Sounds exaggerate visual shape

Timothy D. Sweeny [a,*], Emmanuel Guzman-Martinez [b], Laura Ortega [b], Marcia Grabowecky [b,c], Satoru Suzuki [b,c]

[a] Department of Psychology, University of California – Berkeley, Berkeley, CA, USA
[b] Department of Psychology, Northwestern University, Evanston, IL, USA
[c] Interdepartmental Neuroscience Program, Northwestern University, Evanston, IL, USA

ARTICLE INFO

ABSTRACT

While perceiving speech, people see mouth shapes that are systematically associated with sounds. In particular, a vertically stretched mouth produces a /woo/ sound, whereas a horizontally stretched mouth produces a /wee/ sound. We demonstrate that hearing these speech sounds alters how we see aspect ratio, a basic visual feature that contributes to perception of 3D space, objects and faces. Hearing a /woo/ sound increases the apparent vertical elongation of a shape, whereas hearing a /wee/ sound increases the apparent horizontal elongation of a shape. We further demonstrate that these sounds influence aspect ratio coding. Viewing and adapting to a tall (or flat) shape makes a subsequently presented symmetric shape appear flat (or tall). These aspect ratio aftereffects are enhanced when associated speech sounds are presented during the adaptation period, suggesting that the sounds influence visual population coding of aspect ratio. Taken together, these results extend previous demonstrations that visual information constrains auditory perception by showing the converse – speech sounds influence visual perception of a basic geometric feature.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Mouth shapes are systematically associated with sounds due to the anatomy of vocalization (e.g., Liberman & Mattingly, 1985; Sapir, 1929; Yehia, Rubin, & Vatikiotis-Bateson, 1998). Experiencing these crossmodal associations may lead to neural connectivity or multimodal tuning for visual processing of mouth shapes and auditory processing of speech sounds (Nath & Beauchamp, 2011; Wilson, 2002). Indeed, patches of temporal cortex are activated more strongly by combinations of faces and voices than by either alone (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004), and viewing of silent lip reading activates the auditory cortex (Calvert et al., 1997). Behaviorally, presenting a talking face influences speech perception in infants (Kuhl & Meltzoff, 1982) and improves speech recognition in adults (von Kriegstein et al., 2008). In a classic study, McGurk and MacDonald (1976) demonstrated that auditory perception of a phoneme was altered by a concurrently presented face pronouncing a different phoneme.

Because audition is usually regarded as the primary modality for speech perception, speech related auditory–visual interactions have been evaluated in terms of how looking at the mouth influences hearing of speech. Here we investigated the converse. Does hearing speech sounds alter how we see shapes? We examined the perception of aspect ratio (horizontal or vertical elongation) for two reasons. First, aspect ratio is a fundamental visual feature that is population-coded in the ventral visual pathway (see Suzuki (2005) for a review), contributing to perception of 3D space, objects and faces (e.g., Biederman, 2001; Knill, 1998a, 1998b; Young & Yamane, 1992). Second, horizontal and vertical mouth elongations are ubiquitous in speech

production, with a horizontally elongated mouth typically producing a /wee/ sound and a vertically elongated mouth typically producing a /woo/ sound. We thus used flat (horizontally elongated) and tall (vertically elongated) ellipses as the visual stimuli and /wee/ and /woo/ sounds as the auditory stimuli. We used simple ellipses rather than images of mouths so that observers would be unaware of the relationship between the sounds and aspect ratios. This was important because we wanted to test the hypothesis that consistent auditory–visual coincidences during speech perception develop general auditory–visual associations that influence visual perception at the level of basic shape coding. If the experience of looking at mouth shapes while listening to speech establishes associations between auditory representations of phonemes and visual representations of associated shapes, hearing a /wee/ sound may make a flat ellipse appear even flatter and hearing a /woo/ sound may make a tall ellipse appear even taller (Fig. 1a).

2. Experiment 1: Speech sounds exaggerate appearances of associated visual aspect ratios

We examined perception of a briefly flashed ellipse in three conditions. In the *consistent-sound* condition, an ellipse was presented with a consistent speech sound (a flat ellipse with a /wee/ sound or a tall ellipse with a /woo/ sound). In the *inconsistent-sound* condition, an ellipse was presented with an inconsistent speech sound (a flat ellipse with a /woo/ sound or a tall ellipse with a /wee/ sound). In the control, *environmental-sound*, condition, an ellipse was presented with an environmental sound of no relation to speech or mouth shape (a door shutting or ice cracking). A critical aspect of this experiment was that none of the observers reported awareness of any association between the /wee/ and /woo/ sounds and the flat and tall ellipses during the post-experiment interview. This is not surprising because the ellipses did not resemble mouths and the experimenter did not mention any cross-modal associations during the instructions. Thus, any effect of auditory speech on the visual perception of aspect ratio would have likely occurred implicitly, without an explicit strategy or response bias.

2.1. Method

2.1.1. Observers

In all experiments, undergraduate students from Northwestern University with normal or corrected-to-normal vision and normal hearing gave informed consent to participate, and they were tested individually in a dimly lit room. Seventeen observers participated in this experiment for partial course credit.

