Facilitation of multisensory integration by the "unity effect" reveals that speech is special

Argiro Vatakis
Crossmodal Research Laboratory,
Department of Experimental Psychology,
University of Oxford, Oxford, UK

Asif A. Ghazanfar
Neuroscience Institute and Department of Psychology,
Princeton University, Princeton, NJ, USA

Charles Spence
Crossmodal Research Laboratory,
Department of Experimental Psychology,
University of Oxford, Oxford, UK

Whenever two or more sensory inputs are highly consistent in one or more dimension(s), observers will be more likely to perceive them as a single multisensory event rather than as separate unimodal events. For audiovisual speech, but not for other noncommunicative events, participants exhibit a "unity effect," whereby they are less sensitive to temporal asynchrony (i.e., that are more likely to bind the multisensory signals together) for matched (than for mismatched) speech events. This finding suggests that the modulation of multisensory integration by the unity effect in humans may be specific to speech. To test this hypothesis directly, we investigated whether the unity effect would also influence the multisensory integration of vocalizations from another primate species, the rhesus monkey. Human participants made temporal order judgments for both matched and mismatched audiovisual stimuli presented at a range of stimulus-onset asynchronies. The unity effect was examined with (1) a single call-type across two different monkeys, (2) two different call-types from the same monkey, (3) human versus monkey "cooing," and (4) speech sounds produced by a male and a female human. The results show that the unity effect only influenced participants' performance for the speech stimuli; no effect was observed for monkey vocalizations or for the human imitations of monkey calls. These findings suggest that the facilitation of multisensory integration by the unity effect is specific to human speech signals.

Keywords: cross-modal, multisensory, temporal perception, monkey vocalizations

Citation: Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9):14, 1–11, http://journalofvision.org/8/9/14/, doi:10.1167/8.9.14.

Introduction

People are continually exposed to situations in which two or more simultaneous unisensory inputs specify common environmental events or actions. The integration of multiple sensory inputs into unified multisensory percepts has been demonstrated by research that has used both simple stimuli (Bertelson & de Gelder, 2004; Thomas, 1941; Witkin, Wapner, & Leventhal, 1952) as well as more complex stimuli, such as speech (Easton & Basala, 1982; Jackson, 1953; Walker, Bruce, & O'Malley, 1995). It is generally thought that if sensory inputs are highly consistent in terms of their low-level stimulus dimension(s), such as their spatial location or temporal patterning, observers will be more likely to attribute them to a single multisensory event rather than treating them as multiple separate unimodal events. "Higher-level" (i.e., more cognitive) factors relating to a perceiver's impression that the incoming sensory signals are somehow consistent, and thus ought to "go together," have also been shown to influence multisensory integration. Such results concerning the "unity assumption" are based, at least in part, on the consistency of the information available to each sensory modality (Spence, 2007; Vatakis & Spence, 2007, 2008; Welch, 1999a, 1999b; Welch & Warren, 1980; note that higher-level factors also include phenomena such as perceptual grouping and phenomenal causality; Guski & Troje, 2003; Radeau & Bertelson, 1987).

While most previous investigations of the unity assumption have focused on the role of spatiotemporal variables in the integration of audiovisual speech and nonspeech stimuli (such as studies that have shown that auditory stimuli are typically mislocalized toward visual stimuli, provided that they are presented at approximately the same time; Radeau & Bertelson, 1987), Vatakis and Spence (2007) recently demonstrated that the unity effect (whereby a participant may perceive a multisensory event as unified due to the low-level congruence of two sensory signals) modulates the multisensory integration of audiovisual speech stimuli using a task in which the participants made temporal order judgments (TOJs). The participants

doi: 10.1167/8.9.14 Received November 19, 2007; published July 29, 2008 ISSN 1534-7362 © ARVO

were presented with a series of video clips of speakers uttering speech sounds (i.e., a series of different syllables and words) with a range of different stimulus onset asynchronies (SOAs; note here that the SOA manipulation involved varying the degree of temporal asynchrony between the auditory and the visual streams, thus its manipulation may also have weakened the perceptual unity of a given multisensory event). The auditory- and visual-speech signals were either gender matched (e.g., a female face presented together with the matching female voice) or else gender mismatched (e.g., a female face presented together with a male voice). Vatakis and Spence hypothesized that if the unity effect influences multisensory integration in humans, participants should find it harder to determine whether the visual- or the auditory-speech signals had been presented first when the two stimuli referred to the same underlying perceptual event (matched condition) than when they did not (mismatched condition). In support of this prediction, participants were significantly more sensitive to the temporal order (or alignment) of the auditory- and visual-speech signals in the mismatched conditions than in the matched conditions. This suggests less multisensory integration of the stimuli in the mismatched condition (Vatakis & Spence, 2007; for converging findings from earlier studies of speech perception, see Easton & Basala, 1982; Walker et al., 1995).

It is, however, important to note that speech represents a very familiar stimulus to humans. Indeed, it has been argued elsewhere that it may potentially belong to a "special" class of sensory events (e.g., Bernstein, Auer, & Moore, 2004; Jones & Jarick, 2006; Liberman & Mattingly, 1985; Massaro, 2004; Munhall & Vatikiotis-Bateson, 2004; Tuomainen, Andersen, Tiippana, & Sams, 2005). In order to investigate whether the unity effect would influence the multisensory integration of ecologically valid non-speech stimuli, the participants in a subsequent study were presented with video clips of object actions or musical events that could either be matched (e.g., the sight of a note being played on a piano together with the corresponding sound) or mismatched (e.g., the sight of a note being played on a piano together with the sound of the same note being played on a guitar; Vatakis & Spence, 2008). In contrast to the findings with speech stimuli (Vatakis & Spence, 2007), participants' sensitivity to the temporal alignment of the auditory and visual signals of the object action and musical events did not differ as a function of the matching versus mismatching of the stimuli (Vatakis & Spence, 2008). This null result suggests that the unity effect does not influence people's temporal perception of realistic audiovisual non-speech stimuli (Vatakis & Spence, 2008; cf. Radeau & Bertelson, 1977).

