Psychon Bull Rev (2011) 18:123–128
DOI 10.3758/s13423-010-0027-z

Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli

Shoko Kanaya · Kazuhiko Yokosawa

Published online: 10 November 2010

Abstract Many studies on multisensory processes have focused on performance in simplified experimental situations, with a single stimulus in each sensory modality. However, these results cannot necessarily be applied to explain our perceptual behavior in natural scenes where various signals exist within one sensory modality. We investigated the role of audio-visual syllable congruency on participants' auditory localization bias or the ventriloquism effect using spoken utterances and two videos of a talking face. Salience of facial movements was also manipulated. Results indicated that more salient visual utterances attracted participants' auditory localization. Congruent pairing of audio-visual utterances elicited greater localization bias than incongruent pairing, while previous studies have reported little dependency on the reality of stimuli in ventriloquism. Moreover, audio-visual illusory congruency, owing to the McGurk effect, caused substantial visual interference on auditory localization. Multisensory performance appears more flexible and adaptive in this complex environment than in previous studies.

Keywords Ventriloquism · Speech perception · Multisensory integration

Daily we experience multisensory scenes comprising various signals from different modalities. It is astonishing that in complex environments we can perceive a pair of related multisensory events or identify a source common to several different sensory signals. For example, we effortlessly identify 'what' is spoken (acoustically) with 'who' is speaking (visually) simultaneously amidst a confluence of many irrelevant voices and faces. How is detection of a composite object accomplished? This remains a puzzle. However, a majority of research surrounding this puzzle have concentrated upon only multisensory integration, or binding, between single events within different modalities.

Recently, Roseboom, Nishida, and Arnold (2009) found that the temporal window of audio-visual synchrony for a pair of audio-visual stimuli, or the interval where we are insensitive to inter-sensory asynchrony, can be modulated dynamically by the relationship between multiple sensory events presented within each modality. This indicates that our multisensory system depends upon interactions among several temporally and spatially proximate sensory events; that is, the manner we integrate multisensory information is not simply determined by the relationship between a single pair of isolated stimuli.

In the domain of audio-visual spatial perception, the well-known ventriloquism effect reflects a biasing of localization of the origin of an auditory stimulus toward the location of a concurrent visual stimulus (Jack & Thurlow, 1973). This familiar illusion underlies our perception of a voice as originating from the visible mouth of a puppet when, in reality, the voice belongs to the ventriloquist. This effect has also been demonstrated with non-associative stimuli such as a simple flash and a synchronized burst of noise (Bertelson & Aschersleben, 1998). Determining factors of the ventriloquism effect have generally been assumed to be sensory factors such as temporal synchrony and spatial proximity of auditory and visual stimuli involved (Vroomen & De Gelder, 2004). Negative results have been reported about the role of

S. Kanaya (✉) · K. Yokosawa
Department of Psychology Graduate School of Humanities and Sociology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, Japan
e-mail: skanaya@l.u-tokyo.ac.jp

cognitive factors, such as familiar contexts or congruency in speech information, on ventriloquism. Radeau and Bertelson (1977) combined a voice with two types of visual stimuli, either a realistic stimulus (a visible talker) or a simplified one (a flashing light synchronized with the amplitude peak of the voice). Exposure to these two audio-visual conditions yielded comparable results in a localization task, suggesting a minimal role of realism in the ventriloquism effect. Bertelson, Vroomen, Wiegeraad, and de Gelder (1994) also found little difference in localization bias when they presented a voice with a video of a talking face, either upright or inverted. Moreover, Colin, Radeau, Deltenre, and Morais (2001) demonstrated that congruency of audio-visual speech syllables failed to modify ventriloquism effect size.