2.1.2. Stimuli

We generated ellipses (drawn with dark [54 cd/m^2] 0.057°-thick lines against a white [110 cd/m^2] background) with 11 different aspect ratios (the vertical major axis divided by the horizontal major axis) ranging from flat to tall. The aspect ratios were symmetrically distributed (in log scale) around the circle, −0.485 (1.59° × 0.52°), −0.387
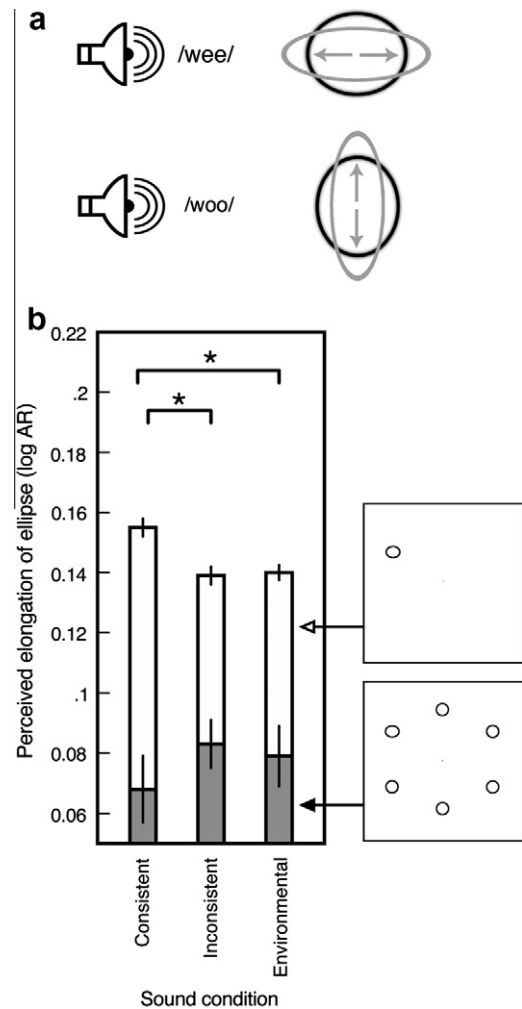
Fig. 1. Speech sounds increased the perceived elongation of consistent ellipses when observers were unaware of the associations between the sounds and shapes. (a) A schematic (not-to-scale) illustration of our primary finding; hearing a /wee/ sound (produced by a horizontally-stretched mouth) made a flat ellipse appear flatter, whereas hearing a /woo/ sound (produced by a vertically-stretched mouth) made a tall ellipse appear taller. (b) Perceived elongation of the ellipses (in the direction of the stimulus elongation, in log aspect ratio) is shown for the consistent-sound, inconsistent-sound, and environmental-sound conditions for the single-stimulus trials (open bars; a display example shown in the top right panel) and the multi-stimulus trials (dark bars; a display example shown in the bottom right panel). The error bars represent ±1 SEM adjusted for the repeated-measures design of the experiment. $^*p < .05$.

(1.44° × 0.59°), −0.271 (1.25° × 0.67°), −0.201 (1.19° × 0.75°), −0.091 (1.05° × 0.85°), 0.0 (circle; 0.95° × 0.95), 0.091 (0.85° × 1.05°), 0.201 (0.75° × 1.19°), 0.271 (0.67° × 1.25°), 0.387 (0.59° × 1.44°), and 0.485 (0.52° × 1.59°). Each ellipse was treated with a Gaussian blur of 2.0-pixel radius to reduce aliasing. The least flat and least tall ellipses (log aspect ratios = ±0.091) were used as the target stimuli because perceived aspect ratios of briefly presented ellipses tend to be exaggerated (e.g., Suzuki & Cavanagh, 1998; Sweeny, Kim, Grabowecky, & Suzuki, 2011). Circles were used as distractors on multi-stimulus trials (see below).

Ten ellipses excluding the circle were presented in the response display for reporting the perceived shape of the flashed ellipse via matching. All ellipses had equivalent areas.

Four sounds were used in the experiment. We created a "flat-mouth" (/wee/) sound by recording a man's voice as he produced a phoneme while stretching the corners of his mouth to make a flat shape. We created a "tall-mouth" (/woo/) sound by recording the same man's voice as he produced a phoneme while bringing the corners of his mouth together to make a tall shape. Two environmental sounds (a door shutting and ice cracking, obtained from a personal sound collection) were used as non-speech control sounds. All sounds were presented via Sennheiser-HD265 headphones, matched for perceived loudness (approximately 53 dB SPL), and differed only slightly in duration (/wee/ = 708 ms, /woo/ = 840 ms, door slamming = 1098 ms, ice cracking = 1156 ms).

Ellipses were presented along an invisible circular orbit (17.9° diameter) around the central fixation cross (0.10°, 62 cd/m^2). The six stimulus locations (top, upper right, lower right, bottom, lower left, and upper left) were symmetrically arranged and evenly spaced (separated by 60° in rotation angle or 4.2° in visual angle) (see Fig. 1b). On each trial, a flat or tall ellipse was presented at one of the six locations with equal probability (16 times at each location, 96 trials total). On half of the trials, the target ellipse was presented alone—*single-stimulus trials* (e.g., Fig. 1b, top right panel), and on the remaining trials, the ellipse was presented among five circles (e.g., Fig. 1b, bottom right panel)—*multi-stimulus trials*. We included multi-stimulus trials because crossmodal interactions have been shown to depend on whether a visual feature is presented alone or in a crowd (Sherman, Sweeny, Grabowecky, & Suzuki, 2012). The multi-stimulus trials also provided a control for response bias (i.e., responding based on sounds irrespective of what is seen); a bias should be more pervasive when perception of the ellipse is less certain, such as in a multi-stimulus trial where the location of the ellipse was difficult to determine in a brief display.

