**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 correspondences 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, & Silvonen, 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., & Matuzawa, 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/rsbp.2000.1576.

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