frontiers in
PSYCHOLOGY

GENERAL COMMENTARY
published: 01 May 2014
doi: 10.3389/fpsyg.2014.00352

The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech

Ryan A. Stevenson[1]*, Mark T. Wallace[2,3,4,5,6] and Nicholas Altieri[7]

[1] *Psychology Department, University of Toronto, Toronto, ON, Canada*
[2] *Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN, USA*
[3] *Department of Hearing and Speech Sciences, Vanderbilt University Medical Center, Nashville, TN, USA*
[4] *Vanderbilt Kennedy Center, Vanderbilt University Medical Center, Nashville, TN, USA*
[5] *Department of Psychology, Vanderbilt University, Nashville, TN, USA*
[6] *Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, USA*
[7] *Department of Communication Sciences and Disorders, Idaho State University, Pocatello, ID, USA*
**Correspondence: ryan.andrew.stevenson@gmail.com*

Edited by:
Bruce D. McCandliss, Vanderbilt University, USA

Reviewed by:
Urs Maurer, University of Zurich, Switzerland

Keywords: multisensory processing, audiovisual integration, speech perception, temporal processing, sensory processing, crossmodal, perceptual binding, speech integration

A commentary on

Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs
by Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., and van Atteveldt, N. (2013). Front. Psychol. 4:331. doi: 10.3389/fpsyg.2013.00331

The amount of research focused on multisensory speech perception has expanded considerably in recent years. Much of this research has focused on which factors influence whether or not an auditory and a visual speech input are "integrated" (i.e., perceptually bound); a special case of how our perceptual systems solve the "binding problem" (Treisman, 1996). The factors that have been identified as influencing multisensory integration can be roughly divided into two groups. First are the low-level stimulus factors that include the physical characteristics of the sensory signals. The most commonly studied of these include the spatial (e.g., Macaluso et al., 2004; Wallace et al., 2004) and temporal (e.g., Miller and D'Esposito, 2005; Stevenson et al., 2011) relationship of the two inputs, and their relative effectiveness (e.g., James et al., 2012; Kim et al., 2012) in driving a neural, perceptual, or behavioral response. The second group of factors can be considered more higher-order or cognitive, and include factors such as the semantic congruence of the auditory and visual signals (Laurienti et al., 2004) or whether or not the gender of the speaker's voice is matched to the face (Lachs and Pisoni, 2004).

While these two categories can be considered conceptually distinct, they are related because of their mutual dependence upon the natural statistics of signals in the environment. When auditory and visual speech signals are closely proximate in time (low-level), they are more likely to have originated from the same speaker, and thus should be integrated (Dixon and Spitz, 1980; Stevenson et al., 2012b). Likewise, if an auditory and a visual speech signal are semantically congruent (high-level), they are more likely to have originated from the same speaker and thus should be integrated (Calvert et al., 2000). Given that these low- and high-level factors are each reflective of the natural statistics of the environmental signals, they will generally co-vary. Taking speech as an example, in a natural setting, the temporally-coincident auditory and visual components of a syllable or word are also semantically congruent (Spence, 2007).

To date, most research has investigated these low- and high-level factors independently. These studies have been highly informative, providing descriptions as to how each of these factors contributes to the process of multisensory integration. What has not received a great deal of focus is the interplay between these factors. A handful of experiments have investigated how low-level factors interact with one another and influence multisensory integration (Macaluso et al., 2004; Royal et al., 2009; Stevenson et al., 2012a), but few have attempted to bridge between low-level stimulus-characteristics and high-level cognitive factors (Vatakis and Spence, 2007). A recent article by Ten Oever et al. (2013), *Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs* addresses this gap in our understanding by investigating the interaction between stimulus timing and semantic congruency modulated by changes in place of articulation or voicing.