On each trial, a consistent sound (a /wee/ sound for a flat ellipse or a /woo/ sound for a tall ellipse), an inconsistent sound (a /wee/ sound for a tall ellipse or a /woo/ sound for a flat ellipse), or an environmental sound (a door-shutting sound or an ice-cracking sound) was presented along with a flat or tall ellipse presented alone or among circles; all combinations of the four sounds, tall and flat ellipses, and single-stimulus and multi-stimulus displays were randomly intermixed and equally frequent across trials. The experiment was controlled with an Apple MacBook OS X using MATLAB (version 2009b) with the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). The visual stimuli were presented on a 19″ CRT monitor at a viewing distance of 115 cm. Approximately 10 practice trials were given prior to the experiment.

2.1.3. Procedure

Each trial began with the presentation of a fixation cross. The experimenter instructed observers to fixate the central cross (though strict central fixation was not crucial for this experiment). Observers were told to focus on the visual task and that the sounds were uninformative. Following 1000 ms of the fixation display, a sound (a /wee/, /woo/, or environmental sound) was initiated, which was immediately followed by a brief presentation of a flat or tall ellipse (alone or among five circles). The sound (950 ms, on average) was started slightly (33 ms) before the ellipse onset to allow processing of the sound's acoustic property prior to presentation of an ellipse. This brief asynchrony, however, was well within the range known to create audio–visual fusion (e.g., Miller & D'Esposito, 2005). The ellipse (alone or among circles) was briefly presented (33 ms). In addition to preventing influences from saccades and high-level deliberative processes, brief presentations make visual perception sensitive to contextual effects (e.g., Suzuki & Cavanagh, 1998; Wolfe, 1984), providing a means to measure crossmodal interactions with increased sensitivity. The fixation cross remained on the screen after the offset of the ellipse for 1000 ms, and was followed by a response display containing a horizontal array of 10 sample ellipses gradually changing from flat to tall. The observer chose the ellipse that appeared most similar to the flashed ellipse by pressing the corresponding button.

2.1.4. Analysis

To reveal crossmodal effects specific to the sound-shape association beyond general effects of a sound or bias in aspect-ratio perception, we averaged the perceived ellipse elongations (in log aspect ratio, with elongations in the veridical direction as positive) across the flat and tall ellipses for each sound condition. Thus, in the consistent-sound condition we averaged the perceived elongations between the tall ellipse presented with the /woo/ sound and the flat ellipse presented with the /wee/ sound, whereas in the inconsistent-sound condition we averaged the perceived elongations between the tall ellipse presented with the /wee/ sound and the flat ellipse presented with the /woo/ sound. In the environmental-sound condition, we averaged the perceived elongations between the tall and flat ellipses presented with the same environmental sound. Because the perceived elongations did not differ between the two environmental sounds, we combined those data together. Perception in this condition provided a baseline against which to evaluate the specificity of the effects with speech sounds, because the environmental sounds had no speech relevance or association with the ellipses (see Section 3.1, *Stimuli*, for a discussion of why this condition provides a more appropriate baseline than a no sound condition).

2.2. Results

On the single-stimulus trials, perceived elongation was larger with the consistent sounds relative to both the inconsistent sounds, $t(16) = 2.457$, $p < .03$, $d = 0.596$, and environmental sounds, $t(16) = 2.802$, $p < .02$, $d = 0.679$ (Fig. 1b). Perceived elongations with the inconsistent and environmental sounds did not differ, $t(16) = .281$, *n.s.* The speech sounds thus increased the perceived elongation of the consistent ellipses (Fig. 1a) without affecting the inconsistent ellipses. No observers reported awareness of the

sound-shape associations or knowledge that the shapes could have been interpreted as mouths. This suggests that the crossmodal shape exaggeration occurs implicitly, and this reasonably rules out response bias.

Note that this cross-modal effect occurred in the context of an overall exaggeration effect. The ellipses appeared more elongated than their veridical aspect ratios (0.091 on average) independently of any effect from the sounds, especially on the single-stimulus trials (Fig. 1b). This overall exaggeration of briefly flashed ellipses was expected based on prior results showing that briefly flashed visual features tend to appear exaggerated (e.g., Suzuki and Cavanagh (1998) for aspect ratio, and Sweeny et al. (2011) for curvature; see Sweeny et al. (2011) for a discussion of possible mechanisms of this exaggeration).

Sounds produced no significant effects on the multi-stimulus trials (t's < 1.213, $n.s.$, with a marginal consistency-by-display-size [one vs. multiple] interaction, $F[2, 32] = 3.093$, $p = .059$, $\eta_p^2 = 0.162$), when the location of the target ellipse was difficult to determine and reported elongations were small (Fig. 1b, shaded bars). This result is consistent with our prior finding that the crossmodal effect of laughter that enhanced the perception of a happy expression for a single face disappeared when the happy face was presented among a crowd of neutral faces (Sherman et al., 2012). This may make sense in the context of the load theory of attention (e.g., Lavie, 2005), which states that when selective attention is not strongly engaged to a target feature, task-irrelevant features receive a relative increase in processing (Pinsk, Doniger, & Kastner, 2004) and are more likely to interfere with perception of the target feature. Thus, the brief presentation and the uncertainty of the location of the target made it unlikely that selective attention was strongly engaged to the target on multi-stimulus trials. It is possible that auditory effects on visual shape coding might depend on attention being focused on the relevant visual target. Note that the absence of the crossmodal effect in the multi-stimulus condition provides further evidence against response bias because if the sounds simply biased observers' responses, they would have equivalently affected responses in the single-stimulus and multiple-stimulus conditions.