Taken together, the results of these two studies suggest that the unity effect in human participants may be specific to speech stimuli or, more generally, to vocalizations. A strong test of these hypotheses consists of presenting non-speech *vocal* stimuli in an identical experimental paradigm. Monkey calls represent an ideal non-speech vocal stimulus for a number of reasons. First, the majority of human participants will have had little or no prior experience with these stimuli. Second, although monkeys cannot produce the same range of vocal sounds as humans, their vocal production mechanisms are nearly identical to those of humans (Fitch & Hauser, 1995). Third, as a result of the similar mechanisms of production, rhesus monkey calls have spectral structure (formants) that is similar to that of human vowel sounds (Fitch, 1997). Fourth, each type of rhesus monkey call is produced with a unique facial configuration (Hauser, Evans, & Marler, 1993) that can be matched to the auditory component by both monkeys (Ghazanfar & Logothetis, 2003) and very young human infants (Lewkowicz & Ghazanfar, 2006). Finally, in the temporal domain, mouth movements occur before the associated voiced component of a given monkey call which represents another parallel between monkey calls and human speech (Ghazanfar, Maier, Hoffmann, & Logothetis, 2005).

If the unity effect is limited to vocalizations then one would predict that human participants should find it harder to determine the temporal alignment of the visual and the auditory signals when the monkey calls are matched than when they are mismatched. Such an outcome would provide the first empirical demonstration that the unity effect can also facilitate the integration of audiovisual non-speech vocalizations.

Materials and methods

Participants

All of the participants (college students; Experiment 1: $N = 14$, 7 female; Experiment 2: $N = 10$, 8 female; Experiment 3: $N = 18$, 9 female; Experiment 4: $N = 15$, 8 female) were naive as to the purpose of the study, and all reported having normal hearing and normal or corrected-to-normal visual acuity. Each experiment was performed in accordance with the ethical standards laid down in the 1990 Declaration of Helsinki. Each experiment lasted 40 min.

Apparatus and materials

The experiments were conducted in a dark sound-attenuated testing booth. The visual stimuli were presented on a 43.18-cm TFT colour LCD monitor (SXGA 1240 × 1024-pixel resolution; 60-Hz refresh rate), placed at eye level, 68 cm from the participant. The auditory stimuli were presented by means of two Packard Bell Flat Panel 050 PC loudspeakers; one placed 25.4 cm to either side of the center of the monitor (i.e., the auditory- and visual-speech stimuli appeared from the same spatial location). The audiovisual stimuli consisted of black and white video clips presented on a black background using Presentation (Version 10.0; Neurobehavioral Systems Inc.).

The video clips of vocalizing monkeys and humans were processed using Adobe Premiere 6.0 (300 × 280-pixel, Cinepak Codec video compression, 16-bit audio sample size, 30 frames/s). The monkey video clips consisted of the following: (a) in Experiment 1, four different video clips of two adult male rhesus monkeys (*Macaca mulatta,* Monkey A and Monkey B, visible from the chin to the upper part of the head) vocalizing a "coo" and a "grunt" call; (b) in Experiment 2, two different video clips of one adult male rhesus monkey (a different individual from Experiment 1) vocalizing a "coo" and a "threat" call (for further details regarding these recordings, see Ghazanfar et al., 2005; all of the clips were 834 ms long; the event was 534 ms, long beginning with the last frame before the first mouth movement to the last frame after the end of vocalization; synchrony was taken to be represented by the initial recording of the clip). "Coo" calls are long tonal signals produced with a wide lip separation and lip protrusion, while "grunts" and "threats" are short, noisy, pulsatile signals produced by a wide lip separation but with limited if any lip protrusion (Hauser et al., 1993). The human video clips consisted of humans producing a "coo-like" sound and speech sounds. For the human coo call, a male was asked to watch the video clip of the monkey cooing and to try to imitate as closely as possible the visual and auditory signals. This ensured that the two clips of the monkey and human were as similar as possible. The speech sounds consisted of closeup views of the faces (visible from the chin to the top of the head) of a British male and female uttering the speech sound /a/ (both clips were 385 ms in duration).

In order to achieve accurate synchronization of the dubbed video clips, each original clip was re-encoded using XviD codec (single pass, quality mode of 100%). Using the multi-track setting in Adobe Premiere, the visual and the auditory components of the to-be-dubbed videos were aligned based on the peak auditory signals of the two video clips and the visual frames of the first movement of the mouth. A final frame-by-frame inspection of the video clips was performed in order to ensure the correct temporal alignment of the auditory and the visual signals. All these steps were followed so that no delays were added to the video clips due to the video processing.

At the beginning and end of each video clip, a still image of the particular stimulus was presented, and background acoustic noise (which was part of the original recording) was presented for a variable duration. The duration of the image and noise was unequal with the difference in their durations being equivalent to the particular SOA tested (values reported below) in each condition. This aspect of the design ensured that the auditory and visual signals always started at the same time, thus avoiding cuing the participants as to the nature of the audiovisual delay with which the auditory and the visual dynamic events were being presented. A 33.33-ms cross-fade was added between the still image and the video clip in order to achieve a smooth transition at the start and end of each video clip. The participants responded using a standard computer mouse, which they held with both hands, using their right thumb for "visual-lip movement first" responses and their left thumb for "auditory-speech/call first" responses (or vice versa, the response buttons were counterbalanced across participants).

Stimulus design

Nine SOAs between the auditory and the visual stimuli were used: ±300, ±200, ±133, ±66, and 0 ms (negative values indicate that the auditory signal was presented first, whereas positive values indicate that the visual signal was presented first). Before starting the experiment, the matched and mismatched video clips were shown to the participants in order to familiarize them with each monkey or human individual/utterance. The participants completed one block of 8 practice trials before the main experimental session in order to familiarize themselves with the task. The practice block was followed by 5 blocks of 144 experimental trials, consisting of two presentations of each of the 8 video clips at each of the 9 SOAs per block of trials. The various SOAs were presented randomly within each block of trials using the method of constant stimuli.

