BRILL

Multisensory Research 33 (2020) 569–598

MULTISENSORY RESEARCH

brill.com/msr

Cross-Modal Correspondence Between Speech Sound and Visual Shape Influencing Perceptual Representation of Shape: the Role of Articulation and Pitch

Yuna Kwak [1], **Hosung Nam** [2,3,*], **Hyun-Woong Kim** [1] **and Chai-Youn Kim** [1,*]

[1] Department of Psychology, Korea University, Seoul 02841, Korea
[2] Department of English Language and Literature, Korea University, Seoul 02841, Korea
[3] Haskins Laboratories, New Haven, CT 06511, USA

Received 14 June 2018; accepted 21 October 2019

Abstract

Cross-modal correspondence is the tendency to systematically map stimulus features across sensory modalities. The current study explored cross-modal correspondence between speech sound and shape (Experiment 1), and whether such association can influence shape representation (Experiment 2). For the purpose of closely examining the role of the two factors — articulation and pitch — combined in speech acoustics, we generated two sets of 25 vowel stimuli — pitch-varying and pitch-constant sets. Both sets were generated by manipulating articulation — frontness and height of the tongue body's positions — but differed in terms of whether pitch varied among the sounds within the same set. In Experiment 1, participants made a forced choice between a round and a spiky shape to indicate the shape better associated with each sound. Results showed that shape choice was modulated according to both articulation and pitch, and we therefore concluded that both factors play significant roles in sound–shape correspondence. In Experiment 2, participants reported their subjective experience of shape accompanied by vowel sounds by adjusting an ambiguous shape in the response display. We found that sound–shape correspondence exerts an effect on shape representation by modulating audiovisual interaction, but only in the case of pitch-varying sounds. Therefore, pitch information within vowel acoustics plays the leading role in sound–shape correspondence influencing shape representation. Taken together, our results suggest the importance of teasing apart the roles of articulation and pitch for understanding sound–shape correspondence.

Keywords

Cross-modal correspondence, bouba–kiki effect, audiovisual interaction, speech sound, vowel sound, articulation, pitch, visual shape

* To whom correspondence should be addressed. E-mails: hnam@korea.ac.kr/chaikim@korea.ac.kr

DOI:10.1163/22134808-20191330

1. Introduction

Converging inputs from our sensory systems are useful for processing external events. The inputs can sometimes provide information on a single object or property, for example when visual and haptic sensory cues provide redundant estimates for shape of the same object (Chen and Spence, 2017; Deroy and Spence, 2016; Ernst and Banks, 2002; Schutz and Kubovy, 2009; Welch, 1972). On the other hand, inputs from different sensory modalities can be converging in the sense that they provide 'corresponding' information, the information not necessarily stemming from the same source and therefore seemingly unrelated. In fact, there has been growing interest in research into 'cross-modal correspondence', a term referring to our brain's tendency of mapping certain features across the senses in a systematic and non-random manner (Marks, 2004; Spence, 2011). For instance, people tend to associate high-pitched sound with light colors rather than dark colors, although pitch itself does not convey lightness/darkness (Hubbard, 1996; Ludwig *et al*., 2011; Marks, 1974, 1987, 1989). Cases of cross-modal correspondence are now reported between many stimulus features. For example, auditory pitch is associated with a variety of stimulus features including visual elevation (Ben-Artzi and Marks, 1995; Bernstein and Edelstein, 1971; Evans and Treisman, 2010; Jamal *et al*., 2017; Melara and O'Brien, 1987), tactile elevation (Occelli *et al*., 2009), visual size (Bien *et al*., 2012; Evans and Treisman, 2010; Gallace and Spence, 2006; Parise and Spence, 2009, 2012; Peña *et al*., 2011), and visual shape (Marks, 1987; Walker *et al*., 2010). Relatively recent studies have also found that vowel sounds are systematically related to visual features such as color and lightness (Kim *et al*., 2018; Moos *et al*., 2014). In addition, tastes/odors are known to be associated with visual shape (Deroy and Valentin, 2011; Ngo *et al*., 2013; Spence and Gallace, 2011) and auditory pitch (Belkin *et al*., 1997; Crisinel and Spence, 2010, 2012).

In the current study, we focus on one of the most representative examples of audiovisual correspondence: the 'maluma–takete' effect documented by Köhler (1929, 1947) later renamed as the 'bouba–kiki' effect (Ramachandran and Hubbard, 2001). This effect refers to the consistency observed in matching the sounds of nonsense words to particular visual shapes: people tend to relate a round shape with the sounds 'maluma'/'bouba' and a spiky shape with the sounds 'takete'/'kiki'. Since Köhler's initial finding, similar patterns have been replicated not only across several languages and cultures (Ahlner and Zlatev, 2010; Bremner *et al*., 2013; Davis, 1961; Tarte, 1974), but also across ages including infants (Maurer *et al*., 2006; Ozturk *et al*., 2013).

Given the robustness of this effect, also referred to as sound–shape correspondence, research has investigated which particular properties in speech sounds are associated with certain shapes. According to previous studies,

speech articulation is the driving force behind the effect (D'Onofrio, 2014; Fort *et al*., 2015). D'Onofrio (2014), for example, showed that vowel frontness, consonant voicing, and consonant place of articulation modulate the associated visual shape. However, not only articulation, but also auditory pitch — an acoustic property that has been frequently studied in cross-modal correspondence literature — possibly leads to the association between speech sound and shape. This is because auditory pitch itself is known to be associated with visual shape (Evans and Treisman, 2010; Marks, 1987). As both articulation and pitch have been reported to bear a relationship with visual shape, and because both of these factors are combined in speech sounds (e.g., low pitch sounds such as /a/ in 'bouba' are associated with a round shape, whereas high pitch sounds such as /i/ in 'kiki' are associated with a spiky shape), which one of them is driving the well-known 'maluma–takete'/'bouba–kiki' effect remains elusive. Particularly in the case of vowel sounds, change in articulation is accompanied by change in auditory pitch; therefore, it is difficult to manipulate these factors separately. It may be possible that sound–shape correspondence is entirely driven by one of these factors, but it is also plausible that both of them play an important role in the association.

Another question yet to be answered in sound–shape correspondence literature is whether the correspondence can exert an influence beyond the conceptual level and affect one's subjective experience of shape. In fact, this question has been examined for other types of cross-modal correspondences. For example, in line with the view that perceptual representation is modulated by pitch–location correspondence, Maeda *et al*. (2004) demonstrated that varying auditory pitch information systematically biases the percept of visual motion direction. However, other studies have shown that cross-modal correspondence does not change one's representation (Hidaka *et al*., 2013; Marks *et al*., 2003), leaving this issue a point of contention within the cross-modal correspondence literature. For instance, Gallace and Spence (2006) found that concurrent auditory information (i.e., high- or low-frequency tone) did not affect the size representation of the visual stimulus. Whether sound–shape correspondence can modulate the shape representation is yet to be established (but see Hung *et al*., 2017).

In the current study, we aimed to examine the two aspects of sound–shape correspondence introduced above: the relative contribution of features combined in speech sounds to sound–shape correspondence (Experiment 1) and the influence of sound–shape correspondence to the perceptual representation of shape (Experiment 2).

In Experiment 1, participants listened to speech sounds and made a forced choice between a round and a spiky shape to indicate the shape more associated with each sound. For the auditory stimuli, an articulatory synthesizer was employed to create simple and short vowel sounds. For the following reasons,

the synthetic vowel sounds served as the appropriate stimulus set for observing the effects of both articulation and pitch — the two factors combined in speech acoustics. First, we could systematically examine the relationship between vowel articulation and shape choice by using a synthesizer to equidistantly manipulate the two articulatory dimensions known to determine vowel acoustics — 'frontness' and 'height' of the tongue body's position. Based on the principles used to create the stimuli, we could observe the effects of height and frontness on the matching pattern between sound and shape. Second, the current stimuli allowed us to test the role of pitch in sound–shape correspondence, separately from vowel articulation. Pitch is an acoustic feature that systematically varies with the tongue body's vertical position — the height dimension — and therefore cannot be easily dissociated from vowel articulation. To touch upon this issue in the current study, we generated two sets of vowel sounds — the pitch-varying (Experiment 1a) and the pitch-constant (Experiment 1b) — and compared the shape choice for the two stimulus sets.

Experiment 2 was designed to explore whether visual shape representation would change depending on the sound information accompanied, based on sound–shape correspondence. As in Experiment 1, we aimed to closely examine the roles of articulation and pitch; and to do so, we conducted three sub-experiments with pitch-varying (Experiment 2a) and pitch-constant (Experiments 2b and 2c) vowel sounds. In all three experiments, an ambiguous visual shape — a neutral morph between a round and a spiky shape — was briefly presented with task-irrelevant vowel sounds. Participants had to adjust the curvature of a random shape in the response display to render it identical to the shape presented with the sound. We hypothesized that if the contents of subjective experience of shape is being influenced by sound–shape correspondence, the visual shape representation should change depending on the sound information. For example, when presented with the vowel sound associated with round shape, the ambiguous visual shape should be represented as being rounder. Comparing the results of the sub-experiments, in which the presented vowel stimulus sets were manipulated in terms of either articulation or pitch, or both, allowed us to explore the relative contribution of the two factors in sound–shape correspondence affecting visual representation.

2. Experiment 1

For the purpose of closely examining the pattern of correspondence between speech sound and shape, we conducted a matching task in Experiment 1. Participants listened to vowel sounds and made a forced choice to indicate the shape better associated with each sound. In Experiment 1a, we presented pitch-varying vowel sounds, among which pitch systematically differed along the height dimension such that sounds with different height possessed different

pitch values. In Experiment 1b, on the other hand, the pitch-constant vowel set of stimuli was utilized, and pitch was set identical among the sounds within the set. Utilizing both stimulus sets allowed us to examine how the articulatory dimensions — frontness and height — interact with pitch in sound–shape correspondence.