Alternatively, Driver (1996) has shown that ventriloquism is only effective on a target speech sequence that is congruent with the visible facial articulations. In his experiment, participants had to repeat one of two verbal sequences including a target and a distracter arising from a common loudspeaker; the target sequence was cued only by a visual display of a talker's lip gestures. Participants performed better when the visual display was spatially separated from the loudspeaker than when they were located at the same position. This facilitation is considered to result from an illusory separation of sound sources for target and the distracter sequences due to the ventriloquism effect; thus, the illusion functions to enhance selective listening much as does a true physical separation of sound sources (Brungart, Kordik, & Simpson, 2005). In Driver's experiment, an irrelevant stream of auditory speech was added to a pair of audio-visual stimuli. Although the task did not require detection of a sound source explicitly, as is typical in ventriloquism research, the findings nonetheless suggest a possible role of audio-visual congruency in speech information, that is a cognitive factor, on multisensory spatial perception, evident largely especially when multiple signals exist in one sensory modality.

In the present research, our hypothesis is that the multisensory system behaves more flexibly and differently when multiple signals occur with a single sensory modality than when isolated stimuli are presented in each modality. In other words, an inter-sensory relationship on a cognitive factor, for example the congruency in speech syllables, might have an influence on ventriloquism when multiple stimuli exist in a sensory modality. Accordingly, we presented a single voice paired with two different visual stimuli, namely bilateral videos of speaking faces. Previous studies indicate that the location of an auditory stimulus, when presented with bilateral visual stimuli, is biased toward the more physically salient of the two stimuli: for example, a larger visual figure or a more intense light (Bertelson, Vroomen, de Gelder, & Driver, 2000; Radeau, 1985). Our manipulation of this physical saliency of visual stimuli involved two dynamic facial displays; one presented a full face with visible mouth movements (‘visible’ utterance) and the other presented a masked face where lips were occluded, revealing only marginal facial movements (‘masked’ utterance). Although both videos provided information through motion synchronized to the presented voice, the visible region of movement was greater in the former, resulting in a difference in the physical saliency of motion information. In this situation, the more salient one of two visual stimuli, i.e., the visible one is predicted to elicit the more pronounced perceptual shift of a sound source toward its spatial location.

In this experiment, we examined whether audio-visual congruency in syllables affects the auditory localization. Our hypothesis predicts that the localization bias owing to a ‘visible’ utterance, which is congruent with a spoken utterance, is stronger than the bias elicited by an incongruent ‘visible’ utterance. Then, the source of an auditory stimulus, situated at either the side of the ‘visible’ or ‘masked’ utterance (visible-side or masked-side), would be precisely detected when it was on the side of the ‘visible’ utterance and likely to be misperceived when it was on the side of the ‘masked’ utterance.

If so, this raises the question: What does this ‘congruency’ reflect? A pair of congruent audio-visual syllables not only elicits a ‘congruent’ percept, but it also exerts greater temporal synchrony as well as other kinds of physical compatibility, relative to incongruent syllable pairs. If a strong ventriloquism effect emerges, it might simply be due to such a physical property. To dissociate strictly these sensory (physical) factors, such as synchrony, from cognitive (perceptual) factors, such as syllable congruency, we introduced a unique condition in which incongruent audio-visual speech information could be perceived illusorily as congruent. In the well-known McGurk effect, a voice speaking a labial syllable, such as /pa/, is dubbed onto a silent video of a face uttering a different non-labial syllable, such as /ka/. Typically, this results in an illusory auditory perception, for instance, /ta/ which is a fusion of the incongruent speech information (McGurk & MacDonald, 1976). Participants are often not aware of audio-visual discrepancy under such an illusion, so their bimodal percept is rather ‘congruent’. We also examined whether this kind of illusory perceptual congruency of audio-visual input by itself can affect the ventriloquism phenomenon, even without strictly physically congruent stimuli.

In sum, using visible and masked facial gestures, we varied both congruency of a voice with the ‘visible’ facial gesture (congruent versus incongruent) and spatial location of a sound source (visible-side versus masked-side). We hypothesized that participants’ auditory localization would be attracted toward the ‘visible’ utterance because of its

physical saliency, thereby facilitating localization performance in the visible-side condition and disturbed performance in the masked-side condition simultaneously. In addition, if audio-visual congruency affects ventriloquism, the performance should be further facilitated or disturbed under a presentation of a voice and a congruent 'visible' utterance, compared to an incongruent 'visible' utterance.

Methods

Participants

Nine male and nine female volunteers participated. All were native Japanese speakers, and reported normal or corrected-to-normal vision and normal hearing ability. Their ages ranged from 20 to 36 years (average 22.3 years) and they were naïve as to the purpose of the experiment.