3. Experiment 2: Speech sounds influence the population coding of aspect ratio

We have demonstrated that speech sounds associated with tall and flat mouth shapes implicitly (i.e., with no explicit awareness of the auditory–visual associations) exaggerate visual aspect ratios of simple ellipses. A potential mechanism of this crossmodal effect is that hearing speech sounds enhances responses of visual neurons tuned to the associated aspect ratios. To psychophysically evaluate this hypothesis, we investigated the speech sounds' influences on aspect-ratio aftereffects; when a tall (or flat) adaptor shape is followed by a symmetric test shape, the test shape appears to be elongated in the orthogonal direction. Previous research suggested that this repulsive aspect-ratio aftereffect reflects an activation (and adaptation) of aspect-ratio-tuned neurons in the ventral visual pathway

(see Section 3.3). Conveniently, aspect-ratio aftereffects occur with brief adaptation, comparable to the brief speech sounds, when the aftereffects are measured with brief test stimuli (e.g., Suzuki, 2005; Suzuki & Cavanagh, 1998). If hearing a speech sound enhances the activation (and thus adaptation) of neurons tuned to the associated aspect ratio, the aspect-ratio aftereffect should be larger when a consistent speech sound is presented during adaptation than when an inconsistent speech sound or an environmental sound is presented. Note that a response bias would predict the opposite pattern. For example, if a /wee/ sound increased the responses and adaptation of flat-tuned neurons while adapting to a flat ellipse, the test stimulus would appear taller. In contrast, if the /wee/ sound simply produced a bias to respond "flat," observers would report the test stimulus as flatter. As in Experiment 1, none of the observers reported awareness of any auditory–visual associations during the post-experiment interview.

3.1. Method

3.1.1. Observers
Eleven new observers were paid to participate after giving informed consent.

3.1.2. Stimuli
Visual and auditory stimuli were similar to those used in Experiment 1 with the following exceptions. The black ellipses (33 cd/m^2) and the white background (109 cd/m^2) were slightly darker. An ellipse was always presented at the center of the screen and had one of 21 aspect ratios symmetrically distributed (in log scale) around the circle, −0.419 (1.67° × 0.63°), −0.374 (1.61° × 0.68°), −0.343 (1.56° × 0.73°), −0.311 (1.51° × 0.78°), −0.285 (1.46° × 0.83°), −0.221 (1.41° × 0.89°), −0.176 (1.35° × 0.94°), −0.131 (1.30° × 0.99°), −0.087 (1.25° × 1.04°), −0.043 (1.20° × 1.09°), 0.0 (circle; 1.15° × 1.15°), 0.043 (1.09° × 1.20°), 0.087 (1.04° × 1.25°), 0.131 (0.99° × 1.30°), 0.176 (0.94° × 1.35°), 0.221 (0.89° × 1.41°), 0.285 (0.83° × 1.46°), 0.311 (0.78° × 1.51°), 0.343 (0.73° × 1.56°), 0.374 (0.68° × 1.61°), 0.419 (0.63° × 1.67°).

The flat or tall adaptor had one of three amounts of elongation, ±0.043, ±0.131, and ±0.311, and the test stimulus was always the circle, backward masked by a random-dot pattern (consisting of 50% black and 50% white pixels covering the central 10° [horizontal] by 7° [vertical] region). We used three different adaptor elongations for the following reasons. On the one hand, the aspect-ratio aftereffect (on a circle) increases with increased adaptor elongation (e.g., Suzuki, 2005), and a larger aftereffect might provide greater sensitivity for detecting crossmodal modulation. On the other hand, a crossmodal boost of the aftereffect might be more salient when the baseline magnitude of the aftereffect is smaller. The use of different adaptor aspect ratios allowed us to demonstrate effects of speech sounds on the population coding of aspect ratio regardless of these influences.

The sounds (/wee/, /woo/, and environmental sounds, at ~62 db SPL) were presented via a pair of JBL speakers (10–25,000 Hz frequency response) placed symmetrically just in front of the visual display screen; each sound was

played through both speakers and was perceived to be co-localized with the centrally presented adaptor ellipse. Each sound was approximately exponentially attenuated after the first 200 ms, becoming inaudible within approximately 500 ms of onset. In addition to trials with the three sound types, trials with no sounds were intermixed to make sure that the presence of a sound per se did not disrupt aspect-ratio aftereffects, and it did not (see footnote 1). Note that the environmental sound condition is a more appropriate control than the no-sound condition because it is matched to the sound conditions in terms of the potential arousing, alerting, and temporal cueing effects of a sound coincident with the brief visual adaptor. We will thus compare aspect-ratio aftereffects among the consistent-sound (/wee/ with a flat adaptor and /woo/ with a tall adaptor), inconsistent-sound (/wee/ with a tall adaptor and /woo/ with a flat adaptor), and environmental-sound conditions as in Experiment 1. Each block included 24 trials (2 adaptor orientations [tall or flat], 3 amounts of elongation, and 4 sound conditions). Each observer was tested in 12 blocks. Sixteen practice trials were given prior to the experiment.

3.1.3. Procedure

Each trial began with a fixation point (0.10° diameter, 62 cd/m^2) lasting 1760 ms. A visual adaptor lasting 176 ms and a sound lasting ~500 ms (consistent, inconsistent, environmental, or none) were simultaneously initiated. The adaptor was followed by a blank display lasting 470 ms (the sound was audible through the first ~324 ms of the blank display), and then by a test circle lasting 47 ms, which was immediately followed by the random-dot mask lasting 294 ms. After a 1760 ms blank interval (we chose this duration so that the test ellipse would be temporally distinct from the preceding sequence of stimuli), a method of adjustment (e.g., Sweeny et al., 2011) began; a circle appeared in the center of the screen and observers pressed the left or right arrow key to gradually change the aspect ratio of the circle to be flatter (coded as negative aspect ratios) or taller (coded as positive aspect ratios), stepping through the 21 aspect ratios indicated above. Once observers satisfactorily matched the image on the screen with their percept of the test shape, they pressed the space bar and the next trial started after 1 s.