2.1. Methods

2.1.1. Participants

Forty individuals (16 males and 24 females, 20–29 years of age) participated in Experiment 1a, and a separate group of 40 participants (16 males and 24 females, 19–39 years of age) ran in Experiment 1b. They gave informed consent approved by the Korea University Institutional Review Board (KU-IRB-17-85-A-1) before participating in the study. All participants had normal or corrected-to-normal vision and hearing. Their native language was Korean.

2.1.2. Stimuli

Visual Stimuli. A round–piky shape pair, which subtended 12.9° × 12.9°, was presented against a black background (see Note 1).

Auditory Stimuli: Pitch-Varying and Pitch-Constant Vowels. In this study, we employed the Haskins laboratories Configurable Articulatory SYnthesis model (CASY; Rubin *et al*., 1996) to generate synthetic vowel-like sounds (Fig. 1; Kim *et al*., 2018). CASY builds upon Mermelstein's articulatory model, in which the model variables specify the spatial position of the articulators (lips, jaw, tongue body, tongue tip, and etc.) in terms of distance and angle, thereby determining the shape of the vocal tract in the mid-sagittal plane (Mermelstein, 1973). The temporal variation of this vocal tract shape is equivalent to a dynamic sequence of speech events.

The vowel stimuli in Experiments 1a and 1b were synthesized by parametrically manipulating the tongue body's center position, which is given by the tongue body articulator variables (Fig. 1; TCL, TCA). All other articulator variables were fixed. We created a set of 25 points of tongue body's center (TC) positions, each point with varying horizontal (i.e., frontness) and vertical (i.e., height) coordinates in the vocal tract space. The distance between all neighboring points was 2.8 mm. Input pitch and duration of vowels were set to 120 Hz and 500 ms, respectively. In Experiment 1a, there were systematic variations in output pitch among sounds generated from different tongue body positions (range: 120.5–131.6 Hz), which reflect the acoustics of vowels resulting from the actual articulatory process (Kim *et al*., 2018; Whalen and Levitt, 1995). For the pitch-constant stimulus set in Experiment 1b, we used Praat software (Boersma and Weenink, 2013) to set the pitch of all vowel sounds at an identical value of 110 Hz (for demonstration of the stimuli sets

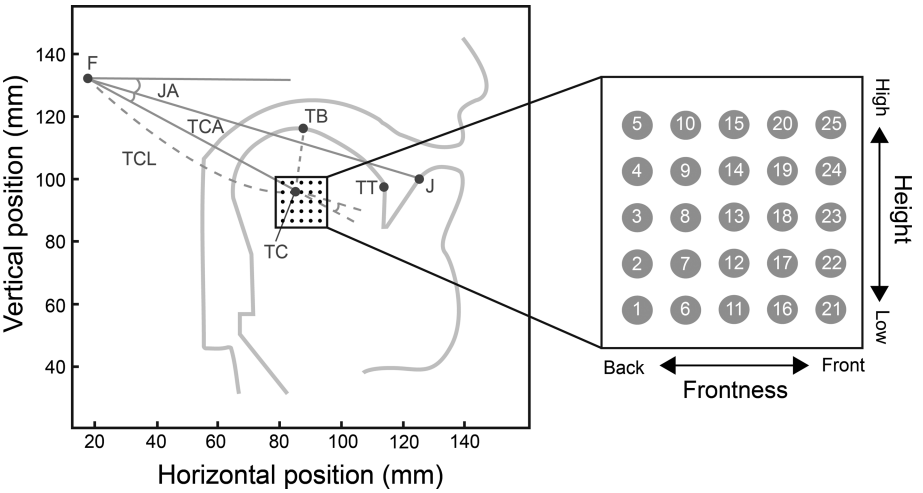

Figure 1. Vowel stimuli presented in Experiment 1. Representation of the vocal tract in CASY with the model's articulatory variables for generating both the pitch-varying (Experiment 1a) and the pitch-constant (Experiment 1b) vowel sounds. F: mandibular condyle, TC: tongue body's center, TB: tongue blade, TT: tongue tip, J: jaw. The position of the tongue body's center (TC) is given by the center of an imaginary circle with a fixed radius representing the tongue body. The center is determined by the angle (TCA) between the line F–J and the line F–TC (TCL). For generating the vowel sounds, the tongue body's articulatory variables (TCA and TCL) were parametrically manipulated to modulate the horizontal and vertical position of the tongue body's center (TC). The tongue blade (TB) and tip (TT) are attached to the tongue body's center. All the other articulatory variables were fixed. The positions of the tongue body's center for generating the 25 stimuli are overlayed on the representation of the vocal tract; a larger version of the positions with indices is shown on the right. The terms 'frontness' and 'height' refer to the horizontal and vertical position of the tongue body's center point in space (for more information, see Kim *et al*., 2018).

used in Experiments 1a and 1b, see Supplementary Movies S1 and S2). The effectiveness of the physical stimulus manipulation in terms of the frontness and height dimensions were verified using multidimensional scaling (for more information, see the Supplementary Text and Supplementary Fig. S1).

2.1.3. Apparatus

Visual stimuli were presented on a 19-inch CRT monitor (1024 × 768 resolution, 60 Hz refresh rate; viewing distance 52 cm). All auditory stimuli were presented through SRH440 headphones. Experiments were conducted in a quiet, dark room using MATLAB version 9.1 (The Mathworks, Inc., Natick, MA, USA) and Psychophysics Toolbox version 3 (Brainard, 1997; Pelli, 1997).

2.1.4. Procedures

In both Experiment 1a and 1b, participants made a two-alternative forced-choice judgment as to the visual shape that was associated with each vowel sound. On each trial, a round and a spiky shape appeared on the left and right sides of the screen (11.018° from the central fixation), and the position of each shape was randomly chosen every trial. The onset of the vowel sound was synchronous with that of the visual display. The visual stimuli were presented on the monitor screen until participants indicated the shape that better matched the sound on the trial with a button press, and there was no limit to the response time. In each experiment, there were 20 repetitions of 25 vowel sounds, resulting in a total of 500 trials. The order of the trials was randomized for each participant.

2.2. Results

Figure 2 shows the mean frequency of choosing the spiky shape for each level of frontness and height of the tongue body's positions when participants listened to pitch-varying (Experiment 1a) and pitch-constant (Experiment 1b) vowel sounds. For the results of both experiments, a two-way ANOVA was conducted with frontness and height as within-participant factors.

2.2.1. Experiment 1a: Pitch-Varying Vowels

Results are shown in Fig. 2A. Both the main effects of frontness and height on shape choice were statistically significant ($F_{(1.477,57.621)} = 20.866$, $p < 0.001$, $\eta^2 = 0.349$; $F_{(1.265,49.324)} = 8.887$, $p < 0.01$, $\eta^2 = 0.186$; both Greenhouse–Geisser corrected). The interaction effect between frontness and height was also significant ($F_{(5.822,227.063)} = 2.728$, $p < 0.05$, $\eta^2 = 0.065$, Greenhouse–Geisser corrected).

In an attempt to determine which means are significantly different from each other, a series of *post-hoc* paired-sample t-tests were conducted. We report only the results of the pairwise comparisons between the extreme levels of each factor (see Fig. 1 for index of vowel stimuli): 1 and 21 [low back vowel and low front vowel; $t(39) = 3.993$, $p < 0.01$, Cohen's $d = 0.753$], 5 and 25 [high back vowel and high front vowel; $t(39) = 4.772$, $p < 0.001$, Cohen's $d = 0.893$], 1 and 5 [low back vowel and high back vowel; $t(39) = 2.791$, $p < 0.05$, Cohen's $d = 0.712$], 21 and 25 [low front vowel and high front vowel; $t(39) = 3.660$, $p < 0.01$, Cohen's $d = 0.846$] [all p values are false discovery rate (FDR)-corrected].

2.2.2. Experiment 1b: Pitch-Constant Vowels

Results are shown in Figure 2B. We observed a significant main effect of frontness on shape choice ($F_{(1.442,56.239)} = 10.521$, $p < 0.01$, $\eta^2 = 0.212$, Greenhouse–Geisser corrected), but not the effect of height ($F_{(1.368,53.359)} =$