Procedure

The participants performed a TOJ task. They were informed that they would be presented with a series of video clips (matched and mismatched clips of the visual and the auditory signals of a pair of stimuli). On each trial, they had to decide whether the auditory-speech/call or visual-lip movement occurred first. (Note here that the participants could have made their judgments either on the basis of the audio-visual event onsets/offsets or else on the basis of the temporal misalignment of the visual lip movement and the corresponding auditory signal.) They were informed that they would sometimes find this task difficult, in which case they should make an informed guess as to the order of presentation of the two signals. The participants did not have to wait until the video clip had finished before making their response, but a response had to be made before the experiment advanced to the next trial.

Analysis

The proportions of "visual-lip movement first" responses were converted to their equivalent *z*-scores under the assumption of a cumulative normal distribution (Finney, 1964). The data from the 7 intermediate SOAs (±200, ±133, ±66, and 0 ms) were used to calculate best-fitting straight lines for each participant for each condition, which, in turn, were used to derive values for the slope and intercept. The r^2 values reflect the correlation between the SOAs and the proportion of "vision-first" responses and hence provide an estimate of the goodness of the data fits. The calculation of the r^2 values for each condition in all four experiments revealed a significant goodness of fit. The ±300 ms data points were excluded from this computation due to the fact that most participants performed near-perfectly at this interval and so did not provide any significant information regarding our experimental manipulations (cf. Spence, Shore, & Klein, 2001; Vatakis & Spence, 2007, 2008, for a similar approach). The slope and the intercept values were used to calculate the JND (JND = 0.675/slope; since ±0.675 represents the 75% and 25% points on the cumulative normal distribution) and the point of subjective simultaneity (PSS = −intercept / slope) values (see Coren, Ward, & Enns, 2004). The PSS data for all experiments are not reported here, given that the "unity effect" has previously been shown not to have any reliable effect on the PSS (cf. Vatakis & Spence, 2007, 2008). However, the statistical outcomes[1] are provided for the sake of completeness.

The JND provides a measure of the participants' sensitivity to the temporal order or alignment of two sensory signals. In particular, it provides a measure of the interval needed in order for participants to judge the temporal order or alignment of the two signals correctly on 75% of the trials. For all of the analyses reported here, Bonferroni-corrected t tests (where $p < .05$ prior to correction) were used for all post hoc comparisons.

Results

Experiment 1 tested whether the unity effect influences the human perception of monkey vocalizations when the *identity* of the caller and the nature of the stimulus is the "unifying" variable. The stimuli consisted of one call type (a "coo" or a "grunt") produced by two different rhesus monkeys (Monkeys A and B) that could either be matched (e.g., matched visual and auditory "coo" vocalization from Monkey A) or mismatched (e.g., the visual "coo" from Monkey A and the "coo" sound from Monkey B; Figure 1A). Figure 2A represents the sigmoid fits of participants' mean responses for each condition at the various SOAs. The JND data for each of the matched and

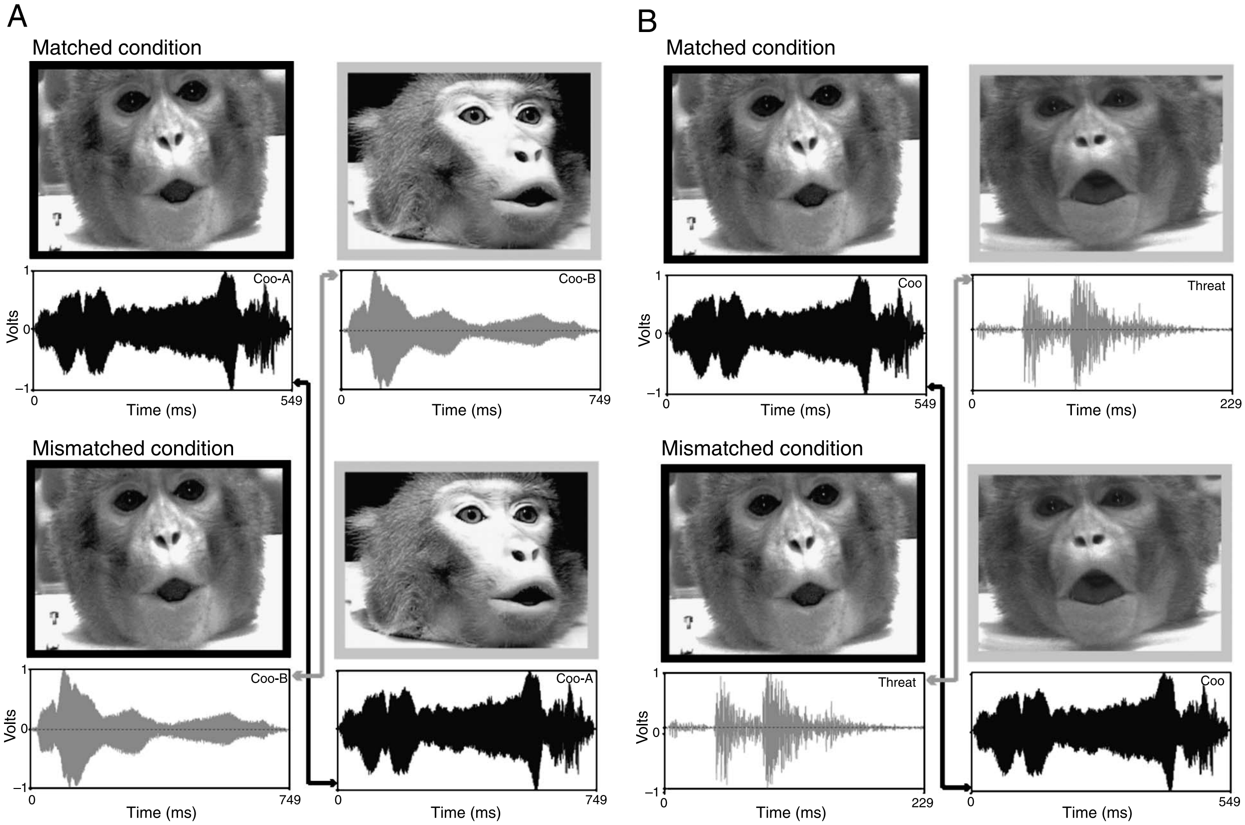

Figure 1. A schematic illustration of the: (A) matched and mismatched "coo" vocalization video clips of the two rhesus monkeys used in Experiment 1 and (B) matched and mismatched "coo" and "threat" vocalization video clips of the rhesus monkey used in Experiment 2.