Stimuli

Visual stimuli were two videos of human faces. In one video, dynamic facial gestures conveyed /ka/ or /pa/ ('visible' utterance). In another video, the visibility of identical facial features was attenuated by a white mask that occluded lip movements; remaining facial gestures were marginal, lacking phonological information ('masked' utterance) (Fig. 1). The auditory stimulus was a one channel recording of a person's voiced utterance of /pa/ or /ka/. Both syllable utterances were extracted from videos of the same monolingual Japanese woman repeating these syllables. Repeated utterances were videotaped with a digital video camera (SONY HDR-UX7), and separated to several audio files and silent video files; each file contained only one audio or visual utterance. Auditory stimuli were digitized at 48 kHz in 16 bit; the average duration of the auditory signal was approximately 160 ms. Visual stimuli were digitized at 29 frames/s in 720 × 480 pixels, and clipped as 300 × 340 pixel moving images showing the speaker's face. The average duration of a single visible syllable was 759 ms. Each sound and video file was selected from different original videotapes on the basis of their intelligibility; they were then combined to create different presentation types. The averaged duration of each presentation was 1,500 ms. Visual stimuli were projected (Canon Power Projector LV-7255) respectively in 37 × 40 cm rectangles to the left and the right of a white screen. These two projections were separated by a strip of 36 cm; this resulted in a 110 × 40 cm rectangle in total. Participants were seated 190 cm from the screen; the rectangular width subtended to 32.29°visual angle. The distance between the midpoints of the lips on the two faces (although one was masked) was 75 cm (subtending to a 22.33°visual angle).

Fig. 1 Bilateral visual stimuli - videos of a "visible" utterance and "masked" utterance

Auditory stimuli were emitted from one of two loudspeakers (BOSE 101VM), located at the bottom left and right of the screen in front of the participant. Center points of loudspeakers were separated by 25 cm and hidden by a cotton sound-penetrable sheet. Each auditory signal was presented at 78 dB SPL, with white noise from both loudspeakers, in a SN ratio of –2 dB measured at the participants' location. Noise was added in order to reduce auditory intelligibility and insure that participants relied upon audio-visual stimuli to perceive speech (Sekiyama & Tohkura, 1991).

Conditions

Four presentation conditions resulted from crossing two levels of the sound source variable (visible-side, masked-side) with two levels of audio-visual congruency variables (congruent, incongruent), as shown in Fig. 2. In every condition, the horizontal arrangement of two visual stimuli ('visible' utterance left and 'masked' utterance right, or the inverse order) was counterbalanced. The sound source variable was used to manipulate whether the auditory stimulus was presented at the same side as the 'visible' utterance (visible-side) or the side of the 'masked' utterance (masked-side). The congruency variable was used to manipulate whether the 'visible' utterance and the voice expressed the same syllable (congruent) or not (incongruent).

In this design, note that combining an auditory labial, /pa/ with an incongruent /ka/ video, a non-labial, often yields the illusory auditory percept of /ta/ (so-called McGurk 'fusion response'). Simply viewing a visual utterance in the center of a visual field is not necessary for this kind of illusion (Paré, Richler, Ten Hove, & Munhall, 2003). When this kind of illusion occurs, participants are often not aware of the incompatibility between auditory and visual speech syllables (Sekiyama & Tohkura, 1994), because the visual /ka/ (non-labial) and the perceived voice, /ta/ (non-labial) is similar with respect to