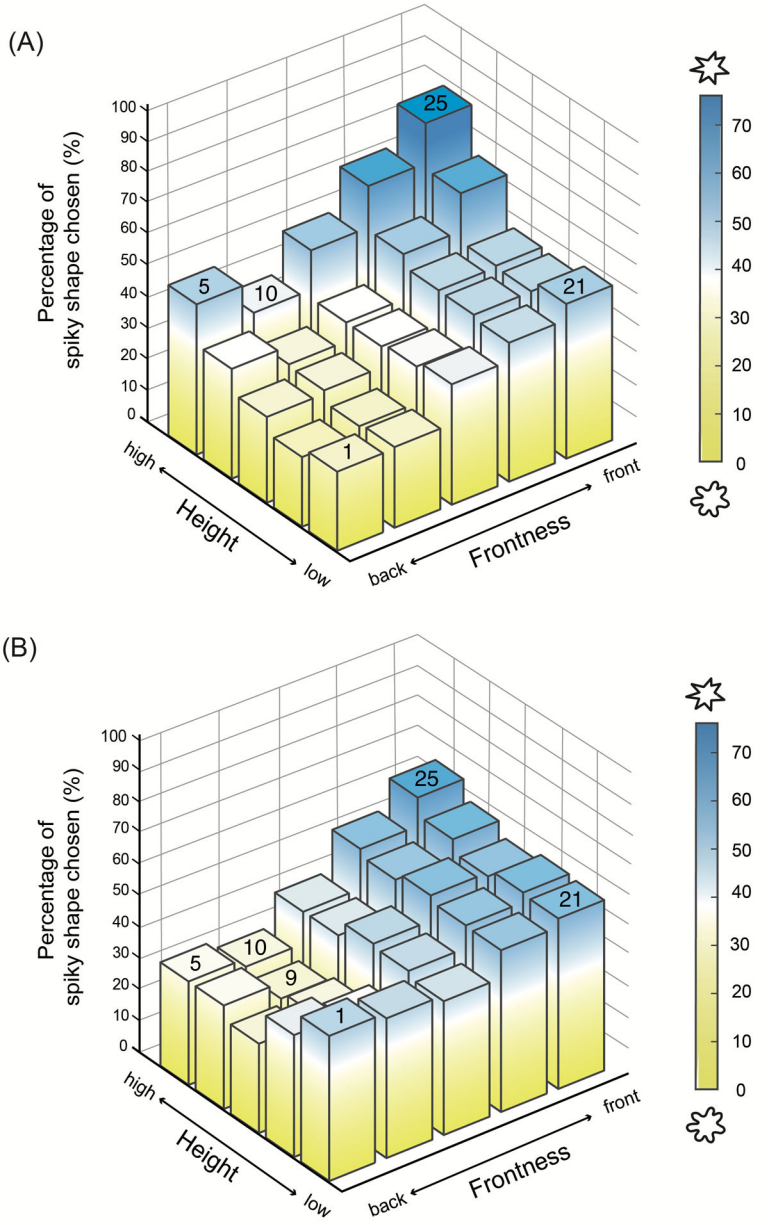

Figure 2. Shape-matching results of Experiments 1a (A) and 1b (B). The *y*-axis shows the percentage of spiky shape choice for 25 vowel sounds. A more yellowish color denotes a higher percentage of round shape choice, whereas a more bluish color denotes a higher percentage of spiky shape choice. The numbers shown on some of the bars indicate the indices of sounds that will be used in Experiment 2.

0.835, $p = 0.399$, Greenhouse–Geisser corrected). The frontness–height interaction was significant ($F_{(4.997,194.892)} = 2.895$, $p < 0.05$, $\eta^2 = 0.069$, Greenhouse–Geisser corrected).

As in Experiment 1a, a series of *post-hoc* paired-sample t-tests were conducted to examine which means are significantly different from each other. When conducting the comparisons between the extreme levels of each dimension, the difference between 5 and 25 (see Fig. 1 for index of vowel stimuli) reached statistical significance [$t(39) = 4.999$, $p < 0.001$, Cohen's $d = 0.895$; FDR-corrected].

2.3. Discussion

Experiment 1 demonstrates that both articulation and auditory pitch are the driving forces behind the sound–shape correspondence effect. In Experiment 1a, the choice between the round and the spiky shape was modulated according to the two articulatory dimensions manipulated to generate vowel acoustics — frontness and height. When the pitch-constant stimuli were presented in Experiment 1b, the frontness dimension still exhibited influence on shape choice; this indicates that articulation plays a role in sound–shape correspondence, even when there is no pitch difference among the vowel sounds. However, that the height dimension had no effect shows that the systematic variation in pitch along the height dimension is also contributing significantly to the association between vowel sound and shape.

The current study indicates that articulatory dimensions determining vowel acoustics modulate the choice between round and spiky shape, leading to well-known phenomena such as the 'bouba–kiki' effect. In particular, manipulating the frontness of the tongue body's position significantly influenced shape choice in both Experiments 1a and 1b, consistently with previous literature (D'Onofrio, 2014; Maurer *et al.*, 2006; McCormick *et al.*, 2015; Spector and Maurer, 2013). Most studies in fact have categorized vowels into front and back vowels and have focused on testing the effect of the frontness dimension. For example, Spector and Maurer (2013) showed that toddlers match front vowels to spiky shape and back vowels to round shape, by presenting the vowels /i/ and /o/. In addition, D'Onofrio (2014) replicated sound–shape correspondence by using a wider range of vowel stimuli and classifying them into front and back vowels to examine the role of vowel frontness.

In terms of the height dimension, Experiments 1a and 1b showed different patterns of results. The main effect of height on shape choice was present only for the pitch-varying stimulus set (Experiment 1a), in which vowel pitch varied systematically with different levels of height. This can be interpreted in the context of previous studies demonstrating pitch–shape correspondence; high pitches are associated with spiky shapes whereas low pitches are associated with round shapes (Marks, 1987; Parise and Spence, 2009; Walker *et*

al., 2010). To elaborate, pitch is an acoustic feature that constitutes vowels (Whalen and Levitt, 1995) and cannot be separated from vowel height; the low vowel /a/ intrinsically has lower pitch compared to the high vowel /i/. By comparing the effect of height on shape choice for pitch-varying and pitch-constant sounds in the current study, we conclude that the pitch–shape correspondence contributes to the association between vowel sound and shape. Vowel height, along with vowel frontness, is an articulatory dimension that determines sound–shape correspondence, but the inherent variation in pitch of vowels with different levels of height seems to be the driving force behind it.

It is also worth noting how the use of speech sounds generated with an articulatory synthesizer can address potential issues arising from stimuli used in previous studies. While most studies have used a limited number of stimuli and have coupled them into congruent and incongruent pairs (e.g., /a/ and /i/ with round and spiky shape), here we broke the dimensions of vowel acoustics into many levels and generated numerous cross-modal pairs that were neither congruent nor incongruent. Such a manipulation can prevent undesired range effects: effects found using a limited number of stimuli may not be generalized to stimuli that are outside of the range tested (Parise, 2016). It is therefore important to know exactly the shape of internal mappings across sensory cues (e.g., non-monotonic mapping) using a wide range of parametrically manipulated stimuli before testing the effect of cross-modal congruency as in previous studies. Failing to do so may lead to issues such as reversed cross-modal congruency effects for different ranges of stimuli. In addition, presenting a large number of stimuli is useful for making the goal of the study less penetrable to participants, compared to presenting only congruent and incongruent stimulus pairs.

Another advantage of the current stimuli is that they could not be easily categorized into cardinal vowels of an extant linguistic system. This would have made it less likely for participants to utilize orthography or other language-specific factors (e.g., sound eliciting activation of lexical items associated with shape in a certain language) when matching a certain shape to speech sounds. In fact, Cuskley *et al*. (2017) demonstrated that the participants' choice between a round and a spiky shape was influenced by the curvature of letters comprising non-words, even when the non-words were presented aurally. It is also reported that orthographic representation is activated when literate participants process phonological information (Slowiaczek *et al*., 2003; Stone *et al*., 1997). In light of the aforementioned studies, presenting a speech sound with which participants easily associate a specific language, thus giving rise to orthographic representation, may preclude a definitive conclusion on the correspondence between shape and speech acoustics *per se*. Thus, although it is impossible to eliminate linguistic context (e.g., categorical perception effects in speech perception are based on linguistic context influencing the perceived

boundary, or category, of simple vocal sounds; Goldstone, 1994; Liberman *et al.*, 1957), our stimuli have advantages over recorded speech sounds, which are confined to a specific language system.

3. Experiment 2

In Experiment 1, we showed that sound is associated with shape in a non-random manner and that both articulation and pitch contribute to this effect. However, although the explicit matching paradigm demonstrates that the association exists, it cannot reveal whether it exerts an influence on how visual shape is subjectively represented. Therefore, Experiment 2 was conducted to further examine whether sound–shape correspondence can induce a shift in perceptual representation. As in Experiment 1, we conducted sub-experiments with pitch-varying (Experiment 2a) and pitch-constant (Experiments 2b and 2c) stimuli to closely examine the contribution of articulation and pitch.

In all three experiments, an ambiguous visual shape was presented with vowel sounds. The task was to adjust the curvature of the shape in the response display to make it as similar as possible to the previously presented, ambiguous visual shape. We hypothesized that if perceptual representation indeed changes based on sound–shape correspondence, it should be affected by the concurrently presented sound information.

The difference among the three experiments was the vowel stimuli presented. In Experiment 2a, the vowel sounds which were associated with the round shape and the spiky shape were chosen from the pitch-varying stimulus set, based on the results of Experiment 1a. Even if we do find the effect of sound on shape perception, however, it is difficult to tease apart the contribution of articulation and pitch in this effect, with the vowels from the pitch-varying set. Therefore, we conducted Experiments 2b and 2c with vowels from the pitch-constant set. The auditory stimuli in Experiment 2b had the same articulatory features as those in Experiment 2a but differed in that pitch was kept constant across the sounds. However, since the results of Experiments 1a and 1b demonstrated that the sound–shape correspondence pattern differs between the sounds used in Experiments 2a and 2b — in other words, the correspondence was present for sounds in Experiments 2a whereas that was not the case for sounds in Experiments 2b — we could not rule out the possibility that such difference, and not the presence and absence of pitch, would influence the results. Therefore, we conducted Experiment 2c with pitch-constant sounds that were comparable in terms of sound–shape correspondence patterns to the sounds used in Experiment 2a, to determine the role of pitch.

3.1. Methods

3.1.1. Participants

Twenty individuals (8 males and 12 females, 20–31 years of age) from Korea University took part in Experiment 2a, and a separate group of 20 individuals (7 males and 13 females, 19–27 years of age) participated in Experiment 2b. A new group of 20 participants (5 males and 15 females) ran in Experiment 2c. None of them participated in Experiment 1. They gave informed consent approved by the Korea University Institutional Review Board (KU-IRB17-85-A-1). All participants had normal or corrected-to-normal vision and hearing, and their native language was Korean.