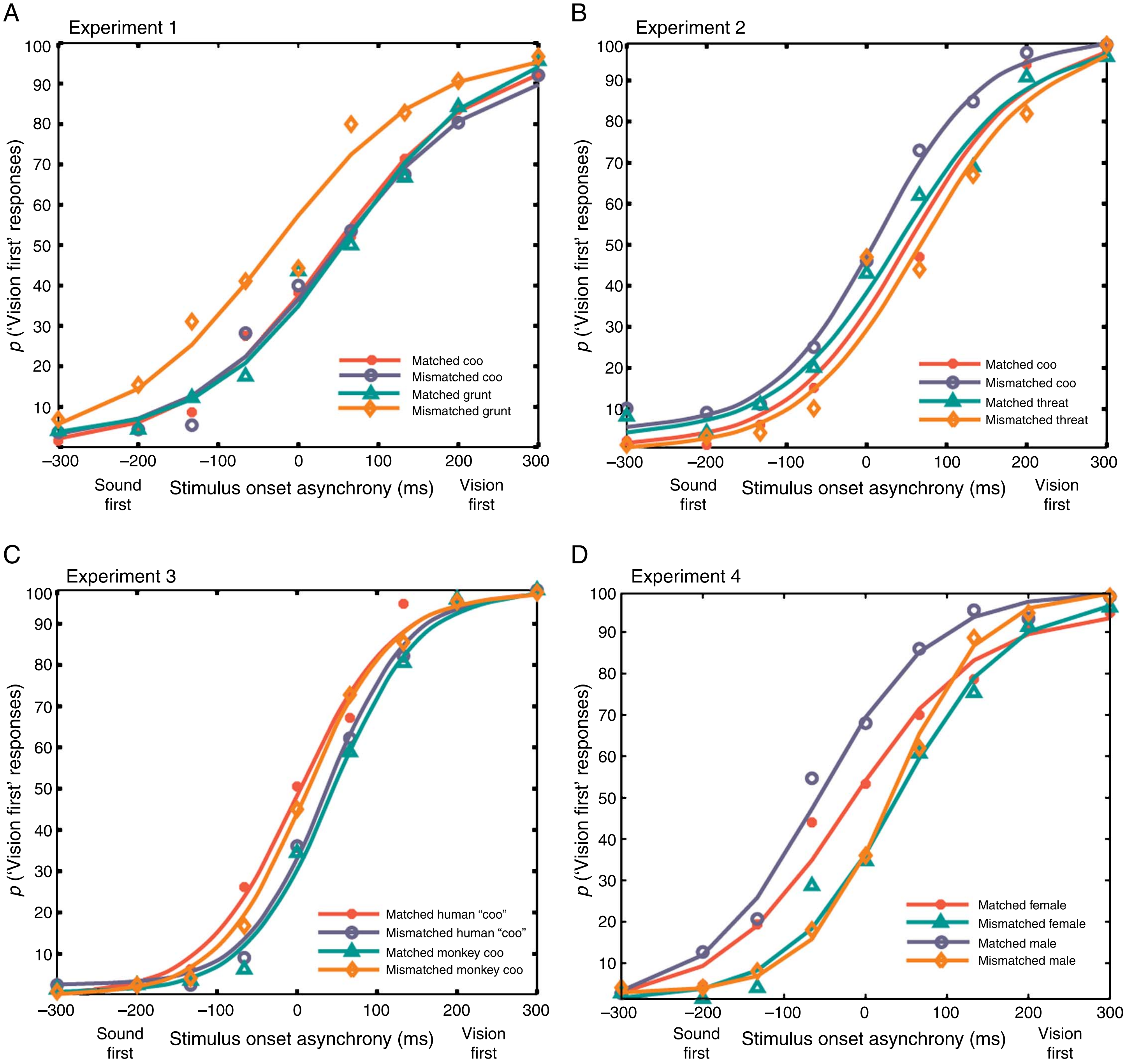

Figure 2. Mean percentage of "vision-first" responses plotted as a function of the stimulus onset asynchrony (SOA) in Experiments 1–4. (A) Matched: audio-visual "coo" and "grunt" vocalizations from Monkeys A and B; mismatched: visual "coo"/"grunt" from Monkey A and the auditory "coo"/"grunt" from Monkey B; (B) matched: audio-visual "coo" and "threat" vocalizations uttered by the same monkey; mismatched: visual "coo"/"threat" and auditory "threat"/"coo" from the same monkey; (C) matched: audio-visual monkey and human cooing; mismatched: visual human/monkey cooing and auditory monkey/human cooing; (D) matched: audio-visual /a/ from a male and from a female speaker; mismatched: visual male/female /a/ and the auditory female/male /a/. The functions represent the sigmoid fits of participants' mean responses for each condition at the various SOAs tested.

mismatched vocalization events (derived from the slopes of the functions) were analyzed using repeated measures analysis of variance (ANOVA) with the factors of face–voice match, type of vocalization (coo or grunt), and monkey seen (Monkey A or Monkey B). The analysis revealed that the participants' sensitivity to the temporal alignment of the auditory and the visual signals was unaffected by whether the monkey's face and auditory vocalization matched (mean = 88 ms) versus mismatched (M = 93 ms) in terms of the identity of the