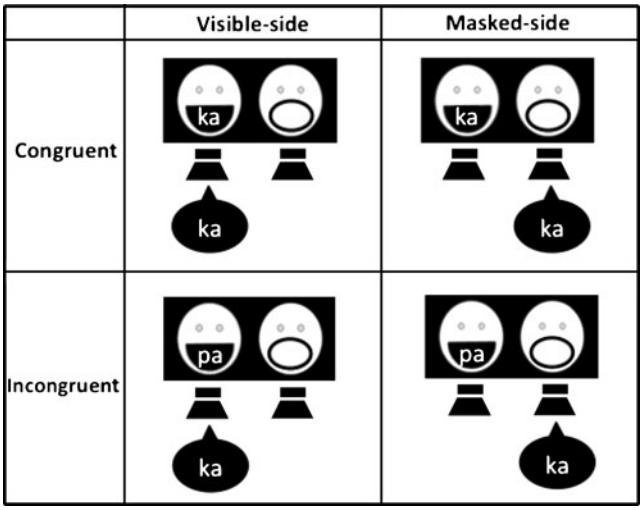

Fig. 2 Examples of audio-visual stimuli in each of the four types of presentations, with auditory /ka/

the place of articulation. So, in this situation the audio-visual speech information turns to be illusorily 'congruent'. However, the reverse combination of an auditory non-labial, /ka/ on a visual labial, /pa/ is known to hardly induce such an illusory fusion, resulting in a correct auditory percept, especially in Japanese participants. This is thought to be due to either the sound structure of Japanese or their nature of less reliance on visual information in speech perception (Sekiyama, 1997). In such a lack of audio-visual speech fusion, they are often aware of the discrepancy in audio-visual speech information, because the visual /pa/ (labial) and the correctly perceived voice, /ka/ (non-labial) is different with respect to the place of articulation. So, incongruent bimodal speech syllables would simply be perceived as incongruent. In brief, when incongruent audio-visual stimuli are presented, it is possible to produce an illusorily 'congruent' percept to visual information only with the auditory /pa/ and not with the auditory /ka/.

Procedure

The experiment was conducted in a darkened soundproof room. All presentations were controlled by a computer (EPSON Endeavor MT7500). On each trial, two faces were bilaterally presented with a single synchronized voice. All stimuli occurred 50 ms after a 100-ms fixation point (+) at a position equally distant from the center point between the two faces. Participants were instructed to continue to gaze at central area pre-specified by the fixation point and discriminate from which side of the area left and right the sound was heard. Then, participants responded by pressing one of two buttons corresponding to the left and the right on a keyboard; all responses were recorded on a computer.

Prior to testing, participants received eight practice trials. Practice trials were identical to those of the main experiment. Participants were not given any feedback.

Design

Each of four conditions (congruency × loudspeaker disposition) occurred 20 times for each of two auditory stimuli (/ka/ and /pa/). The absolute position of the sound-emitting loudspeaker (left or right) was counterbalanced over trials. Also, presentation order of all stimulus combinations was randomized over a total of 160 trials within a single test block.

Results

Previous studies concerning effects of spatial factors on audio-visual localization and speech perception have reported robust symmetry for left and right sound sources (Bertelson et al., 1994; Colin et al., 2001). Therefore, we collapsed over levels of this counter-balanced factor in the following analyses.

The percentages of responses in which the participants correctly discriminated the left or right sound source (Fig. 3) were analyzed using a two factor repeated measures ANOVA (congruency × sound source) separately for auditory-/ka/ trials and auditory-/pa/ trials, respectively. In a pilot experiment, in which we presented these auditory syllables without visual stimulation (using 18 volunteers who did not participate in the main experiment), percent correct localizations differed significantly depending on the syllable (77.8% for /ka/ and 86.7% for /pa/). Therefore, we analyzed auditory /ka/ and /pa/ trials separately.

In both auditory-/ka/ trials and auditory-/pa/ trials, the main effect of sound source was significant, $F(1, 17) = 10.30$, $p < 0.01$, partial $\eta^2 = .38$, and $F(1, 17) = 33.19$, $p < 0.01$, partial $\eta^2 = .66$. In both trial types, localization performance was better in the visible-side condition than in the masked-side condition. The congruency × sound source interaction was also significant for both /pa/ and /ka/ trials, $F(1, 17) = 8.72$, $p < 0.01$, partial $\eta^2 = .34$, and $F(1, 17) = 13.92$, $p < 0.01$, partial $\eta^2 = .45$; the performance in congruent conditions was better in the visible-side condition whereas the reverse was true with incongruent conditions where performance was better in the masked-side condition. This interaction differed somewhat as function of /ka/ and /pa/ trials, however. Participants performed better in the visible-side than in the masked-side condition with congruent audio-visual syllables in both auditory /ka/ and /pa/ trials, and incongruent syllables in auditory /pa/. However, their performance as a function of sound source did not differ significantly with incongruent syllables in auditory /ka/ trials.