3.1.2. Stimuli

Visual Stimuli. The visual stimuli were identical for all three experiments.

To generate an ambiguous shape that could not be categorized as round or spiky, linear interpolation was carried out on the x, y coordinates of the points on the contour of a round and a spiky shape. We searched for points on the contours to which the line from the center of the square matrix embedding each shape formed an angle of 0° to 360° (with steps of 1°) with the horizontal axis passing through the center. As a result, we extracted 360 points each along the contour of the two shapes. The interval for linear interpolation was defined as the Euclidean distance between the two points in each shape that corresponded to one another in terms of the angle formed with the horizontal axis. The interval was divided by 14 to extract coordinates for 13 intermediate shapes. This process was repeated for 360 pairs of coordinates (each pair contains one coordinate from the round shape and one coordinate from the spiky shape), which were connected to form the contour of the intermediate shapes. Therefore, we were able to generate a total of 15 visual stimuli including the roundest shape and the spikiest shape (Fig. 3A). The reason we chose to create 15 shapes is because we concluded that the number was appropriate for levels of shape adjustment in the response phase (see Section 3.1.4) based on results of a pilot test, and the size of the steps along the parametric shape space was determined accordingly [step size = (the difference between corresponding coordinates on the round and the spiky shape)/(the total number of shapes − 1)].

Based on the method with which the visual stimuli were generated, the intermediate shape indexed as 8 (shape index 1 indicates round shape and index 15 indicates spiky shape) was assumed to be the most ambiguous stimulus and thus was presented as the test stimulus in the main experiment. Although there is a possibility that the test shape might not be the exact point of perceptual ambiguity, selecting the middle index was justified based on the fact that there was a large number of steps resulting in a small step size. When calculating

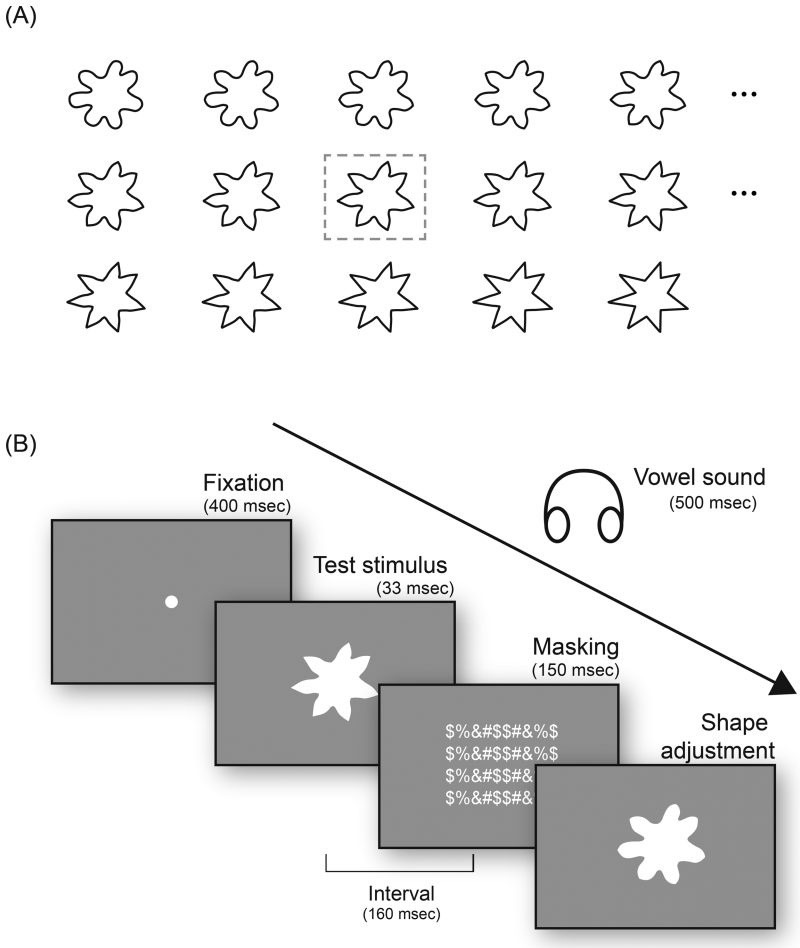

Figure 3. Stimuli and trial sequence in Experiment 2. (A) The 15 shapes used in Experiments 2a, 2b, and 2c. Top left and bottom right show a round and a spiky shape, respectively. The dashed box indicates the shape chosen as the ambiguous test stimulus. (B) Example of a trial sequence in Experiments 2a, 2b, and 2c. For illustration purposes, the shape stimuli are shown in full contrast. After an initial fixation point of 400 ms, the test stimulus was introduced for 33 ms and was accompanied by a vowel sound with a duration of 500 ms. The masking stimulus appeared shortly after and was presented for 150 ms. Finally, a randomly selected shape was presented as a starting point for the shape adjustment task. The trial sequence for the shape adjustment task and the sound detection task are identical up to this point, but the two tasks require different response keys. After participants made a response, visual feedback (correct/incorrect) was provided only for the sound detection task. Participants were instructed to press the spacebar for the next trial.

the Euclidean distance between all the corresponding points (in terms of angles from the horizontal axis) in any two most similar steps, the value ranged from 0 to approximately 0.17° of visual angle in the display. Furthermore, most of the large values came from comparing shape index 1 (the roundest) and 2, or shape index 15 (the spikiest) and 14, which indicates that there was much less difference in shape in the case of middle shape indices.

The visual stimuli were presented on a gray background (27.8 cd/m^2 in luminance). All shapes presented during the experiment spanned 9.5° of visual angle in width and 7.3° of visual angle in height. The contrast (alpha value) of all shapes was 12% (33.6 cd/m^2 in luminance).

Auditory Stimuli: Pitch-Varying and Pitch-Constant Vowels. Auditory stimuli in Experiment 2a were chosen from the vowel sounds presented in Experiment 1a (see Fig. 1 for index of vowel sounds which will be used throughout the paper). Based on the results of Experiment 1a, we selected a low back vowel (LB; index 1 in Fig. 1) which was associated with the round shape, and a high front vowel (HF; index 25 in Fig. 1) which was associated with the spiky shape, significantly above the chance level (see Fig. 2A for the index and the shape-matching results of the sounds used in Experiment 2a). In addition, a low front vowel (LF; index 21 in Fig. 1) associated with neither of the two shapes was presented as a control stimulus. All paired differences of the shape choice for the sounds were statistically significant, and the shape choices for the low back vowel and the high front vowel were significantly different from chance (see Section 2.2.1). To maintain participants' attention on the auditory stimuli during the experiment, catch trials were inserted in which a high back vowel (HB; index 10 in Fig. 1) served as the catch stimulus to be detected (see Section 3.1.4). The pitch of LB (1), HB (10), LF (21), and HF (25) were 120.5 Hz, 131.6 Hz, 120.5 Hz, and 131.6 Hz, respectively.

In Experiment 2b, participants were presented with LB (1), HB (10), LF (21), and HF (25) from the pitch-constant stimulus set. The pitch of the sounds was set to 110 Hz. As shown in the results of Experiment 1b (Fig. 2B), the shape choice for LB (1), LF (21), and HF (25) sounds were not significantly different from chance and did not differ among each other (see Fig. 2B for the index and the shape-matching results of the sounds used in Experiment 2b). Although the vowel stimuli in Experiments 2a and 2b differed in terms of the associated shape, LB (1), HB (10), LF (21), and HF (25) were selected to maintain the articulation constant across the two experiments while manipulating pitch. The vowel sounds in Experiment 2b served as the appropriate comparison to those in Experiment 2a in that pitch manipulation differed while controlling for articulation. However, unlike those in Experiment 2a, the stimuli in Experiment 2b were not associated with a particular shape above the chance level, and sound–shape correspondence was absent in this case.

Therefore, in Experiment 2c, we chose vowel sounds from the pitch-constant set that were significantly associated with the round shape and the spiky shape, although they differed in terms of the tongue body's positions with the vowel stimuli in Experiments 2a and 2b (see Fig. 2B for the index and the shape-matching results of the sounds used in Experiment 2c). Based on the matching results of Experiment 1b (see Fig. 2B), the sound indices 9 and 25 were selected because they each showed the strongest association with the round shape and the spiky shape, respectively, among the 25 sounds in the pitch-constant set. Sound index 21 was presented as the neutral sound, and index 1 was the catch sound. Such choices were made because we attempted to utilize sounds that had the same tongue body positions with Experiments 2a and 2b, as much as possible. As in Experiment 2b, the pitch of all sounds was set to 110 Hz.

3.1.3. Apparatus

The apparatus was identical to that in Experiment 1.

3.1.4. Procedures

The procedures were identical for all three experiments.

Before the main experiment, participants went through a practice session which served to familiarize them with the task instructions. Only the catch sound was explicitly given to participants during the instructions.

The procedures of the main experiment are shown in Fig. 3B. Each trial began with a fixation dot (subtending 0.5° × 0.5°) lasting 400 ms, after which the test stimulus was presented for 33 ms (subtending 9.5° × 7.3°). A masking stimulus consisting of 4 × 10 arrays of random symbols (subtending 11.5° × 10°) appeared 160 ms after the onset of the test stimulus (Hung *et al.*, 2017), and the exposure time was 150 ms. On each trial, a vowel sound was presented together with the test stimulus for 500 ms, and we presented the sounds at three different stimulus onset asynchronies (SOAs): 50 ms before the test stimulus (−50 ms), simultaneously with the test stimulus (0 ms), and 50 ms after the test stimulus (+50 ms). The manipulation of SOA was to examine whether perceptual representation, if influenced by sound–shape correspondence, depends on the temporal synchrony between the audiovisual features (for more details, see Section 3.3). In the response display, participants had to perform one of two tasks, the shape adjustment task or the sound detection task. For trials with no sound, LB/sound matched with round shape, LF/neutral sound, and HF/sound matched with spiky shape, participants had to perform the shape adjustment task by reporting the perceived shape of the test stimulus using the method of adjustment. Participants could alter the level of the shape stimulus presented on the response display to render it as similar as possible to the perceived test stimulus. The shape from which participants started the adjustment was randomly determined among the 15 shapes every trial (see Section 3.1.2).