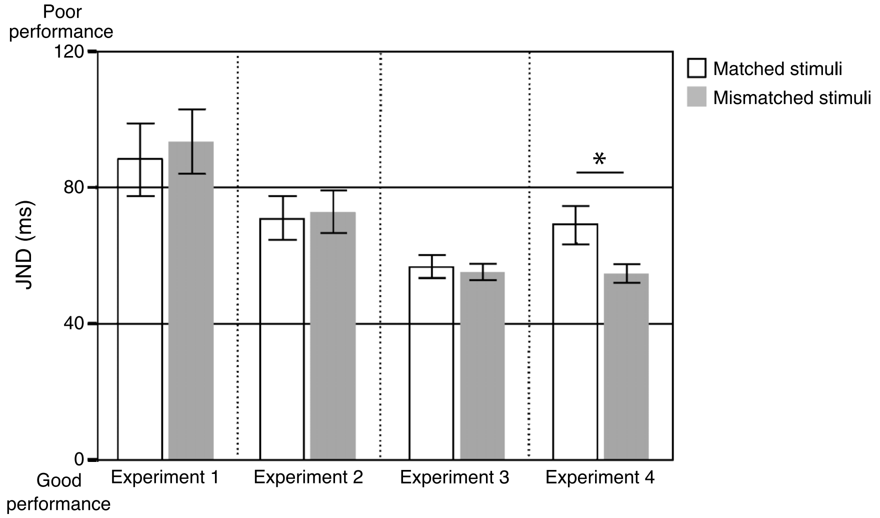

Figure 3. (A) Average JNDs for the matched and the mismatched audiovisual stimuli (monkey/human vocalizations and human speech) presented in Experiments 1–4. The error bars represent the standard errors of the means. Significant differences ($p < .05$) are highlighted by an asterisk.

monkey ($F(1,13) < 1$, ns) (Figure 3). None of the other effects were significant either. Thus, the results of Experiment 1 showed no difference in participants' sensitivity to the temporal alignment of the auditory and the visual vocalization stimuli as a function of whether they came from the same monkey or not. The results therefore fail to provide any support for the unity effect in the case of caller identity changes across monkey vocalizations.

It is possible that the null effect of matching reported in Experiment 1 was due to the participants simply being unaware of the mismatch between the auditory and the visual signals (cf. Saldaña & Rosenblum, 1993; Vatakis & Spence, 2007). We therefore conducted a follow-up study that was identical to Experiment 1 with the sole exception that the participants now indicated whether the auditory and visual signals matched (rather than indicating which sensory modality signal appeared to have been presented first). The results showed that participants ($N = 5$) correctly discriminated between the matched and mismatched monkey vocalization video clips on 84% of the trials. Thus, the null effect of matching on the JND in Experiment 1 cannot simply be accounted for by any lack of awareness on the part of our participants with regard to the presence of a discrepancy between what they were seeing and hearing in the mismatched video clips.

In Experiment 2, two different vocalizations (a "coo" and a "threat" call) from the same monkey were used as stimuli in order to confirm (by replicating the findings of Experiment 1) that the unity effect has no behavioral consequences for the temporal integration of audiovisual vocalizations. The stimuli consisted of a "coo" and "threat" vocalization made by the same rhesus monkey that could either be matched (e.g., matched visual and auditory "coo" vocalization) or mismatched (e.g., the visual "coo" and the "threat" sound from the same monkey; Figure 1B). If the pattern of results in this experiment were found to be similar to that reported in Experiment 1, then it would provide a replication of the previously observed absence of the unity effect.

Figure 2B shows the sigmoid fits of participants' mean responses for each condition tested in Experiment 2 at the various SOAs. The analysis of the JND data revealed no significant difference in participants' sensitivity to the temporal alignment of the auditory and the visual signals in the case of matched ($M = 71$ ms) versus mismatched video clips ($M = 72$ ms; see Figure 3) ($F(1,9) < 1$, ns). There was no significant main effect of the Type of vocalization ("coo" or "threat") ($F(1,9) = 2.12$, $p = .19$) nor any interaction between face–voice match and type of vocalization ($F(1,9) < 1$, ns). Thus, the results of Experiment 2 failed to provide any support for the unity effect for different calls from the same monkey. These results are therefore entirely consistent with the null effects of matching reported in Experiment 1.

Given the possibility that the null effect of matching was being driven by the participants simply being unaware of the audiovisual mismatch of the video clips presented, we conducted another study to ensure that the matched videos could be distinguished from the mismatched videos. This follow-up study was identical to Experiment 2 with the sole exception that the participants ($N = 5$) had to indicate whether the auditory and the visual vocalization signals matched or not. The participants were able to perform this task very accurately ($M = 95\%$ correct), thus showing that the audiovisual discrepancy on trials where the auditory and the visual signals belonged to different vocalizations from the same monkey was easily detectable.

In order to examine whether speech signals are more strongly integrated than nonspeech signals, we compared the data obtained from Experiment 2 with those of a similar experiment conducted using videos of humans in our previous study (see Experiment 3 in Vatakis & Spence, 2007, where we presented the same human individual uttering two different speech tokens). A mixed ANOVA with the within-subjects factor of matching and the between-subjects factor of Experiment (Experiment 2 vs. Experiment 3 of Vatakis & Spence, 2007) revealed a significant main effect of experiment ($F(1,20) = 7.30$, $P < .01$), with the sensitivity of participants' TOJs in Experiment 2 ($M = 72$ ms) with monkey calls being significantly worse than that obtained in Vatakis and Spence's (2007) previous study ($M = 58$ ms) with human speech. The analysis also revealed an interaction between matching and experiment ($F(1,20) = 4.33$, $p = .05$), with the sensitivity of participants' TOJs in the mismatching condition being significantly worse in Experiment 2 ($M = 72$ ms) than in the Experiment 3 of Vatakis and Spence ($M = 52$ ms). No main effect of matching was obtained ($F(1,20) = 3.29$, $p = .90$). This analysis therefore demonstrates that the JNDs were higher (i.e., performance was worse) for the case of monkey calls as compared to human speech stimuli, thus implying that participants exhibited greater sensitivity when discriminating the temporal alignment of the audiovisual signals for the speech stimuli than when discriminating the nonspeech stimuli. More importantly, the JNDs in the mismatched conditions were significantly higher (i.e., participants were less sensitive) in the case of monkey calls than when presented with mismatched human speech stimuli.