3.2. Results

In order to analyze the magnitude of aspect-ratio aftereffects beyond the variability due to individual differences in the baseline bias (i.e., a tendency to see briefly presented shapes as horizontally or vertically elongated), we computed an *aftereffect index* (in log-aspect-ratio units). Specifically, we subtracted the mean perceived aspect ratio of the test shape following adaptation to a tall ellipse from that following adaptation to the flat ellipse for each observer for each amount of adaptor elongation and for each sound condition. A larger positive value of this index indicates a larger magnitude of the aspect-ratio aftereffect.

A two-factor ANOVA with sound condition (consistent, inconsistent, and environmental) and adaptor elongation (three magnitudes) as the independent variables and the aftereffect index as the dependent variable, yielded
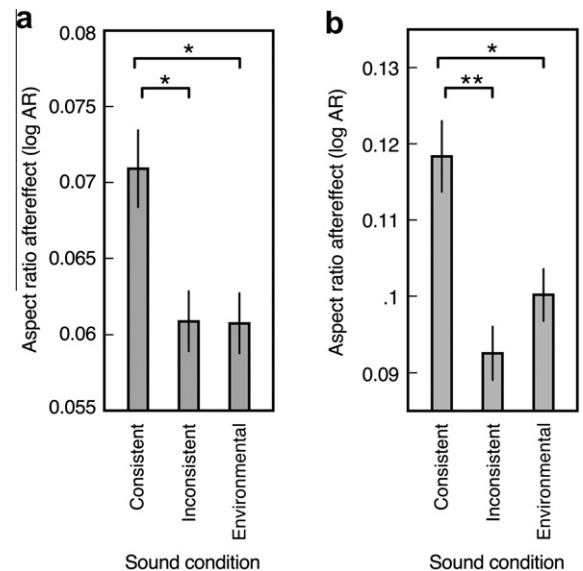
Fig. 2. Speech sounds presented during adaptation to consistent ellipses increased aspect ratio aftereffects. The aftereffect index (see main text for details) is shown for the consistent-sound, inconsistent-sound and environmental-sound conditions. A larger positive value indicates a stronger aftereffect. (a) The results averaged across all amounts of adaptor elongation. (b) The results for the most elongated adaptors. Error bars represent ±1 *SEM* adjusted for the repeated-measures design of the experiment. $^*p < .05$, and $^{**}p < .01$.

significant main effects of sound condition, $F(2, 20) = 4.783$, $p < .02$, $\eta_p^2 = .286$, and adaptor elongation, $F(2, 20) = 33.408$, $p < .0001$, $\eta_p^2 = .771$. The latter indicates that a more elongated adaptor produced a larger aspect-ratio aftereffect (e.g., Suzuki, 2005). The former indicates that the sounds significantly influenced the magnitude of aspect-ratio aftereffects. Follow-up analyses showed that the aspect-ratio aftereffect was significantly larger with the consistent sounds relative to both the inconsistent sounds, $t(10) = 2.701$, $p < .023$, $d = 0.814$, and environmental sounds, $t(10) = 2.653$, $p < .025$, $d = 0.800$, while the aftereffect was equivalent with the inconsistent and environmental sounds, $t(10) = .035$, n.s. (Fig. 2a).[1]

The effect of the consistent sound appears to be especially strong for the most elongated adaptors (Fig. 2b) as reflected in a marginal interaction between sound condition and adaptor elongation, $F(4, 40) = 2.297$, $p < .08$, $\eta_p^2 = 0.200$. The consistent sound enhanced aspect-ratio aftereffects relative to both the inconsistent sound, $t(10) = 5.067$, $p < .001$, $d = 1.528$, and environmental sounds, $t(10) = 2.403$, $p < .037$, $d = 0.724$, with no significant difference between the inconsistent and environmental sounds, $t(10) = 1.311$, n.s.

---

[1] The magnitude of the aspect-ratio aftereffect in the no-sound condition was intermediate (0.067, SE = 0.003). This was confirmed by the fact that the contrast reflecting the hypothesis that the no-sound condition was exactly intermediate, that is {+5, −3, −3, +1} for {Consistent, Inconsistent, Environmental, no-sound}, was significant, $t(10) = 3.597$, $p < .005$, $d = .990$. However, as discussed in the Methods section, the interpretation of the no-sound condition is ambiguous because it differs from the sound conditions on several factors such as arousal, alertness, and temporal cueing, each of which may increase or decrease the aftereffect.

Overall, consistent speech sounds presented during adaptation enhanced aspect-ratio aftereffects relative to inconsistent speech sounds and environmental sounds.

3.3. Discussion

Our results complement the classic McGurk effect by showing that hearing speech sounds distorts visual perception of shape. Importantly, our results demonstrate that associations between auditory and visual features are not merely metaphorical (e.g., Köhler, 1947; Marks, 1996; Sapir, 1929) or limited to influencing response times, accuracy (e.g., Bernstein & Edelstein, 1971; Gallace & Spence, 2006; Marks, 1987) or temporal and spatial integration (e.g., Parise & Spence, 2009), but that they also change a visual feature's appearance.