As participants responded with left and right arrow keys, the visual stimulus was constantly updated so that the shape became more round or spiky. When participants reached the roundest/spikiest shape, a message appeared to inform that they had done so. On the other hand, when the catch sound was presented, participants had to perform a sound detection task by pressing the 'f' key. Visual feedback ('correct' or 'incorrect') was provided only for responses to those trials. Participants were correct when responding with the 'f' key to the catch sound, but incorrect when pressing the arrow keys for the catch sound or pressing the 'f' key for other sounds. Response time was not limited for either task type. A one-minute break was given three times during the experiment.

Each sound condition, except for the catch sound, was repeated 20 times for each of the three SOA conditions (3 sounds $\times$ 20 repetitions $\times$ 3 SOAs = 180 trials). Catch sounds were randomly inserted 20 times during the experiment (the mean and the standard error of the trials in which participants gave a correct response were 16.95 $\pm$ 0.749 in Experiment 2a, 15.25 $\pm$ 1.066 in Experiment 2b, and 16 $\pm$ 0.696 in Experiment 2c). Also, there were 60 trials in which only the test stimulus was presented without any vowel sounds, adding up to a total of 260 trials. However, incorrect trials were re-added at the end of the experiment to match the number of trials in each condition, and therefore the total number of trials differed among participants (the mean and the standard error of the number of incorrect trials across participants were 5.6 $\pm$ 1.134 in Experiment 2a, 12.3 $\pm$ 2.474 in Experiment 2b, and 4.25 $\pm$ 0.739 in Experiment 2c).

3.1.5. Data Analysis

The data analysis was identical for all three experiments.

Since the purpose of inserting catch trials was to make participants process sound information, the trials in which participants heard the catch sound were excluded from further analysis. For the conditions that were included in the analysis (3 SOAs $\times$ 3 sounds + 1 no-sound), we calculated the average of the shape indices the participants reported to be the same as the test stimulus, in other words, the average of the final shapes at which participants stopped the adjustment. Instead of comparing among conditions with the arbitrary shape indices (1 to 15), the shape index was rendered more informative by subtracting the average of the no-sound trials from the average for each combination of sound and SOA condition — LB/sound matched with round shape, LF/neutral sound, HF/sound matched with spiky shape, each coupled with three SOAs. This modified shape index was used for comparison among the conditions. The index 0 indicates the perceived shape of the test stimulus when no sound was presented with it. Negative and positive indices indicate rounder and spikier shapes, respectively.

3.2. Results

Mean values of the modified shape index are displayed in Fig. 4. For each experiment, a two-way repeated-measures ANOVA was conducted with sound (LB/sound matched with round shape, LF/neutral sound, HF/sound matched with spiky shape) and SOA (−50 ms, 0 ms, 50 ms) as within-participant variables.

3.2.1. Experiment 2a: Pitch-Varying Vowels

Results are shown in Fig. 4A. The main effect of sound type on perceived shape of the test stimulus was significant ($F_{(2,38)} = 4.308$, $p < 0.05$, $\eta^2 = 0.185$, Greenhouse–Geisser-corrected η^2). Neither the main effect of SOA nor the SOA–sound interaction effect reached statistical significance ($F_{(2,38)} = 0.505$, $p = 0.564$; $F_{(4,76)} = 0.668$, $p = 0.616$; both Greenhouse–Geisser-corrected).

Post-hoc pairwise comparisons using paired t-tests were conducted on data collapsed across the SOA conditions: we observed significant differences between LB (1) and HF (25), and LF (21) and HF (25) [$t(19) = 2.400$, $p < 0.05$, Cohen's $d = 0.664$; $t(19) = 2.370$, $p < 0.05$, Cohen's $d = 0.615$; both FDR-corrected]. These results indicate that the test stimulus in trials with the low back vowel and the low front vowel was perceived to be rounder than when the high front vowel was presented. However, there was no statistical difference in the perceived shape of the test stimulus between the low back vowel trials and the low front vowel trials [$t(19) = 0.215$, $p = 0.832$; FDR-corrected].

We also compared each sound condition with the no-sound condition by conducting a series of one-sample t-tests against 0. Note that the value 0 of the modified index refers to the perceived shape of the test stimulus without any sound (see Section 3.1.5). The results revealed a significant difference between LB (1) and 0 [$t(19) = 2.991$, $p < 0.05$, Cohen's $d = 0.669$; FDR-corrected], suggesting that the test stimulus was perceived to be a rounder shape when the low back vowel was presented, compared to when the test stimulus was not accompanied by vowel sounds. The difference in perceived shape between other conditions and 0 did not reach statistical significance [LF (21): $t(19) = 2.239$, $p = 0.056$; HF (25): $t(19) = 0.537$, $p = 0.597$; both FDR-corrected].

These results indicate that sound information is influencing visual shape perception in a direction consistent with sound–shape correspondence. That there is a shift in the content of visual perception indicates that sound–shape correspondence is affecting the perceptual processing stage.

3.2.2. Experiment 2b: Pitch-Constant Vowels

Figure 4B demonstrates the results of Experiment 2b. The main effect of sound type on perceived shape of the test stimulus was not significant ($F_{(2,38)} =$

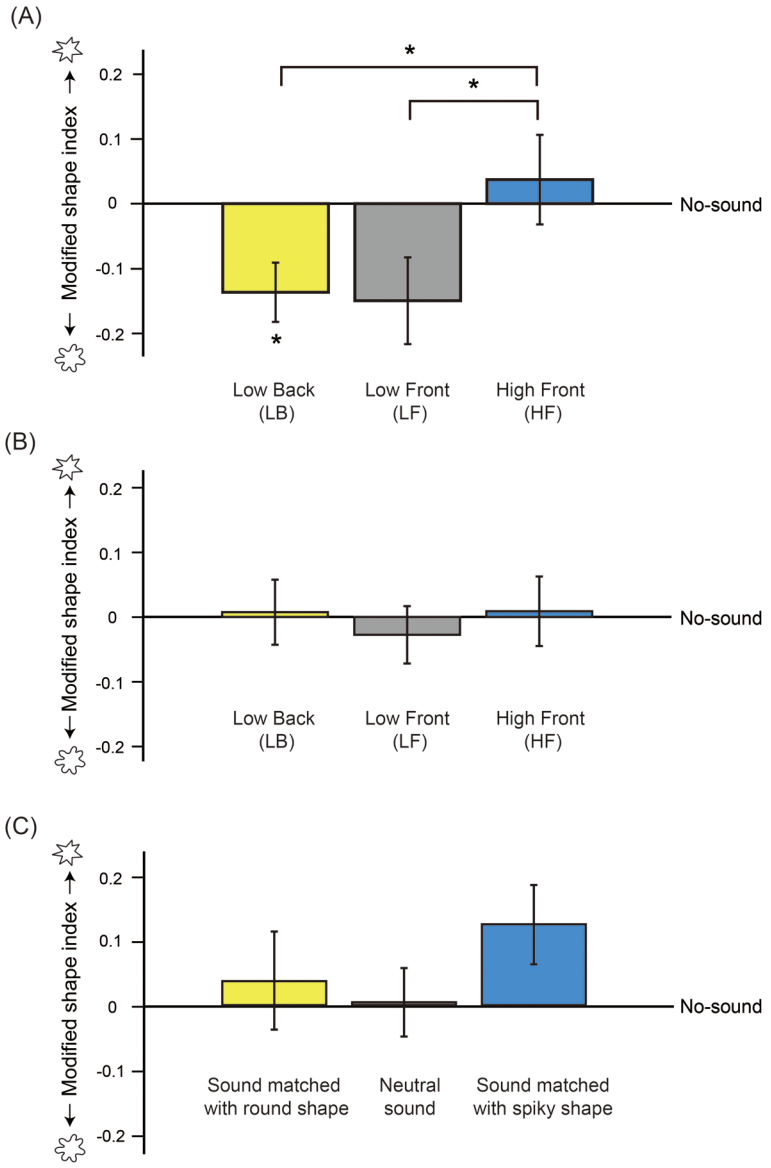

Figure 4. Shape adjustment results of Experiments 2a (A), 2b (B), and 2c (C). The results are averaged across participants for each of the three sound conditions (LB/sound matched with round shape, LF/neutral sound, and HF/sound matched with spiky shape). The y-axis indicates the modified shape index, with 0 indicating the perceived shape for the no-sound condition. Error bars represent ± 1 SEM. Asterisks indicate significant differences ($^{*}p < 0.05$, corrected for false discovery rate).

0.288, $p = 0.751$, Greenhouse–Geisser-corrected). Neither the main effect of SOA nor the SOA–sound interaction effect reached statistical significance ($F_{(2,38)} = 0.703$, $p = 0.501$; $F_{(4,76)} = 1.291$, $p = 0.281$; both Greenhouse–Geisser-corrected).

The sounds used in Experiment 2b had the same tongue body positions — articulation — as the sounds in Experiment 2a, but the difference was that the pitch value was kept identical among the sounds. Therefore, comparing the results of Experiments 2a and 2b suggests the possibility that the difference in pitch among the vowel sounds is driving the effect of sound on visual shape perception. However, since the sounds in Experiment 2b also failed to show significant association with a particular shape in the matching task in Experiment 1b, the absence of sound–shape correspondence might be the reason for the discrepancy in the results of Experiments 2a and 2b. To confirm the role of pitch and rule out the difference in terms of sound–shape correspondence effects, we conducted Experiment 2c.