Taken together, the results of Experiments 1 and 2 suggest that the unity effect has no behavioral consequences for the human perception of monkey calls, suggesting its specificity to human vocalizations. Because both the human face and the human speech are thought to represent "special" signals for human perceivers (Diamond & Carey, 1986; Liberman & Mattingly, 1985), Experiment 3 tested whether the unity effect would influence the perception of human coos, monkey coos, and mismatched audiovisual combinations across the two species. Specifically, the stimuli consisted of (a) an adult male rhesus monkey and a human producing a "coo" and (b) the same video clips but with the auditory channels swapped over so that the sight of the rhesus monkey cooing was paired with the human's "coo" sound and vice versa. By using these stimuli, we were able to assess whether the sole presence of a human face was driving the unity effect that has been demonstrated for speech signals (Vatakis & Spence, 2007).

The sigmoid fits of participants' mean responses for each condition at the various SOAs presented in Experiment 3 are shown in Figure 2C. The analysis of the JND data revealed no significant difference in sensitivity to the temporal alignment of the auditory and visual signals in the case of matched (e.g., the visual- and auditory-signal of a human cooing; $M = 56$ ms) versus mismatched video clips (e.g., the visual-signal of a human cooing paired with the auditory-signal of a monkey cooing; $M = 55$ ms; see Figure 3) ($F(1,17) < 1$, ns). There was no significant main effect of Caller (human vs. monkey) ($F(1,17) < 1$, ns) nor any interaction between face–voice match and caller ($F(1,17) < 1$, ns).

Given that the null effect of matching that we obtained might have been driven by the participants simply being unaware of the audiovisual mismatch present in the video clips, we conducted a control study (similar to the control studies reported for Experiments 1 and 2). The 5 participants performed the discrimination task very accurately ($M = 95\%$ correct), thus showing that, as expected, the audiovisual discrepancy on trials where the auditory- or visual-signal belonged to the monkey or the human vocalization was easily detectable by participants.

The results of Experiment 3 are entirely consistent with those of Experiments 1 and 2 in showing no evidence for a unity effect for nonspeech sounds within and across humans and another primate species. The mere presence of a human face does not therefore in-and-of-itself necessarily promote the matching effect observed for audiovisual speech. This suggests that the unity effect may be driven solely by the presence of auditory-speech sounds or else that the integration of the highly correlated auditory-speech and visual-speech signals may have been necessary to promote the unity effect seen in our previous study (Vatakis & Spence, 2007).

The three independent null results reported in Experiments 1–3 represent what Frick (1995) would call "a good effort" to investigate whether the unity effect is present for monkey vocalizations, given the number of participants tested (32 participants in total), the number of experimental trials conducted per participant, the use of an experimental manipulation that has been shown to provide an effective and sensitive measure of performance regarding the existence of the unity effect (Vatakis & Spence, 2007, 2008), and the avoidance of both floor and ceiling effects (by varying the SOA, and thus the difficulty of the task; cf. Frick, 1995, on the interpretation of null effects).

According to Frick (1995), a null effect can be considered meaningful as long as a significant effect can be found in very similar circumstances. We therefore conducted a final experiment in order to replicate our previous speech findings (Vatakis & Spence, 2007). Specifically, we presented video clips of a British male and female uttering the speech sound /a/. The stimuli could either be matched (e.g., matched visual lip movements and auditory-speech of the female uttering /a/) or mismatched (e.g., the visual lip movements of the female together with the auditory-speech sound of the male uttering /a/). Based on the results of our 4 previous experiments (Vatakis & Spence, 2007), the expectation was that the sensitivity of participants' TOJs would be lower (i.e., JNDs would be significantly higher) for the matched than for the mismatched video clips due to the greater cross-modal integration occurring for the matched auditory and visual signals as compared to the mismatched signals.

The participants' mean responses for each condition at the various SOAs presented in Experiment 4 are depicted in the sigmoid fits of Figure 2D. Analysis of the JND data revealed a significant main effect of face–voice match ($F(1,14) = 7.50$, $P < .01$), with the sensitivity of participants' responses being significantly higher when the face and the voice were mismatched (i.e., when they came from different gender speakers; $M = 54$ ms) than when they matched (i.e., when they referred to the same underlying multisensory speech event; $M = 69$ ms; see Figure 3). There was no significant main effect of the gender of the speaker (male or female) ($F(1,14) < 1$, ns) nor any interaction between face–voice match and gender ($F(1,14) < 1$, ns).

In order to examine whether speech signals are more strongly integrated than nonspeech audiovisual signals, we compared the data obtained from Experiments 1 and 4 (where different individuals from the same species were used as stimuli; i.e., different monkeys for Experiment 1 and different humans in Experiment 4). A mixed ANOVA with the within-participants factor of matching and the between-participants factor of experiment (Experiment 1 vs. Experiment 4) revealed a significant main effect of experiment ($F(1,27) = 8.26$, $P < .001$), with the participants in Experiment 1 being significantly less sensitive to the temporal alignment of the stimuli ($M = 91$ ms) than those tested in Experiment 4 ($M = 62$ ms). There was also a significant interaction between matching and experiment ($F(1,27) = 4.98$, $P < .05$), with the sensitivity of participants' TOJs being significantly worse in Experiment 1 ($M = 93$ ms) than in Experiment 4 ($M = 55$ ms) but only for the case of mismatched stimuli. No main effect of matching was obtained ($F(1,27) = 1.19$, $p = .285$). Overall, participants were more sensitive (i.e., their JNDs were lower) to the asynchronously present in speech as compared to nonspeech stimuli. Additionally, participants were also more sensitive to the asynchrony present in the mismatched speech as compared to the mismatched animal call stimuli, thus implying a higher degree of integration for the mismatched speech stimuli than for the mismatched animal call stimuli.