Aspect ratio is a fundamental visual feature presumably coded by relative activation of a population of neurons tuned to different aspect ratios in the ventral visual pathway (e.g., Kayaert, Biederman, & Vogels, 2003; Regan & Hamstra, 1992; Suzuki, 2005). Our results suggest that audiovisual speech experience facilitates feature specific interactions between auditory processing of spectral patterns and ventral-visual processing of aspect ratio. Specifically, a /wee/ sound (typically associated with a horizontally elongated mouth) and a /woo/ sound (typically associated with a vertically elongated mouth) might make a shape appear flatter or taller (Experiment 1) by crossmodally boosting the activity of neurons tuned to flat and tall aspect ratios, respectively. We behaviorally tested this possibility in Experiment 2 by evaluating the effects of sounds on the aspect-ratio aftereffect.

Viewing a tall (or flat) adaptor shape makes a subsequently presented symmetric shape appear elongated in the orthogonal direction. Comparison between the psychophysical properties of this aftereffect and known physiological properties of cortical visual neurons suggests that this aftereffect reflects an adaptive population coding of aspect ratio in the ventral visual pathway (e.g., Regan & Hamstra, 1992; Suzuki, 2003, 2005; Suzuki & Cavanagh, 1998). For example, viewing a tall shape would strongly activate tall-tuned neurons but only weakly activate flat-tuned neurons, thus strongly adapting (desensitizing) tall-tuned neurons but only weakly adapting flat-tuned neurons. When a symmetric shape is subsequently presented, the strongly adapted tall-tuned neurons would be less activated than the weakly adapted flat-tuned neurons so that the symmetric shape would appear flat. Stronger activation (causing stronger adaptation) of the tall-tuned neurons would produce a larger aftereffect (e.g., making the symmetric shape appear flatter to a greater degree). The magnitude of the aspect-ratio aftereffect therefore provides a behavioral measure of activation of aspect-ratio tuned neurons (Regan & Hamstra, 1992; Suzuki, 2003, 2005; Suzuki & Cavanagh, 1998). A speech sound presented during adaptation to a consistently elongated adaptor increased the aftereffect, suggesting that a /wee/ sound increases activation and adaptation of flat-tuned neurons and a /woo/ sound increases activation and adaptation of tall-tuned neurons.

Although a psychophysical investigation cannot directly reveal how auditory processing might influence the activity of visual neurons, the data can inform and constrain some possibilities. Auditory neurons could boost the responses of visual neurons through excitatory connections, either directly from auditory to visual areas, or indirectly, through a multisensory integration area with connections to auditory and visual cortices (e.g., Nath & Beauchamp, 2011). Another equally plausible account is provided by the motor theory of speech perception (Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985). Listening to speech sounds is known to recruit motor areas underlying speech production (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Watkins, Strafella, & Paus, 2003). Here, covert motor simulation of mouth shapes consistent with speech sounds could have influenced the visual encoding of associated aspect ratios through feedback connectivity (Skipper, Nusbaum, & Small, 2005; Skipper, van Wassenhove, Nusbaum, & Small, 2007). While our data cannot discriminate among these alternatives, a future investigation, which directly addresses them, is warranted.

Interestingly, consistent sounds increased perceived elongation (Experiment 1) and adaptation (Experiment 2), but inconsistent sounds had little effect on either measure. This may suggest that auditory or motor-simulation input boosts responses of associated visual neurons that are already activated, but that it does not drive visual neurons on its own or inhibit responses of unassociated visual neurons. Thus, the underlying crossmodal influences may be excitatory and multiplicative.

In conclusion, what we see is shaped by what we hear, not only when seeing meaningful objects such as faces (Smith, Grabowecky, & Suzuki, 2007), but also when seeing a simple geometric feature, adding to growing evidence that perceptual reality is fundamentally multimodal (e.g., Schroeder & Foxe, 2005).

Acknowledgments

References

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7*, 1190–1192.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87*, 241–247.

Biederman, I. (2001). Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision, 13*, 241–253.

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

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lip reading. *Science, 276*, 593–596.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience, 15*, 399–402.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech reviewed. *Psychonomic Bulletin & Review, 13*, 361–377.

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

Kayaert, G., Biederman, I., & Vogels, R. (2003). Shape tuning in macaque inferior temporal cortex. *The Journal of Neuroscience, 23*, 3016–3027.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, 36 (ECVP Abstract Supplement).

Knill, D. C. (1998a). Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues. *Vision Research, 38*, 1655–1682.

Knill, D. C. (1998b). Discriminating surface slant from texture: Comparing human and ideal observers. *Vision Research, 38*, 1683–1711.

Köhler, W. (1947). *Gestalt psychology* (2nd ed.). New York: Liveright.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences, 9*, 75–82.

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

Marks, L. E. (1987). Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 384–394.

Marks, L. E. (1996). On perceptual metaphors. *Metaphor and symbolic activity, 11*, 39–66.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience, 25*, 5884–5893.

Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of Neuroscience, 31*, 1704–1714.

Parise, C. V., & Spence, C. (2009). When birds of a feather flock together: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*, e5664.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Pinsk, M. A., Doniger, G. M., & Kastner, S. (2004). Push–pull mechanism of selective attention in human extrastriate cortex. *Journal of Neurophysiology, 92*, 622–629.

Regan, D., & Hamstra, S. J. (1992). Shape discrimination and the judgment of perfect symmetry: Dissociation of shape from size. *Vision Research, 32*, 1845–1864.

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

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology, 15*, 454–458.

Sherman, A., Sweeny, T. D., Grabowecky, M., & Suzuki, S. (2012). Laughter exaggerates happy and sad faces depending on visual context. *Psychonomic Bulletin & Review, 1069–9384*, 1–7.

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage, 25*, 76–89.

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*, 2387–2399.