3.2.3. Experiment 2c: Pitch-Constant Vowels

Figure 4C demonstrates the results of Experiment 2c. The main effect of sound type on perceived shape of the test stimulus was not significant ($F_{(2,38)} = 1.577$, $p = 0.220$, Greenhouse–Geisser corrected). Neither the main effect of SOA nor the SOA–sound interaction effect reached statistical significance ($F_{(2,38)} = 0.118$, $p = 0.557$; $F_{(4,76)} = 0.247$, $p = 0.911$; both Greenhouse–Geisser-corrected).

These results, along with those of Experiments 2a and 2b, suggest that the difference in pitch among the vowel sounds is important for sound information to influence visual shape perception. In the current experiment setting where sound–shape correspondence is not made explicit and sound is task-irrelevant, the influence of pitch seems to override that of articulation.

3.3. Discussion

Experiment 2 was conducted to examine whether concurrent presentation of sound information shifts visual shape representation based on sound–shape correspondence. In Experiment 2a, the low back vowel and the low front vowel biased the shape representation toward the round shape compared to the high front vowel. The results of Experiments 2b and 2c further show that this effect — the effect of sound on shape representation — depends largely on the variation in pitch among the vowel sounds.

The observation that pitch and its correspondence with visual shape affects perceptual representation is intriguing because studies examining such effects are focused on its correspondence with visual motion direction or elevation (Guo *et al*., 2017; Maeda *et al*., 2004; Orchard-Mills *et al*., 2016), although pitch is one of the features most frequently studied in the cross-modal correspondence literature due to the wide range of features it is associated with

(e.g., visual lightness/darkness, elevation, motion direction, size); indeed, visual motion and elevation are features possessing a tighter link with pitch compared to visual shape (Parise *et al*., 2014). Therefore, the current results are meaningful in that shift in the content of visual representation has not been explored for pitch–shape correspondence, not to mention the correspondence between vowel sound and shape (but see Hung *et al*., 2017; Parise and Spence, 2009).

One of the influential frameworks of multisensory research is modeling factors modulating multisensory integration with the Bayesian integration theory (Burr and Alais, 2006; Ernst, 2007; Ernst and Bülthoff, 2004; Seilheimer *et al*., 2014). According to this model, humans combine sensory information in an efficient manner by utilizing prior beliefs and weighting the sensory signals depending on their reliability, also referred to as 'cue combination'. We interpret the current results in the context of cue combination. Although research has prevailed on how sensory cues providing redundant or complementary estimates of a single property are integrated, Spence (2011) raised the possibility of modeling cross-modal correspondence within the framework of the Bayesian integration theory. It is plausible that cross-modal features, although not stemming from the same source, may be combined based on the prior information we have of the systematic mapping across the senses (this prior knowledge is also referred to as the 'coupling prior'; Ernst, 2007; Parise, 2016). In Experiment 2a, sound–shape correspondence acted as prior information, modulating audiovisual interaction of vowel sound and shape. Experiments 2b and 2c further show that it is the pitch component embedded within the vowel sound that interacts with shape information. In other words, pitch–shape correspondence — which can be interpreted as a form of sound–shape correspondence — is the coupling prior modulating audiovisual interaction, thereby leading to change in perceptual representation.

That sound and shape are integrated based on their correspondence, even without participants' explicitly evaluating the association between the features, shows that the current results reflect multisensory interaction. When asked after the experiment, none of the participants were aware that the sounds and shapes could bear a relationship, and none of them were conscious of the irrelevant sound information influencing their response in the shape adjustment task. Indeed, sound–shape correspondence is less intuitive compared to correspondence between stimulus features that share similar neural codes (e.g., both loudness and brightness are coded in terms of an increase/decrease in the magnitude of stimulation) or are associated frequently in nature (e.g., larger objects tend to produce low-frequency sounds), in which the relationship may be more cognitively penetrable (Deroy and Spence, 2016; Evans and Treisman, 2010; Sidhu and Pexman, 2018; Spence and Deroy, 2012). Taking these into account, we argue that the driving force behind our results is that

sound–shape correspondence modulates audiovisual interactions even in the absence of explicit awareness of the association. Our finding is stronger evidence of perceptual representation modulated by cross-modal correspondence compared to some of the previous studies, which have shown that cross-modal correspondence influences performance and perception only when its saliency is boosted (e.g., explicitly informing participants of the association or orienting attention toward the association; Chiou and Rich, 2012; Klapetek *et al*., 2012; Orchard-Mills *et al*., 2016).

In the present study, we did not find the effect of SOA, although temporal synchrony/asynchrony between features is known to be an important factor determining multisensory integration. It is important to consider, however, that the relationship between sound and shape is different from that of features belonging to a single property which explicitly require spatiotemporal integration and are susceptible to SOA manipulation (e.g., audiovisual speech/action; see Chen and Vroomen, 2013 for review; Obermeier and Gunter, 2015). Therefore, it is somewhat surprising that despite such characteristics, Hung *et al*. (2017) reported an interaction effect between SOA and cross-modal congruency using similar stimuli as in our study (e.g., the sound 'bubu' and round visual shape): the effect of cross-modal congruency was observed only when the auditory stimulus preceded the visual stimulus by 150 ms. Yet, such results might reflect a priming effect, where a clearly audible sound presented earlier brings the masked, low-contrast visual shape congruent with it to conscious perception. This may not be a mere speculation since a previous study demonstrated the priming effect of non-words presented aurally, on performance of a subsequent task related to shape (Sidhu and Pexman, 2017).

4. General Discussion

The present study demonstrates the non-arbitrary relationship between speech sound and shape, and further implies that perceptual representation is influenced by the association. Importantly, the synthetic vowel sounds we created enabled us to more closely examine the separate roles of articulation and pitch in these effects. In Experiment 1, we found that the frontness articulation plays an important role in the correspondence between vowel and shape, and that the effect of height articulation can be attributed to pitch effects. In other words, both articulation and pitch are contributing to sound–shape correspondence. Results of Experiment 2 show that sound–shape correspondence affects representation of an ambiguous shape accompanied by task-irrelevant vowel sounds, providing compelling evidence for audiovisual interaction of vowel sound and shape. This effect was observed only when the concurrently presented vowel sounds possessed different pitch, leading to the conclusion that

the correspondence between pitch within vowel acoustics and shape is driving this effect. The different pattern of results elicited by pitch-varying and pitch-constant vowels in both Experiments 1 and 2 suggests the importance of taking into account the factors mingled in speech sounds in investigating sound–shape correspondence.

Previous studies have utilized various experimental paradigms to explore cross-modal correspondences. For example, a number of studies have reported that response times or accuracy in speeded detection/classification tasks are modulated according to the congruency of cross-modal stimuli (Ben-Artzi and Marks, 1995; Bernstein and Edelstein, 1971; Brunel *et al*., 2015; Brunetti *et al*., 2018; Chiou and Rich, 2012; D'Ausilio *et al*., 2014; Evans and Treisman, 2010; Gallace and Spence, 2006; Getz and Kubovy, 2018; Jamal *et al*., 2017; Marks, 1987, 2004; Melara and O'Brien, 1987; Walker and Walker, 2012). Other examples include the Implicit Association Test (Greenwald *et al*., 1998), through which Parise and Spence (2012) demonstrated that response time is affected by the cross-modal congruency of stimuli assigned to the same response key, or an adjustment paradigm where participants had to adjust the pitch/loudness of a tone to reach the tone that matched a certain level of visual brightness (Marks, 1974). Furthermore, a handful of studies employ the explicit matching paradigm, as in Experiment 1 of the current study (D'Onofrio, 2014; Maurer *et al*., 2006; McCormick *et al*., 2015; Spector and Maurer, 2013). It should be noted, however, that our study departs from others by presenting synthesized speech sounds based on Articulatory Phonology (Ohala *et al*., 1986), which specifies human speech in terms of coordinated articulatory movements of vocal tract organs. Using an articulatory synthesizer enabled us to examine sound–shape correspondence in a more systematic manner compared to previous studies by parsing the role of features combined in vowel acoustics. In addition, the parametric manipulation of articulatory gestures allowed us to present a large number of stimuli. This could help minimize major challenges arising from employing a matching paradigm, such as undesired range effects and cognitive penetrability (see Section 2.3 for details).

Although results from the aforementioned paradigms allow us to argue that cross-modal correspondence influences human performance, whether it affects perceptual representation via multisensory interaction remains unresolved. Several studies support the claim that corresponding cross-modal stimuli induce a shift in perceptual representation (Guo *et al*., 2017; Maeda *et al*., 2004; O'Leary and Rhodes, 1984; Takeshima and Gyoba, 2013). Parise and Spence (2009) found that congruent audiovisual pairs increased the Just Noticeable Differences (JNDs) of participants' temporal-order judgments, which shows that the coupling between sound and shape reduced the reliability of the unisensory estimates and promoted multisensory integration. Another study

worth noting is Hung *et al*. (2017), which demonstrates that congruent sound–shape pairs lowered the threshold for detecting the masked, low-contrast visual shape in one of the SOA conditions. However, although these results are important pieces of evidence indicating that sound and shape can be integrated into perceptual representation, whether and how the 'content' of perceptual representation in one modality is altered by information from another modality through multisensory interaction still remains elusive. To demonstrate the change in perceptual content, we took a different approach by parametrically manipulating the roundness/spikiness dimension in Experiment 2. By breaking the dimension into many levels and employing the adjustment technique to sensitively measure one's subjective experience of perceptual representation, we could observe in which direction it shifts depending on each sound condition. This approach enabled us to conclude that the content of the visual shape representation is altered through sound and shape cues being combined based on sound–shape correspondence.

At this point, it is natural to wonder about the brain mechanism subserving cross-modal correspondence and its influence on perceptual representation. Some studies have shown that activity in the superior temporal sulcus (STS) is modulated by cross-modal congruency, although the direction of this effect is equivocal (Barraclough *et al*., 2005; Calvert *et al*., 2000; Lüttke *et al*., 2016; Meyer *et al*., 2011; Van Atteveldt *et al*., 2004). Even considering the dearth of evidence, the involvement of STS in cross-modal correspondence is not far-fetched, with research suggesting that this area refers to the 'content' or 'identity' of sensory information when combining unimodal inputs, rather than similarity in the physical properties, such as spatiotemporal congruence, of very simple sensory stimuli (Calvert, 2001). In addition to such brain regions regarded as multisensory convergence zones, a body of research has revealed activity changes in sensory-specific regions, even primary sensory cortices, induced by signals from a different sensory modality or multisensory co-activation (see Driver and Noesselt, 2008 for review; Calvert *et al*., 1997; Giard and Peronnet, 1999; Pekkola *et al*., 2005; Wallace *et al*., 2004). Noteworthy, these findings utilized stimuli which provide information on the same property or identity (e.g., audiovisual speech/action). Recent studies, however, have begun to make progress in directly addressing the neural mechanism underlying audiovisual 'correspondence' (McCormick *et al*., 2018). For example, Bien *et al*. (2012) showed that temporarily disrupting the intraparietal cortex with transcranial magnetic stimulation (TMS) resulted in the loss of the pitch-size congruency effect on the performance in an auditory spatial localization task. Their results are also in line with an anecdotal report that the 'bouba–kiki' effect is not generalized to patients with damage in the angular gyrus (Ramachandran and Hubbard, 2003). Peiffer-Smadja and Cohen (2019), a recent imaging study on the 'bouba–kiki' effect, found that

the congruency of sound–shape correspondence modulated activations in the bilateral prefrontal cortex and the occipitotemporal visual cortex. Together with the aforementioned studies, the current results demonstrating a shift in visual shape representation caused by sound, suggest the possibility of signals from auditory-specific areas modulating the activity of visual-specific areas, through feedback influences from higher-order regions such as the STS, parietal, and even frontal regions (Macaluso *et al*., 2000; Ramachandran and Hubbard, 2001). Future research is needed to determine the exact locus of cross-modal correspondence and the interaction among sensory and higher-order cortices. The extent to which the underlying neural mechanism is shared or differs across different types of cross-modal correspondences is also an intriguing question that will hopefully be answered through further research.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (No. NRF-2016R1A2B4011 267 and No. NRF-2017M3C7A1029659) awarded to C-YK and by National Institutes of Health-National Institute on Deafness and Other Communication Disorders (No. DC-002717) to Haskins Laboratories. We also thank Christian Wallraven for helpful comments on the research and on the manuscript.

Supplementary Material

Supplementary material is available online at:
https://doi.org/10.6084/m9.figshare.11346611

Note

1. In Experiment 1a, we presented two round–spiky pair types: a pair resembling stimuli in previous studies (Ahlner and Zlatev, 2010; Cuskley *et al*., 2017; Ramachandran and Hubbard, 2001) and a more bulbous pair (D'Onofrio, 2014). Both shape pairs maintain a round–spiky distinction but differ in the degree of overall roundness, and we employed both pair types to examine whether they yield different results. Pair type was a between-participant factor, which means that each participant was presented with only one of the two pair sets throughout the whole experiment. However, as we did not find a significant effect of pair type on shape choice, we kept the pair type constant across the participants in Experiment 1b.

References

Ahlner, F. and Zlatev, J. (2010). Cross-modal iconicity: a cognitive semiotic approach to sound symbolism, *Sign Syst. Stud.* **38**, 298–348.

Ashby, F. G., Maddox, W. T. and Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model, *Psychol. Sci.* **5**, 144–151.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W. and Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions, *J. Cogn. Neurosci.* **17**, 377–391.

Belkin, K., Martin, R., Kemp, S. E. and Gilbert, A. N. (1997). Auditory pitch as a perceptual analogue to odor quality, *Psychol. Sci.* **8**, 340–342.

Ben-Artzi, E. and Marks, L. E. (1995). Visual–auditory interaction in speeded classification: role of stimulus difference, *Percept. Psychophys.* **57**, 1151–1162.

Bernstein, I. H. and Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time, *J. Exp. Psychol.* **87**, 241–247.

Bien, N., ten Oever, S., Goebel, R. and Sack, A. T. (2012). The sound of size: crossmodal binding in pitch–size synesthesia: a combined TMS, EEG and psychophysics study, *NeuroImage* **59**, 663–672.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/.

Brainard, D. H. (1997). The psychophysics toolbox, *Spat. Vis.* **10**, 433–436.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J. and Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners, *Cognition* **126**, 165–172.

Brunel, L., Carvalho, P. F. and Goldstone, R. L. (2015). It does belong together: cross-modal correspondences influence cross-modal integration during perceptual learning, *Front. Psychol.* **6**, 358. DOI:10.3389/fpsyg.2015.00358.

Brunetti, R., Indraccolo, A., Del Gatto, C., Spence, C. and Santangelo, V. (2018). Are cross-modal correspondences relative or absolute? Sequential effects on speeded classification, *Atten. Percept. Psychophys.* **80**, 527–534.

Burr, D. and Alais, D. (2006). Combining visual and auditory information, in: *Visual Perception — Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso and P. U. Tse (Eds), *Progress in Brain Research, Vol. 155*, pp. 243–258. Elsevier, Amsterdam, The Netherlands.

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies, *Cereb. Cortex* **11**, 1110–1123.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., William, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D. and David, A. S. (1997). Activation of auditory cortex during silent lipreading, *Science* **276**, 593–596.

Calvert, G. A., Campbell, R. and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex, *Curr. Biol.* **10**, 649–657.

Chen, L. and Vroomen, J. (2013). Intersensory binding across space and time: a tutorial review, *Atten. Percept. Psychophys.* **75**, 790–811.

Chen, Y.-C. and Spence, C. (2017). Assessing the role of the 'unity assumption' on multisensory integration: a review, *Front. Psychol.* **8**, 455. DOI:10.3389/fpsyg.2017.00445.

Chiou, R. and Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting, *Perception* **41**, 339–353.

Cooke, T., Jäkel, F., Wallraven, C. and Bülthoff, H. H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects, *Neuropsychologia* **45**, 484–495.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*, 2nd edn. Chapman and Hall, London, UK.

Crisinel, A.-S. and Spence, C. (2010). A sweet sound? Food names reveal implicit associations between taste and pitch, *Perception* **39**, 417–425.

Crisinel, A.-S. and Spence, C. (2012). A fruity note: crossmodal associations between odors and musical notes, *Chem. Senses* **37**, 151–158.

Cuskley, C., Simner, J. and Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect, *Psychol. Res.* **81**, 119–130.

D'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J. J. and Fadiga, L. (2014). Vision of tongue movements bias auditory speech perception, *Neuropsychologia* **63**, 85–91.

Davis, R. (1961). The fitness of names to drawings. A cross-cultural study in Tanganyika, *Br. J. Psychol.* **52**, 259–268.

Deroy, O. and Spence, C. (2016). Crossmodal correspondences: four challenges, *Multisens. Res.* **29**, 29–48.

Deroy, O. and Valentin, D. (2011). Tasting liquid shapes: investigating the sensory basis of cross-modal correspondences, *Chemosens. Percept.* **4**, 80–90.

D'Onofrio, A. (2014). Phonetic detail and dimensionality in sound–shape correspondences: refining the Bouba–Kiki paradigm, *Lang. Speech* **57**, 367–393.

Driver, J. and Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments, *Neuron* **57**, 11–23.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch, *J. Vision* **7**, 7. DOI:10.1167/7.5.7.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**, 429–433.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept, *Trends Cogn. Sci.* **8**, 162–169.

Evans, K. K. and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features, *J. Vision* **10**, 6. DOI:10.1167/10.1.6.

Fort, M., Martin, A. and Peperkamp, S. (2015). Consonants are more important than vowels in the bouba–kiki effect, *Lang. Speech* **58**, 247–266.

Gaißert, N., Wallraven, C. and Bülthoff, H. H. (2010). Visual and haptic perceptual spaces show high similarity in humans, *J. Vision* **10**, 2. DOI:10.1167/10.11.2.

Gallace, A. and Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size, *Percept. Psychophys.* **68**, 1191–1203.

Getz, L. M. and Kubovy, M. (2018). Questioning the automaticity of audiovisual correspondences, *Cognition* **175**, 101–108.

Giard, M. H. and Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study, *J. Cogn. Neurosci.* **11**, 473–490.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination, *J. Exp. Psychol. Gen.* **123**, 178–200.

Greenwald, A. G., McGhee, D. E. and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test, *J. Pers. Soc. Psychol.* **74**, 1464–1480.

Guo, L., Bao, M., Guan, L. and Chen, L. (2017). Cognitive styles differentiate crossmodal correspondences between pitch glide and visual apparent motion, *Multisens. Res.* **30**, 363–385.

Hidaka, S., Teramoto, W., Keetels, M. and Vroomen, J. (2013). Effect of pitch–space correspondence on sound-induced visual motion perception, *Exp. Brain Res.* **231**, 117–126.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval, *Am. J. Psychol.* **109**, 219–238.