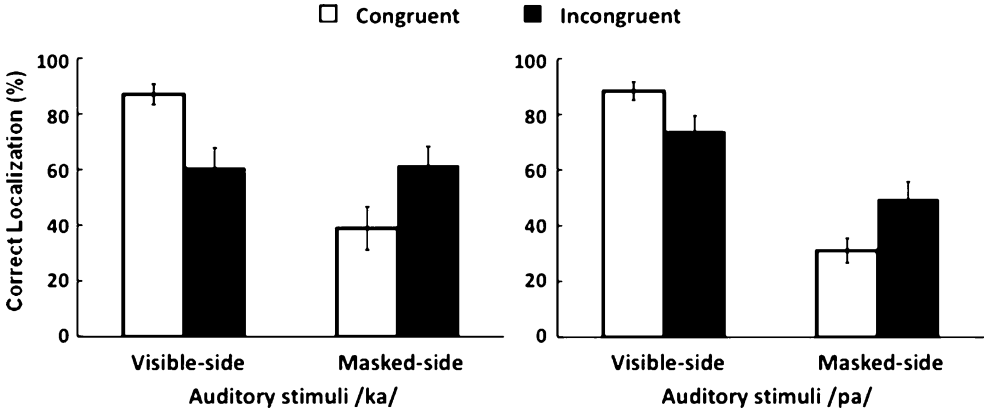

Fig. 3 Mean proportions of responses in which participants correctly reported the sound-source left or right. Error bars correspond to the standard errors of the means

Discussion

We investigated whether audio-visual congruency using spoken syllables affects ventriloquism when one auditory stimulus is simultaneously paired with two visual stimuli. One of two faces provided a 'visible' utterance whereas the other did not, although both faces featured gestures synchronized to the auditory stimulus. The 'visible' utterance was either congruent or incongruent with the presented voice. We observed a substantial 'saliency effect,' manifested as biased localization favoring the more salient of the two visual stimuli, consistent with other reports (Bertelson et al, 2000; Radeau, 1985). Of the two visual stimuli, the more salient, 'visible' utterance, stimulus provided more dynamic information than the 'masked' one. Participants performed better when the sound source corresponded with the side of the 'visible' utterance, and poorly in the 'masked' condition, suggesting a perceptual shift of the sound source toward the 'visible' utterance that facilitated performance.

If inter-sensory congruency in audio-visual stimuli gives rise to a strong ventriloquism effect, a congruent combination of a voice and a 'visible' utterance should strengthen the 'saliency effect.' With congruent, but not incongruent audio-visual utterances, we would observe greater facilitation in localization performance in the visible-side condition, and greater disturbance in the masked-side condition. In the present study, we obtained these expected results in both auditory /ka/ and /pa/ trials.

One difference between in auditory /ka/ and /pa/ trials was evident. It relates to the possibility of illusory fusion of incongruent audio-visual speech information. In particular, Japanese participants perceive incongruent audio-visual stimuli as illusorily congruent only in auditory /pa/, not in auditory /ka/ trials. If this kind of illusory congruency of audio-visual speech information in perception works in a fashion similar to true stimulus congruency, we would expect to find a larger localization bias in the condition where an incongruent 'visible' face was presented with auditory /pa/ trials than with /ka/ trials. On auditory /ka/ trials, participants' did not differ in performance in the incongruent condition as a function of sound source (visible-side or masked-side); this suggests an absence of localization bias toward the 'visible' utterance, i.e., a null 'saliency effect.' In contrast, on auditory /pa/ trials performance was significantly better in the visible-side condition than in the masked-side condition, suggesting a substantial 'saliency effect.' Ventriloquism was also affected by the audio-visual illusory congruency in perception, not only by the true congruency in stimuli.