Smith, E., Grabowecky, M., & Suzuki, S. (2007). Auditory–visual crossmodal integration in perception of face gender. *Current Biology, 17*, 1680–1685.

Suzuki, S. (2003). Attentional selection of overlapped shapes: A study using brief aftereffects. *Vision Research, 43*, 549–561.

Suzuki, S. (2005). High-level pattern coding revealed by brief shape aftereffects. In C. Clifford & G. Rhodes (Eds.). *Fitting the mind to the world: Adaptation and aftereffects in high-level vision. Advances in Visual Cognition Series* (Vol. 2). Oxford University Press.

Suzuki, S., & Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 24*(5), 1315–1341.

Sweeny, T. D., Kim, Y. J., Grabowecky, M., & Suzuki, S. (2011). Internal curvature signal and noise in low- and high-level vision. *Journal of Neurophysiology, 105*, 1236–1257.

von Kriegstein, K., Dogan, O., Giraud, A.-L., Kell, C. A., Gruter, T., Kleinschmidt, A., et al. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *Proceedings of the National Academy of Sciences, USA, 105*, 6747–6752.

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia, 41*, 989–994.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*, 625–636.

Wolfe, J. (1984). Short test flashes produce large tilt aftereffects. *Vision Research, 24*, 1959–1964.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal tract and facial behavior. *Speech Communication, 26*, 23–43.

Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science, 256*, 1327–1331.

i-Comment

Hearing mouth shapes: Sound symbolism and the reverse McGurk effect

Charles Spence
Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, Oxford, UK; e-mail:
charles.spence@psy.ox.ac.uk

Ophelia Deroy
Centre for the Study of the Senses, University of London, London, UK; e-mail: ophelia.deroy@gmail.com

Abstract. In their recent article, Sweeny, Guzman-Martinez, Ortega, Grabowecky, and Suzuki (2012) demonstrate that heard speech sounds modulate the perceived shape of briefly presented visual stimuli. Ovals, whose aspect ratio (relating width to height) varied on a trial-by-trial basis, were rated as looking wider when a /woo/ sound was presented, and as taller when a /wee/ sound was presented instead. On the one hand, these findings add to a growing body of evidence demonstrating that audiovisual correspondences can have perceptual (as well as decisional) effects. On the other hand, they prompt a question concerning their origin. Although the currently popular view is that crossmodal correspondences are based on the internalization of the natural multisensory statistics of the environment (see Spence, 2011), these new results suggest instead that certain correspondences may actually be based on the sensorimotor responses associated with human vocalizations. As such, the findings of Sweeny et al. help to breathe new life into Sapir's (1929) once-popular "embodied" explanation of sound symbolism. Furthermore, they pose a challenge for those psychologists wanting to determine which among a number of plausible accounts best explains the available data on crossmodal correspondences.

Keywords: sound symbolism, crossmodal correspondence, multisensory integration, audition, vision.

The participants in a new study by Sweeny et al. (2012) were briefly presented (for 30 ms) with an outline oval shape from one of six positions arranged in a virtual circle around fixation. The aspect ratio of the oval varied randomly on a trial-by-trial basis, sometimes the oval was wider than it was tall, whereas on other trials it was taller than it was wide. Just before the onset of the visual stimulus, a speech sound (either a /wee/ or a /woo/) or an environmental sound (either the sound of a door closing or ice cracking) was presented over headphones. At the end of each trial, the participants had to choose, from an array of 10 ovals displayed on the computer screen, the one that had just been presented. Even though the sound was completely task-irrelevant, participants nevertheless still rated the oval as looking taller on the /wee/ trials (compared with when any of the other sounds had been presented) and as looking significantly wider on the /woo/ trials (again when compared with the presentation of any of the other sounds).

The Sweeny et al. (2012) experiment can be thought of as constituting a kind of reverse McGurk effect. McGurk and MacDonald (1976) famously demonstrated that the lip movements that one sees can influence the speech sounds that one hears. Here, heard speech sounds influence the "mouth-like" shape that one sees. Sweeny et al. put forward two hypotheses concerning the origin of this crossmodal effect of speech sounds on shape perception: Either participants pick up on the statistical association that exists between speech sounds and the mouth shapes that are seen on a locutor's lips, or else the correspondence emerges from the automatic processing of articulatory movements in the motor areas underlying speech production (Liberman & Mattingly, 1985). Making a /wee/ sound requires a speaker to form a wider oval shape with his or her mouth than when uttering a /woo/ sound (which requires a taller, narrower mouth shape). Whichever hypothesis is correct, this particular crossmodal effect appears to operate at an implicit level because, when questioned after the experiment, none of the participants reported that they had ever thought of the outline ovals in terms of mouth shapes. This, in

turn, suggests that the correspondence operates in a relatively involuntary and automatic fashion (see Spence & Deroy, submitted).

In a follow-up experiment, Sweeny et al. (2012) went on to demonstrate that the crossmodal consequences of speech sounds on the visual perception of object shape were likely operating at a relatively low level of the visual system. Specifically, they showed that the repeated presentation of the speech sound that was more consistent with a given aspect ratio oval was capable of inducing an aspect-ratio aftereffect (as compared, once again, with the inconsistent sound or environmental sound conditions). Neurophysiological research suggests that such phenomena likely result from low-level distributed shape coding mechanisms, with individual visual neurons coding for specific aspect ratios (Kayaert, Biederman, & Vogels, 2003).

The results of Sweeny et al. (2012) add to a growing body of research (see Guzman-Martinez et al., 2012; Parise & Spence, 2009), demonstrating that crossmodal correspondences can have genuinely *perceptual* effects (in addition to the more *decisional* effects likely targeted by earlier research using speeded discrimination tasks; see Marks, 2004, for a review) and operate in a relatively early and automatic manner. They also raise an important question regarding the origins of such crossmodal correspondences, and the link between "sound symbolism" and other forms of audiovisual correspondence. Back in 1929, Sapir, the founding father of sound symbolism research, first demonstrated a connection between speech sounds and the size of their referents. He showed that the majority of people thought that the larger of two round tables should be called "mal," whereas the smaller table should be called "mil" (a finding, incidentally, that has been replicated in many different languages/countries). Most people also match angular shapes with the word "takete" while matching rounded shapes with the word "maluma" (see Köhler, 1929; Ramachandran & Hubbard, 2001).

A likely origin for the former effect is the link between sounds and the size of their sources. In recent years, the mil-mal sound symbolism effect has been assimilated to the widely documented crossmodal correspondence between object size and the pitch of the sound it makes when struck, sounded, voiced etc. (see Parise & Spence, 2009; Spence, 2011). Other things being equal, larger objects/animals tend to make lower-pitched sounds than do smaller objects/animals.

However, according to an alternative account (incidentally one that was first put forward by Sapir, 1929), the mil-mal phenomenon is actually speech specific and results from the fact that the mouth has to make a wider opening when uttering an /a/ sound than when uttering an /i/ sound. Once related to speech, the correspondence between sound and size can also be explained not merely in terms of audiovisual but also in terms of audiomotor, associations, linking the sounds that one hears to the automatic articulatory movements generated when listening to speech (Galantucci, Fowler, & Turvey, 2006). If the latter account were to be correct, this crossmodal correspondence would then become embodied (Pezzulo et al., 2011), grounded in sensorimotor associations, rather than based on an external association between two sensory experiences, whose resemblance would be processed in an amodal manner.

Although both statistical and embodied accounts can explain the sound–size correspondence, the latter theory of sound symbolism would appear to provide a more plausible explanation for the existence of sound–shape correspondences. The fact that most people match angular shapes with the word "takete" while matching rounded shapes with the word "maluma" (Ramachandran & Hubbard, 2001) may most parsimoniously be explained by the suggestion that it is the sharp vocal transitions made by the mouth when uttering the plosive sounds in "takete" that people map onto the sharp/angular shape. There does not seem to be an obvious alternative account in terms of the natural statistics of the environment (unless, that is, it should turn out that angular objects give rise to sounds that are relevantly different from rounded objects when, for example, explored haptically; see Guzman-Martinez et al., 2012). One way in which to distinguish between the statistical and embodied accounts here would be to test whether this correspondence exists only in cases or species where the vocalizing follows the takete-sharp mouth movements rule (contrast this with the correspondence between sound–size of the source that can be found across species, independent of their rules of vocalization; see Ludwig et al., 2011).

One of the challenges for future research will be to try and figure out whether crossmodal correspondences such as the one between the relative pitch of a speech sound and the relative size of an object are better accounted for in terms of the statistical account (according to which an organism internalizes the multisensory statistics of the environment) versus an embodied account (which posits that such mappings result from the physical constraints on how speech sounds are generated). It will

further be interesting here to determine whether the sound–shape and sound–size crossmodal corre-spondences are related, and whether the latter has multiple origins (perhaps originating both in speech and in external associations).

The question of the origin of the association matters when it comes to deciding on the question of the unity versus disunity of the category of crossmodal correspondences. Understanding the role of embodied versus external associations would certainly help to link the results of Sweeny et al. (2012) to others, showing that the shapes we see (and respond to) can also influence the pitch (or fundamental frequency) of the speech sounds we utter (Parise & Pavani, 2011) or that making a mouth movement (consistent with "ba" or "da") can give rise to a McGurk effect when listening to speech sounds, just as when actually viewing someone else's mouth movements uttering those sounds (see Sams, Mottonen, & Sihvonen, 2005).

References

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*, 361–377. doi:10.3758/BF03193857.

Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., & Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Current Biology, 22*, 383–388. doi:10.1016/j.cub.2012.01.004.

Kayaert, G., Biederman, I., & Vogels, R. (2003). Shape tuning in macaque inferior temporal cortex. *The Journal of Neuroscience, 23*, 3016–3027. doi:10.1167/3.9.514.

Köhler, W. (1929). *Gestalt psychology.* New York: Liveright.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech revisited. *Cognition, 21*, 1–36. doi:10.1016/0010-0277(85)90021-6.

Ludwig, V. U., Adachi, I., & Matzuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences USA, 108*, 20661–20665. doi:10.1073/pnas.1112605108.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85–105). Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748. doi:10.1038/264746a0.

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research, 214*, 373–380. doi:10.1007/s00221-011-2836-3.

Parise, C., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*(5), e5664. doi:10.1371/journal.pone.0005664.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. J. (2011). The mechanics of embodiment: A dialog on embodiment and computational modelling. *Frontiers in Psychology, 2*, 5. doi:10.3389/fpsyg.2011.00005.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3–34. doi:10.1098/rspb.2000.1576.

Sams, M., Mottonen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research, 23*, 429–435. doi:10.1016/j.cogbrainres.2004.11.006.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225–239. doi:10.1037/h0070931.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995. doi:10.3758/s13414-010-0073-7.

Spence, C., & Deroy, O. (submitted). How automatic are crossmodal correspondences? *Consciousness & Cognition*.

Sweeny, T. D., Guzman-Martinez, E., Ortega, L., Grabowecky, M., & Suzuki, S. (2012). Sounds exaggerate visual shape. *Cognition, 124*, 194–200. doi:10.1016/j.cognition.2012.04.009.