General discussion

These results demonstrate that the unity effect modulates the multisensory integration of audiovisual speech and suggest that the effect is driven by either the acoustic speech signal or the integration of the visual- and auditory-speech signals (Experiments 3 and 4). In contrast, the unity effect does not seem to influence the multisensory integration of monkey vocalizations or human vocalizations when nonspeech sounds are uttered (i.e., when humans imitate monkey calls; see Experiments 1–3). More specifically, human participants' sensitivity to the temporal alignment of the auditory- and visual-speech stimuli was higher for the mismatched as compared to the matched conditions (Experiment 4). This was not the case for the monkey and the human-imitated monkey vocalization stimuli, where no difference in participants' sensitivity to audiovisual temporal alignment was observed between the matched and the mismatched video clips. As a whole, the results reported here demonstrate that while the unity effect can influence the multisensory integration of audiovisual speech stimuli, it does not appear to have any such effect on the integration of audiovisual nonspeech vocal stimuli.

To date, support for the unity effect has come primarily from studies of audiovisual speech perception (i.e., Easton & Basala, 1982; Vatakis & Spence, 2007, 2008; Walker et al., 1995; Warren, Welch, & McCarthy, 1981), but not from studies that have utilized nonspeech events (such as musical and object action events; see Radeau & Bertelson, 1977; Vatakis & Spence, 2008). The null results reported here extend these findings to the case of monkey vocalizations and human nonspeech vocalizations. Naturally, this raises the question of what drives the unity effect observed for speech but not for nonspeech stimuli? The answer to this question may relate to the putatively "special" nature of speech processing (e.g., Bernstein et al., 2004; Jones & Jarick, 2006; Liberman & Mattingly, 1985; Massaro, 2004; Munhall & Vatikiotis-Bateson, 2004; Tuomainen et al., 2005). In particular, the special nature of audiovisual speech may lie in the existence of a "specific mode of perception" that refers to the structural and functional processes related with the articulatory gestures of speech and/or to the perceptual processes associated with the phonetic cues that are present in speech signals (Tuomainen et al., 2005). The putatively special nature of speech processing is presumably driven by the fact that speech represents a very important stimulus for human interaction. Thus, according to the findings presented here, the mere presence of a human producing a nonsense sound (i.e., imitating a monkey vocalization; "coo") is, by itself, insufficient to elicit the unity effect. However, the presence of an acoustic speech signal and/or the presentation of a visual- and auditory-speech signal is sufficient to facilitate multisensory integration (probably giving rise to a temporal ventriloquism effect, whereby auditory stimuli affect the perceived time at which visual stimuli are perceived to occur; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Vatakis & Spence, 2007; Vroomen & Keetels, 2006).

One could argue that the difference between the results of Experiments 1–3 and those of Experiment 4 may be explained in terms of the high levels of people's exposure to, and production of, audiovisual speech. The validity of this argument could be supported by the fact that extensive exposure to speech may result in people being very accurate in recognizing articulatory gestures and facial movements as compared to other types of complex audiovisual events (such as music; see Summerfield, 1987). The notion of a heightened familiarity with speech could also explain why the participants in this study did not exhibit any facilitation when watching a human making cooing sounds. Additionally, for the case of mismatched speech, the participants may have been particularly sensitive to the incompatibility between the movements of the speakers' face and the speech sounds that they heard. Thus, participants may somehow have used this information as a cue to help them discriminate the nature of the audiovisual asynchrony that they were being exposed to.

One could possibly argue that the matching effect observed previously for human speech stimuli might also have been driven by the higher levels of sensorimotor experience that humans have with speech. However, developmental data on the perception of asynchrony argue against this account. In particular, Lewkowicz (1996, 2000) has shown that human infants are able to detect asynchrony in audiovisual speech signals from 6 to 8 months of age. Lewkowicz has also shown that infants are more sensitive to asynchrony for audiovisual nonspeech than for audiovisual speech stimuli, suggesting that they experience a greater unification of audiovisual speech events. Given that human infants do not fully develop speech until the second year of life, it seems probable that the matching effect obtained for human speech is not driven by sensorimotor experience per se, given that infants can detect asynchrony prior to having any meaningful sensorimotor experience of speech. The infant data are therefore inconsistent with the notion that the matching effect for audiovisual speech in our own research is attributable to human sensorimotor experience.

One must also note, however, that the matching effect reported previously (Vatakis & Spence, 2007) cannot be driven *solely* by people's high exposure/familiarity with speech. That is, while a lack of familiarity could be argued to account for the results obtained with the monkey vocalizations in this study, earlier studies of the unity effect for nonspeech failed to demonstrate any such effect across a variety of settings and stimulus-types (e.g., object action and musical stimuli) that were very familiar. A more plausible alternative account to the "speech is special" argument for the unity effect would therefore seem to be that people find audiovisual speech events to be somehow more "attention-grabbing" than non-speech events (e.g., see Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Theeuwes & Van der Stigchel, 2007).

Finally, one could argue that the observed unity effect for speech could have been driven by specific low-level stimulus features. There are many possible low-level acoustic differences between monkey calls and human speech; however, several studies have also shown that there is quite an overlap between the two kinds of stimuli. For example, formant frequencies that define vowel sounds are important cues for assessing the size and the gender of human speakers (Smith & Patterson, 2005), which are also present in rhesus monkey where some vocalizations have the same size-related formant pattern (Rendall, Owren, & Rodman, 1998) that can be used for size discrimination (Ghazanfar et al., 2007). This is not surprising since the general mechanism of vocal production is identical between humans and monkeys (Fitch & Hauser, 1995). In order to examine the possibility that low-level cues might have been driving the results presented here, we analyzed our stimuli based on three low-level cues: duration, fundamental frequency (that defines pitch), and overall bandwidth. All three measures, however, showed that there was considerable overlap in the low-level acoustics of the human and monkey vocalizations used in this study (see Table 1; It must be noted that the F0 was observed to be higher for the monkey as compared to humans). We also add here that only a single exemplar for each of the stimulus comparisons was used; thus, it could be argued that the JNDs reported here could have been influenced by idiosyncrasies in the particular stimulus set. That is, differences in the way an utterance is made by the same speaker at different instances could have affected the JNDs reported (we note here that for at least the human stimulus creation, multiple instances of the same utterance were recorded in order to control for such deviations).