In this study, participants were presented with single-syllable stimuli, with auditory, visual, and audiovisual syllables systematically manipulated according to place of articulation and voicing. In addition, the temporal alignment of the audiovisual presentations was also parametrically varied. Hence, semantic content was varied through changes both in the auditory (voicing) and visual (place

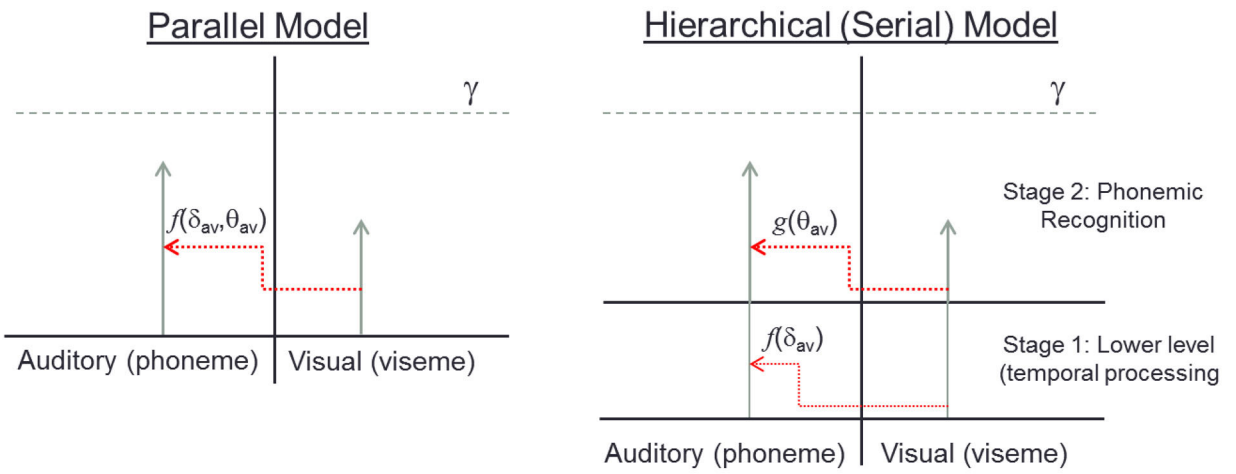

FIGURE 1 | The left panel shows a "parallel accumulator" model with auditory and visual evidence racing toward threshold (γ). The amount of visual influence on the auditory signal is a function, *f*, of parameters δ, and θ which represent temporal coincidence detection and phonetic congruence respectively, both of which contribute evidence to a single accumulator. The serial model on the right shows two separate stages where integration is affected first by temporal, then by semantic processing. Hence, in stage 1, visual information influences auditory processing only as a function *f* of temporal coincidence. In stage 2, visual information influences auditory processing solely as a function,*g*, of phonemic compatibility.

of articulation) signals, while at the same time, the relative timing of the auditory and visual stimuli were systematically varied. While the results specific to these factors are interesting on their own, most germane to this commentary is how these two factors interacted. The authors measured the window of time within which the visual cue influenced the syllable that was heard. This probabilistic construct, referred to as the "time window of integration" or the "temporal binding window," has been shown to vary greatly according the type of stimulus being integrated (Vatakis and Spence, 2006; Stevenson and Wallace, 2013). In the Ten Oever et al. study, semantically congruent stimuli were found to be associated with a wider temporal binding window than semantically incongruent stimuli. That is, stimulus components that are semantically matched have higher rates of integration at more temporally disparate offsets.

The result is surprising in that it runs counter to predictions generated by hierarchical serial models. In such models, lower-level properties such as stimulus timing are processed initially, and are then followed by the processing of the linguistic (i.e., semantic) content in the auditory and visual signals. However, the current results, by illustrating an interaction between timing and congruency, suggest that hierarchical models are insufficient to explain the data. Rather, we posit that these results are better interpreted within a "parallel accumulation of evidence" framework (**Figure 1**). In this model, the temporal relationship of two sensory inputs provides important information about the likelihood that those two inputs originated from the same speaker and should be integrated. In addition, the semantic congruence of these inputs also provides information as to whether or not the two sensory inputs should be bound. Importantly, these two types of evidence are pooled into a single decision criterion. Thus, within such a framework, when stimuli are semantically congruent, a decreased amount of temporal alignment is needed in order to cross a decision bound that would result in these two inputs being integrated, manifesting in a broader temporal binding window for semantically congruent speech stimulus pairs.