In order to verify that participants perceived the expected congruency and the incompatibility with audio-visual speech information in each condition, we performed an additional experiment. Eighteen new participants observed the same audio-visual presentation as in the main experiment; they had to report the perceived location of the sound source (as in the main experiment) and also identify a heard syllable on each of 160 trials (response order was counterbalanced) simultaneously. Participants generally responded correctly for auditory syllables in congruent conditions for both auditory /ka/ and /pa/ trials (correct response percentages were 99.2 and 94.3 in auditory /ka/ and /pa/ trials, respectively). However, in incongruent conditions, only the auditory /pa/ syllables were frequently misperceived (mainly as /ta/), and auditory /ka/ syllables were primarily perceived correctly (correct response percentages were 33.5 and 99.5 for auditory /pa/ and /ka/ trials, respectively). Their localization performance was statistically identical to results in the main experiment.

One clear inference from these results is that audio-visual congruency in speech syllables affects the way visual information attracts participants' auditory localization. Congruent visual stimuli elicited a greater localization bias than incongruent stimuli. It was not only physical aspects of audio-visual congruency that mattered but even an illusory congruency in perception had an impact on participants' auditory spatial judgments. This demonstrates a contribution of cognitive factors; it suggests that some top-down, or higher-order, processing affects audio-visual spatial judgments, which have previously been considered to be essentially driven by bottom-up, or lower-order, processing based on purely physical features of multisensory environments.

An important point to be discussed is whether these results genuinely reflect the role of audio-visual congruency. For example, perhaps in the two visual stimuli presented, the 'visible' utterance attracts much attention than does the 'masked' one, especially when it is congruent with a presented voice. However, it has been demonstrated that attention plays no role in ventriloquism (Bertelson et al., 2000; Vroomen, Bertelson, & de Gelder, 2001). In these studies the direction of participants' attention was manipulated either overtly or covertly, but the direction of localization bias was determined solely by stimulus saliency (i.e., the size of bilateral visual stimuli), independent of attentional direction. Therefore the location to which they had consciously attended or the feature that automatically captured their attention cannot be assumed to determine participants' responses in the present results.

We suggest a possible relationship between audio-visual localization and speech perception, reflected in ventriloquism and the McGurk effect respectively. Previous studies have reported dissociable mechanisms under these effects, in that the perceived location of the sound source is regulated only by sensory factors such as temporal or spatial proximity between a pair of audio-visual stimuli, independent of any cognitive factor such as ecological association or congruency. However, here we discovered that participants' auditory localization bias can be affected by the relationship between the identities of stimuli or their congruency in syllables. The outcome of audio-visual speech processing has regulated participants' audio-visual spatial judgments. These results contradict previous reports, and this may be due to our use of multiple stimuli within one sensory modality. Our cognitive mechanisms are highly adaptive to the everyday environment, where we can recognize multiple signals relevant to one particular event correctly and effectively. This complex nature of the multisensory system should be a subject for further research.

References

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*, 482–489.

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62*, 321–332.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of the International Conference on Spoken Language Processing, 2*, 559–562. Yokohama.

Brungart, D. S., Kordik, A. J., & Simpson, B. D. (2005). Audio and visual cues in a two-talker divided attention speech-monitoring task. Human factors. *The Journal of the Human Factors and Ergonomics Society, 47*, 562–573. doi:10.1518/001872005774860023

Colin, C., Radeau, M., Deltenre, P., & Morais, J. (2001). Rules of intersensory integration in spatial scene analysis and speech reading. *Psychologica Belgica, 41*, 131–144.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381*, 66–68. doi:10.1038/381066a0

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills, 37*, 967–979.

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

Paré, M., Richler, R. C., Ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, 65*, 553–567.

Radeau, M. (1985). Signal intensity, task context, and auditory-visual interaction. *Perception, 14*, 571–577.

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics, 22*, 137–146.

Roseboom, W., Nishida, N., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9*(12), 1–8. doi:10.1167/9.12.4

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*, 73–80.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America, 90*, 1797–1805.

Sekiyama, K., & Tohkura, Y. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E), 15*, 143–158.

Vroomen, J., & de Gelder, B. (2004). Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In G. Calvert, C. Spence, & B. Stein (Eds.), *The handbook of multisensory processes* (pp. 141–150). Cambridge MA: MIT Press.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*, 651–659.