Hung, S.-M., Styles, S. J. and Hsieh, P.-J. (2017). Can a word sound like a shape before you have seen it? Sound–shape mapping prior to conscious awareness, *Psychol. Sci.* **28**, 263–275.

Jamal, Y., Lacey, S., Nygaard, L. and Sathian, K. (2017). Interactions between auditory elevation, auditory pitch, and visual elevation during multisensory perception, *Multisens. Res.* **30**, 287–306.

Kim, H.-W., Nam, H. and Kim, C.-Y. (2018). [i] is lighter and more greenish than [o]: intrinsic association between vowel sounds and colors, *Multisens. Res.* **31**, 419–437.

Klapetek, A., Ngo, M. K. and Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search?, *Atten. Percept. Psychophys.* **74**, 1154–1167.

Köhler, W. (1929). *Gestalt Psychology*. Liveright, New York, NY, USA.

Köhler, W. (1947). *Gestalt Psychology*, 2nd edn. Liveright, New York, NY, USA.

Kovic, V., Plunkett, K. and Westermann, G. (2010). The shape of words in the brain, *Cognition* **114**, 19–28.

Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries, *J. Exp. Psychol.* **54**, 358–368.

Ludwig, V. U., Adachi, I. and Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans, *Proc. Natl Acad. Sci. USA* **108**, 20661–20665.

Lüttke, C. S., Ekman, M., van Gerven, M. A. J. and de Lange, F. P. (2016). Preference for audiovisual speech congruency in superior temporal cortex, *J. Cogn. Neurosci.* **28**, 1–7.

Macaluso, E., Frith, C. D. and Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention, *Science* **289**, 1206–1208.

Maeda, F., Kanai, R. and Shimojo, S. (2004). Changing pitch induced visual motion illusion, *Curr. Biol.* **14**, R990–R991.

Marks, L. E. (1974). On associations of light and sound: the mediation of brightness, pitch, and loudness, *Am. J. Psychol.* **87**, 173–188.

Marks, L. E. (1987). On cross-modal similarity: auditory–visual interactions in speeded discrimination, *J. Exp. Psychol. Hum. Percept. Perform.* **13**, 384–394.

Marks, L. E. (1989). On cross-modal similarity: the perceptual structure of pitch, loudness, and brightness, *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 586–602.

Marks, L. E. (2004). Cross-modal interactions in speeded classification, in: *The Handbook of Multisensory Processes*, G. A. Calvert, C. Spence and B. E. Stein (Eds), pp. 85–105. Bradford Book, Cambridge, MA, USA.

Marks, L. E., Ben-Artzi, E. and Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination, *Int. J. Psychophysiol.* **50**, 125–145.

Masson, L. M., Bulthé, J., Op de Beeck, H. P. and Wallraven, C. (2016). Visual and haptic shape processing in the human brain: unisensory processing, multisensory convergence, and top-down influences, *Cereb. Cortex* **26**, 3402–3412.

Maurer, D., Pathman, T. and Mondloch, C. J. (2006). The shape of boubas: sound–shape correspondences in toddlers and adults, *Dev. Sci.* **9**, 316–322.

McCormick, K., Kim, J. Y., List, S. and Nygaard, L. C. (2015). Sound to meaning mappings in the bouba–kiki effect, in: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings and P. P. Maglio (Eds), pp. 1565–1570, Cognitive Science Society, Austin, TX, USA.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C. and Sathian, K. (2018). Neural basis of the crossmodal correspondence between auditory pitch and visuospatial elevation, *Neuropsychologia* **112**, 19–30.

Melara, R. D. and O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions, *J. Exp. Psychol. Gen.* **116**, 323–336.

Mermelstein, P. (1973). Articulatory model for the study of speech production, *J. Acoust. Soc. Am.* **53**, 1070. DOI:10.1121/1.1913427.

Meyer, G. F., Greenlee, M. and Wuerger, S. (2011). Interactions between auditory and visual semantic stimulus classes: evidence for common processing networks for speech and body actions, *J. Cogn. Neurosci.* **23**, 2291–2308.

Moos, A., Smith, R., Miller, S. R. and Simmons, D. R. (2014). Cross-modal associations in synaesthesia: vowel colours in the ear of the beholder, *i-Perception* **5**, 132–142.

Ngo, M. K., Velasco, C., Salgado, A., Boehm, E., O'Neill, D. and Spence, C. (2013). Assessing crossmodal correspondences in exotic fruit juices: the case of shape and sound symbolism, *Food Qual. Prefer.* **28**, 361–369.

Obermeier, C. and Gunter, T. C. (2015). Multisensory integration: the case of a time window of gesture–speech integration, *J. Cogn. Neurosci.* **27**, 292–307.

Occelli, V., Spence, C. and Zampini, M. (2009). Compatibility effects between sound frequency and tactile elevation, *NeuroReport* **20**, 793–797.

Ohala, J. J., Browman, C. P. and Goldstein, L. M. (1986). Towards an articulatory phonology, *Phonol. Yearb.* **3**, 219–252.

O'Leary, A. and Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception, *Percept. Psychophys.* **35**, 565–569.

Orchard-Mills, E., Van der Burg, E. and Alais, D. (2016). Crossmodal correspondence between auditory pitch and visual elevation affects temporal ventriloquism, *Perception* **45**, 409–424.

Ozturk, O., Krehm, M. and Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound–shape cross-modal correspondences in 4-month-olds, *J. Exp. Child Psychol.* **114**, 173–186.

Parise, C. V. (2016). Crossmodal correspondences: standing issues and experimental guidelines, *Multisens. Res.* **29**, 7–28.

Parise, C. V. and Spence, C. (2009). "When birds of a feather flock together": synesthetic correspondences modulate audiovisual integration in non-synesthetes, *PLoS ONE* **4**, e5664. DOI:10.1371/journal.pone.0005664.

Parise, C. V. and Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test, *Exp. Brain Res.* **220**, 319–333.

Parise, C. V., Knorre, K. and Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing, *Proc. Natl Acad. Sci. USA* **111**, 6104–6108.

Peiffer-Smadja, N. and Cohen, L. (2019). The cerebral bases of the bouba–kiki effect, *NeuroImage* **186**, 679–689.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A. and Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T, *NeuroReport* **16**, 125–128.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies, *Spat. Vis.* **10**, 437–442.

Peña, M., Mehler, J. and Nespor, M. (2011). The role of audiovisual processing in early conceptual development, *Psychol. Sci.* **22**, 1419–1421.

Ramachandran, V. S. and Hubbard, E. M. (2001). Synaesthesia — a window into perception, thought and language, *J. Conscious. Stud.* **8**, 3–34.

Ramachandran, V. S. and Hubbard, E. M. (2003). Hearing colors, tasting shapes, *Sci. Am.* **288**, 52–59.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M. and Browman, C. (1996). CASY and extensions to the task-dynamic model, in: *Proceedings of the 1st ESCA ETRW on Speech Production Modeling and 4th Speech Production Seminar*, Autrans, France, pp. 125–128.

Schutz, M. and Kubovy, M. (2009). Causality and cross-modal integration, *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1791–1810.

Seilheimer, R. L., Rosenberg, A. and Angelaki, D. E. (2014). Models and processes of multisensory cue combination, *Curr. Opin. Neurobiol.* **25**, 38–46.

Sidhu, D. M. and Pexman, P. M. (2017). A prime example of the maluma/takete effect? Testing for sound symbolic priming, *Cogn. Sci.* **41**, 1958–1987.

Sidhu, D. M. and Pexman, P. M. (2018). Five mechanisms of sound symbolic association, *Psychon. Bull. Rev.* **25**, 1619–1643.

Slowiaczek, L. M., Soltano, E. G., Wieting, S. J. and Bishop, K. L. (2003). An investigation of phonology and orthography in spoken-word recognition, *Q. J. Exp. Psychol. A* **56**, 233–262.

Spector, F. and Maurer, D. (2013). Early sound symbolism for vowel sounds, *i-Perception* **4**, 239–241.

Spence, C. (2011). Crossmodal correspondences: a tutorial review, *Atten. Percept. Psychophys.* **73**, 971–995.

Spence, C. and Deroy, O. (2012). Crossmodal correspondences: innate or learned?, *i-Perception* **3**, 316–318.

Spence, C. and Gallace, A. (2011). Tasting shapes and words, *Food Qual. Prefer.* **22**, 290–295.

Stone, G. O., Vanhoy, M. and Van Orden, G. C. (1997). Perception is a two-way street: feedforward and feedback phonology in visual word recognition, *J. Mem. Lang.* **36**, 337–359.

Takeshima, Y. and Gyoba, J. (2013). Changing pitch of sounds alters perceived visual motion trajectory, *Multisens. Res.* **26**, 317–332.

Tarte, R. D. (1974). Phonetic symbolism in adult native speakers of Czech, *Lang. Speech* **17**, 87–94.

Van Atteveldt, N., Formisano, E., Goeble, R. and Blomert, L. (2004). Integration of letters and speech sounds in the human brain, *Neuron* **43**, 271–282.

Walker, P. and Walker, L. (2012). Size–brightness correspondence: crosstalk and congruity among dimensions of connotative meaning, *Atten. Percept. Psychophys.* **74**, 1226–1240.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A. and Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences, *Psychol. Sci.* **21**, 21–25.

Wallace, M. T., Ramachandran, R. and Stein, B. E. (2004). A revised view of sensory cortical parcellation, *Proc. Natl Acad. Sci. USA* **101**, 2167–2172.

Welch, R. B. (1972). The effect of experienced limb identity upon adaptation to simulated displacement of the visual field, *Percept. Psychophys.* **12**, 453–456.

Whalen, D. H. and Levitt, A. G. (1995). The universality of intrinsic F_0 of vowels, *J. Phon.* **23**, 349–366.