Through this interaction between stimulus timing and semantic congruence, Ten Oever and colleagues provided compelling evidence that low-level stimulus and high-level cognitive factors are not processed in a completely serial manner, but rather interact with one another in the formation of a perceptual decision. These results have significant implications in informing our view as to the neurobiological substrates involved in real-world multisensory perceptual processes. Most importantly, the work suggests that significant feedforward and feedback circuits are engaged in the processing of naturalistic multisensory stimuli, and that these circuits work in a parallel and cooperative fashion in evaluating the statistical relations of the stimuli to one another on both their low-level (i.e., stimulus feature) and high-level (i.e., learned semantic) correspondences.

REFERENCES

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. doi: 10.1016/S0960-9822(00)00513-3

Dixon, N. F., and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception* 9, 719–721. doi: 10.1068/p090719

James, T. W., Stevenson, R. A., and Kim, S. (2012). "Inverse effectiveness in multisensory processing," in *The New Handbook of Multisensory Processes*, ed B. E. Stein (Cambridge, MA: MIT Press), 207–221.

Kim, S., Stevenson, R. A., and James, T. W. (2012). Visuo-haptic neuronal convergence demonstrated with an inversely effective pattern of BOLD activation. *J. Cogn. Neurosci.* 24, 830–842. doi: 10.1162/jocn_a_00176

Lachs, L., and Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecol. Psychol.* 16, 159–187. doi: 10.1207/s15326969eco1603_1

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., and Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Exp. Brain Res.* 158, 405–414. doi: 10.1007/s00221-004-1913-2

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *Neuroimage* 21, 725–732. doi: 10.1016/j.neuroimage.2003.09.049

Miller, L. M., and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25, 5884–5893. doi: 10.1523/JNEUROSCI.0896-05.2005

Royal, D. W., Carriere, B. N., and Wallace, M. T. (2009). Spatiotemporal architecture of cortical receptive fields and its impact on multisensory interactions. *Exp. Brain Res.* 198, 127–136. doi: 10.1007/s00221-009-1772-y

Spence, C. (2007). Audiovisual multisensory integration. *Acoust. Sci. Technol.* 28, 61–70. doi: 10.1250/ast.28.61

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., and Wallace, M. T. (2012a). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Exp. Brain Res.* 219.1, 121–137. doi: 10.1007/s00221-012-3072-1

Stevenson, R. A., VanDerKlok, R. M., Pisoni, D. B., and James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *Neuroimage* 55, 1339–1345. doi: 10.1016/j.neuroimage.2010.12.063

Stevenson, R. A., and Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Exp. Brain Res.* 227, 249–261. doi: 10.1007/s00221-013-3507-3

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012b). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *J. Exp. Psychol. Hum. Percept. Perform*. 38.6, 1517. doi: 10.1037/a0027339

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., and van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Front. Psychol.* 4:331. doi: 10.3389/fpsyg.2013.00331

Treisman, A. (1996). The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178. doi: 10.1016/S0959-4388(96)80070-5

Vatakis, A., and Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Res.* 1111, 134–142. doi: 10.1016/j.brainres.2006.05.078

Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756. doi: 10.3758/BF03193776

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., and Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Exp. Brain Res.* 158, 252–258. doi: 10.1007/s00221-004-1899-9

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 January 2014; accepted: 03 April 2014; published online: 01 May 2014.

Citation: Stevenson RA, Wallace MT and Altieri N (2014) The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. Front. Psychol. ***5****:352. doi: 10.3389/fpsyg.2014.00352*

This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology.