetic experience of pronouncing vowels. We ensured that the participants opened their mouths in the same way they would if pronouncing each vowel by asking them to hold either of two types of solid object in their teeth: one was egg-shaped and the other was board-shaped. In order to hold the egg-shaped object with their teeth, participants had to open their mouth widely, and the resultant lip shape was similar to that when pronouncing the vowel /a/. On the other hand, holding the board-shaped object required participants to open their mouth slightly along the vertical axis and pull their lips sideways. This shape mimicked the lip shape when pronouncing the vowel /i/. If it is the case that the proprioception of the size of oral cavity elicits images of size without auditory experience, the reaction times should have been shorter when the participants were holding the egg-shaped object and the larger target disk was presented, and they are holding the board-shaped object and the smaller target disk was presented, compared with the opposite combinations.

Method

Participants Twenty-four Japanese-speaking adults (13 males, 11 females; mean age, 26.8 years; range 22-42 years) took part in the experiment.

Apparatus and Materials The visual stimuli were presented on a laptop computer (Dell Inspiron 1526) with a 15.4-inch screen, controlled by Cedrus SuperLab 4.0. The visual stimuli were the same as Experiment 1. Two solid objects (egg-shaped and board-shaped) made from styrofoam were used to ensure the participants opened their mouths in the same way they would if pronouncing each vowel. The egg-shaped object was 5.5 cm in maximum diameter and 8 cm long, and the board-shaped object was 7.5 cm by 15 cm long and 0.5 cm thick. Twenty-four sets of the two objects were prepared so that each participant could use a new one.

Procedure As in Experiment 1, participants sat at a desk, and the experimenter instructed them to indicate the relative size of the target disk as soon as possible by pressing the keys. The sequence of the visual events in each trial was the same as Experiment 1.

The participants completed 12 practice trials before the experiment. The experiment was composed of six blocks of 72 trials, with a short break at the end of each block. Each block had 12 trials, three trials for each size of the target disk (±10%, ±20%), and the order of the trials was randomized in each block for each participant. Six blocks were divided into three phases, which had two blocks each. In one phase, participants were instructed to open their mouth naturally, and hold the smaller side of the egg-shaped object in their teeth. In the other phase, participants were instructed to open their mouth slightly sideways, and hold the longer side of the board-shaped object in their teeth. In the remaining phase, participants were instructed to complete the task in the same way as the practice trials, i.e., to hold no object in their mouth. The order of the three phases was counterbalanced across participants. Each of the trials was classified into three conditions, 24 trials for each condition: congruent condition (i.e., the participants are holding the egg-shaped object and the larger target disk is presented, or they are holding the board-shaped object and the smaller target disk is presented, incongruent condition (i.e., the opposite combination to the congruent condition), or control condition (i.e., the participants are not holding anything when the target disk is presented).

Results

The means and standard errors of reaction times and number of wrong decisions as a function of condition and size of the target disk are shown in Table 1. As in Experiment 1, because the error rate was quite low (*M* = 3.1%, *SD* = 2.9%), subsequent analysis was performed only on the reaction times.

Reaction times As in Experiment 1, the reaction times for the wrong decision (3.1%) and above +3SD from the mean reaction times of each participant (1.2%) were excluded from the analysis.

We performed repeated measures of variance on the reaction times as a function of condition (3) and size of the target disk (4). The analysis revealed a significant main effect of size of the target disk ($F(3,69) = 14.8$, $p < .001$),[1] but no significant main effect of condition ($F(2,46) = .05$, $p > .10$),[1] and no significant interaction between condition and size ($F(6,138) = .438$, $p > .10$),[1] A post hoc Bonferroni test of size revealed significant differences between -10% and the other sizes (all *p*s < .01) with the slowest responses occurring in -10%, and a marginally significant difference between +10% and +20% with slower responses in +10% ($p = .08$).

These results indicate that the condition did not affect the reaction times, although the size of the target disk affected them as in Experiment 1.

Discussion

In Experiment 2, we attempted to investigate whether the proprioception of the size of the oral cavity could elicit bigger/smaller images on its own without the subject actually hearing any vowel sounds, using the same task as Experiment 1. The reaction times did not differ significantly between in the congruent condition and in the incongruent condition. The results indicate that the proprioception of the size of oral cavity when pronouncing /a/ and /i/ may not, on its own, elicit the image of bigger/smaller sizes. However, it should be pointed out that in this experiment we controlled the absolute size of the oral cavity, in other words, we investigated the effect of the *static* kinesthetic experience of pronunciation. It is possible that the *dynamic* kinesthetic experience of pronunciation, that is, the temporal change of the relative size of the mouth, plays an important role in eliciting the image of bigger/smaller sizes.

It is also worth noting that the lack of uncertainty about the variation of stimuli in the irrelevant dimension may have weakened the effect of treatment (Gallace & Spence, 2006). In Experiment 1, there was an uncertainty about the variation of stimuli in the irrelevant dimension, induced by trial-by-trial variation. In contrast, in Experiment 2, the stimuli in the irrelevant dimension were fixed during each of the blocks.

In sum, the results in Experiment 2 indicate that the static kinesthetic experience (i.e., the proprioception of the absolute size of oral cavity) may not contribute to the vowel-size relationship, although it is possible that the dynamic kinesthetic experience could contribute to it. In addition, the lack of uncertainty about the variation of the irrelevant dimension may have weakened the effect of treatment.

General Discussion

In the present study, we attempted to re-examine the vowel-size relationship in a way that distinguished between two possible factors, formant frequencies of the vowels (Experiment 1) and kinesthetic experience while pronouncing the vowels (Experiment 2), using the speeded classification paradigm.

The results of Experiment 1 indicate that the component formant frequencies of vowels on their own can explain the vowel-size relationship, and the results of Experiment 2 indicate that the static kinesthetic experience (proprioception of the absolute size of oral cavity) may not contribute to the vowel-size relationship.

However, in the results of Experiment 2, the possibility remains that the dynamic kinesthetic experience (the temporal change of the relative size of the mouth) might have elicited bigger/smaller images and had an influence on the results. Furthermore, we cannot completely eliminate the possibility that the dynamic kinesthetic experience may have affected the results of Experiment 1 from the viewpoint of motor theory (e.g., Liberman & Mattingly, 1985), which understands the perception of speech as vocal tract gestures. From this viewpoint, the dynamic kinesthetic experience automatically generated from hearing vowels may have affected the judgments of size, and supported the results of Experiment 1.

Taking the above into account, the vowel-size relationship can be mainly explained by the component formant frequencies and the static kinesthetic experience may not contribute to it, but the dynamic kinesthetic experience of pronunciation may play some role. Further research is needed to evaluate the role of component formant frequencies more exactly by controlling the kinesthetic experience more rigidly, and to investigate the role of the dynamic kinesthetic experience of pronunciation.

References

Gallace, A., & Spence, C. (2006). Multisensory synthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*(7), 1191-1203.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Newman, S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology, 45,* 53-75.

Tarte, R. D., & Baritt, L. S. (1971). Phonetic symbolism in adult native speakers of English: Three studies. *Language and Speech, 14,* 158-168.

Tarte, R.D. (1982). The relationship between monosyllables and pure tones: an investigation of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior, 21*, 352-360.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225-239.