Stimulus type	F0 (Hz)	Max frequency (Hz)	Duration (ms)
Monkey A "grunt"	676	13505	179
Monkey A "coo"	580	12746	737
Monkey B "threat"	907	10024	267
Monkey B "coo"	567	9414	551
Human "coo"	377	10914	710
Female /a/	283	8923	298
Male /a/	217	10020	321

Table 1. The low-level features (duration, fundamental frequency, and overall bandwidth) of the human speech and monkey vocalizations stimuli used in this study demonstrate the overlap between these stimuli.

Overall, the absence of any differences in participants' JNDs for matched and mismatched nonspeech stimuli demonstrates that the unity effect does not appear to influence the multisensory integration of non-speech vocal stimuli. In contrast, the unity effect facilitates the multisensory integration of audiovisual speech stimuli as a function of the acoustic speech signal and/or the integration of the auditory–visual speech stimulus. Resolving the relative contribution of auditory speech and/or the integration of auditory and visual speech to the unity effect represents an important issue for future research.

Acknowledgments

The rhesus monkeys used to create the video clips in this study were all socially housed at the Max Planck Institute for Biological Cybernetics, Tuebingen, Germany. We thank Chandramouli Chandrasekaran for his help with data fitting. A.V. was supported by a Newton Abraham Studentship from the Medical Sciences Division, University of Oxford. Correspondence regarding this article should be addressed to A.V. at the Department of Experimental Psychology, University of Oxford, 9 South Parks Road, Oxford, OX1 3UD.

Commercial relationships: none.

Corresponding author: Argiro Vatakis.
Email: argiro.vatakis@gmail.com.
Address: Department of Experimental Psychology, University of Oxford, 9 South Parks Road, Oxford, OX1 3UD, UK.

Footnote

[1] As mentioned in the main text, the PSS data for each of the experiments were not reported in detail since it has been demonstrated that the "unity effect" does not have any reliable effect on the PSS (cf. Vatakis & Spence, 2007, 2008). For the sake of completeness, we report here the statistical results obtained for the PSS for each of the four experiments (same factors as in JND analysis): Experiment 1: ($F(1,13) = 26.63$, $p < .01$); Experiment 2: ($F(1,9) < 1$, ns); Experiment 3: ($F(1,17) = 1.76$, $p = .20$); Experiment 4: ($F(1,14) = 9.45$, $p < .01$).

References

Bernstein, L. E., Auer, E. T., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association? In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 203–223). Cambridge, MA: MIT Press.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception in crossmodal space and crossmodal attention. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 141–177). Oxford: Oxford University Press.

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 6,* 1048–1053. [PubMed]

Coren, S., Ward, L. M., & Enns, J. T. (2004). *Sensation & perception* (6th ed.). Fort Worth: Harcourt Brace.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 107–117. [PubMed]

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32,* 562–570. [PubMed]

Finney, D. J. (1964). *Probit analysis: Statistical treatment of the sigmoid response curve*. London, UK: Cambridge University Press.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America, 102,* 1213–1222. [PubMed]

Fitch, W. T., & Hauser, M. D. (1995). Vocal production in nonhuman primates: Acoustics, physiology and functional constraints on 'honest' advertising. *American Journal of Primatology, 37,* 191–219.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138. [PubMed]

Ghazanfar, A. A., & Logothetis, N. K. (2003). Neuroperception: Facial expressions linked to monkey calls. *Nature, 423,* 937–938. [PubMed]

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25,* 5004–5012. [PubMed] [Article]

Ghazanfar, A. A., Turesson, H. K., Maier J. X., van Dinther, R., Patterson, R. D., & Logothetis, N. K. (2007). Vocal-tract resonances as indexical cues in rhesus monkeys. *Current Biology, 17,* 425–430. [PubMed] [Article]

Guski, R., & Troje, N. F. (2003). Audiovisual phenomenal causality. *Perception & Psychophysics, 65,* 789–800. [PubMed]

Hauser, M. D., Evans, C. S., & Marler, P. (1993). The role of articulation in the production of rhesus monkeys, *Macaca mulatta,* vocalizations. *Animal Behaviour, 45,* 423–433.

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology, 5,* 52–65.

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research, 174,* 588–594. [PubMed]

Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1094–1106. [PubMed]

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development, 71,* 1241–1257. [PubMed]

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 6771–6774. [PubMed] [Article]

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

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 153–176). Cambridge, MA: MIT Press.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17,* 154–163. [PubMed]

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 177–188). Cambridge, MA: MIT Press.

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

Radeau, M., & Bertelson, P. (1987). Auditory–visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research, 49,* 17–22. [PubMed]

Rendall, D., Owren, M. J., & Rodman, P. S. (1998). The role of vocal tract filtering in identity cueing in rhesus monkey (*Macaca mulatta*) vocalizations. *Journal of the Acoustical Society of America, 103,* 602–614. [PubMed]

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics, 54,* 406–416. [PubMed]

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age. *Journal of the Acoustical Society of America, 118,* 3177–3186. [PubMed] [Article]

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28,* 61–70.

Spence, C., Shore, D. I., & Klein, R. M. (2001). Multisensory prior entry. *Journal of Experimental Psychology: General, 130,* 799–832. [PubMed]

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). London: LEA.

Theeuwes, J., & Van der Stigchel, S. (2007). Faces capture attention: Evidence from inhibition-of-return. *Visual Cognition, 13,* 657–665.

Thomas, G. J. (1941). Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology, 28,* 163–177.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96,* B13–B22. [PubMed]

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756. [PubMed]

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127,* 12–23. [PubMed]

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 1063–1071. [PubMed]

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics, 57,* 1124–1133. [PubMed]

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory 'compellingness' in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics, 30,* 557–564. [PubMed]

Welch, R. B. (1999a). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371–387). Amsterdam: Elsevier Science, BV.

Welch, R. B. (1999b). The advantages and limitations of the psychophysical staircase procedure in the study of intersensory bias: Commentary on Bertelson in Cognitive contributions to the perception of spatial and temporal events. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 363–369). Amsterdam: Elsevier Science, BV.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667. [PubMed]

Witkin, H. A., Wapner, S., & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology, 43,* 58–67. [PubMed]