Perception & Psychophysics
2007, 69 (5), 744-756

Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli

Argiro Vatakis and Charles Spence
University of Oxford, Oxford, England

We investigated whether the "unity assumption," according to which an observer assumes that two different sensory signals refer to the same underlying multisensory event, influences the multisensory integration of audiovisual speech stimuli. Syllables (Experiments 1, 3, and 4) or words (Experiment 2) were presented to participants at a range of different stimulus onset asynchronies using the method of constant stimuli. Participants made unspeeded temporal order judgments regarding which stream (either auditory or visual) had been presented first. The auditory and visual speech stimuli in Experiments 1–3 were either gender matched (i.e., a female face presented together with a female voice) or else gender mismatched (i.e., a female face presented together with a male voice). In Experiment 4, different utterances from the same female speaker were used to generate the matched and mismatched speech video clips. Measuring in terms of the just noticeable difference the participants in all four experiments found it easier to judge which sensory modality had been presented first when evaluating mismatched stimuli than when evaluating the matched-speech stimuli. These results therefore provide the first empirical support for the "unity assumption" in the domain of the multisensory temporal integration of audiovisual speech stimuli.

When presented with two stimuli, one auditory and the other visual, an observer can perceive them either as referring to the same unitary audiovisual event or as referring to two separate unimodal events. The binding versus segregation of these unimodal stimuli—what Bedford (2001) calls the object identity decision; see also Radeau and Bertelson (1977)—depends on both low-level (i.e., stimulus-driven) factors, such as the spatial and temporal co-occurrence of the stimuli (Calvert, Spence, & Stein, 2004; Welch, 1999), as well as on higher level (i.e., cognitive) factors, such as whether or not the participant assumes that the stimuli should "go together." This is the so-called "unity assumption," the assumption that a perceiver makes about whether he or she is observing a single multisensory event rather than multiple separate unimodal events—a decision that is made, at least in part, on the basis of the consistency of the information available to each sensory modality (Welch & Warren, 1980, p. 663; see also Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004),[1] and on the basis of perceptual grouping (Lyons, Sanabria, Vatakis, & Spence, 2006; Radeau & Bertelson, 1987; Sanabria, Soto-Faraco, Chan, & Spence, 2004; Spence, Sanabria, & Soto-Faraco, 2007; Thomas, 1941) and phenomenal causality (Guski & Troje, 2003; Michotte, 1946). It has been argued that whenever two or more sensory inputs are perceived as being highly consistent (i.e., as being related in a way that they appear to "go together"),[2] observers will be more likely to treat them as referring to a single audiovisual event (Jackson, 1953; Welch & Warren, 1980). They will then be more likely to assume that such inputs have a common spatiotemporal origin, and will therefore be more likely to bind them into a single multisensory perceptual object or event (Bedford, 2001; cf. Corballis, 1994).

Research on the audiovisual binding of sensory information, specifically with regard to the question of *stimulus localization*, dates back more than 60 years (see Jackson, 1953; Thomas, 1941; Witkin, Wapner, & Leventhal, 1952). By now, many different studies have shown that auditory stimuli are typically mislocalized toward visual stimuli, if they are presented at approximately the same time (Hairston et al., 2003; Jack & Thurlow, 1973; Radeau & Bertelson, 1977; Slutsky & Recanzone, 2001). The magnitude of this spatial ventriloquism effect has been shown to decrease as the separation between the auditory and visual stimuli is increased (Jack & Thurlow, 1973; Jackson, 1953; see Bertelson & de Gelder, 2004, for a review). Researchers have recently suggested that such crossmodal integration may in fact result from the statistically optimal combination of the unimodal inputs (Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2004; Ernst & Banks, 2002; Heron, Whitaker, & McGraw, 2004; Roach, Heron, & McGraw, 2006).

In the last few years, a phenomenon analogous to spatial ventriloquism has also been demonstrated in the temporal domain whereby, within a certain temporal window, visual stimuli are "pulled" into approximate temporal alignment with the corresponding auditory stimuli (Fend-

A. Vatakis, argiro.vatakis@psy.ox.ac.uk

rich & Corballis, 2001; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Scheier, Nijhawan, & Shimojo, 1999; Vroomen & Keetels, 2006). Also, a combination of spatial and temporal factors has been shown to affect audiovisual judgments of phenomenal causality on the basis of the saliency of the constituent perceptual events (i.e., without strict audiovisual synchrony being required; Guski & Troje, 2003). Common motion and other Gestalt grouping principles have also been shown to facilitate crossmodal binding (Lyons et al., 2006; Sanabria et al., 2004; Soto-Faraco, Kingstone, & Spence, 2003; Spence et al., 2007; Vroomen & de Gelder, 2000).

Given that a great deal is now known about the low-level spatiotemporal constraints on the multisensory integration of simple arbitrary stimulus pairs (as indexed, for example, by research on the ventriloquism effect, both spatial and temporal, with auditory beeps and visual flashes), it is important that, by using more realistic, ecologically valid, and meaningful stimuli, researchers start to explore the multisensory integration of events that more closely resemble our everyday experience. The use of more naturalistic stimuli should allow researchers to determine whether or not enhanced crossmodal integration will occur when people are presented with more meaningful combinations of auditory and visual stimuli (an idea often presented under the rubric of the "unity assumption"; Bertelson & de Gelder, 2004; Jack & Thurlow, 1973; Jackson, 1953; Vroomen, 1999; Welch, 1999; Welch & Warren, 1980).

The literature on spatial ventriloquism suggests that a greater visual bias of perceived auditory location occurs for more "meaningful" (or "compelling"; see Warren, Welch, & McCarthy, 1981) combinations of audiovisual stimuli—such as, for example, the sight of a steaming kettle and a whistling sound (Jackson, 1953)—than for the kinds of arbitrary, or nonmeaningful, combinations of stimuli, such as flashing lights and brief tones, that typically have been presented in the majority of previous laboratory studies. However, whereas several empirical studies have been taken to provide support for the "unity assumption" in the domain of spatial ventriloquism (Jack & Thurlow, 1973; Jackson, 1953; Warren et al., 1981; but see also Radeau & Bertelson, 1977, for evidence that the "realism" of the stimuli used has no effect on adaptation effects), the possibility that these results reflect nothing more than a response bias has not, as yet, been ruled out (Bertelson & Aschersleben, 1998; Bertelson & de Gelder, 2004; Caclin, Soto-Faraco, Kingstone, & Spence, 2002; Welch, 1999). For example, it is possible that the participants in Jackson's early study might simply have been biased to assume that hearing a steam whistle and seeing a steaming kettle at the same time meant that these two events ought to go together. Consequently, participants may have been more likely simply to decide to point to the location where they saw the steaming kettle, despite having been verbally instructed to point to the source of the whistling sound. Note that the participants presumably did not have any such preconceptions of unity when confronted with an arbitrary pairing of auditory and visual stimuli—such as, for example, the sound of a bell and the onset of a light source, as used in Jackson's other spatial ventriloquism experiment.

It therefore remains uncertain, from the findings of these previous studies, whether the "unity assumption" actually influenced the *perceptual* experience of participants in a relatively automatic manner, or whether it simply biased their *decisional* strategies instead (note also the possibility that other forms of decisional bias, such as criterion shifts, may also have influenced participants' performance in these former studies; see Shaw, 1980). It is also possible, of course, that the effects reported in these previous studies of spatial ventriloquism relevant to the evaluation of the "unity assumption" may equally reflect some unknown combination of perceptual and decisional effects.

It should also be pointed out that the presentation of "informationally rich" stimuli such as the sight and sound of a kettle (i.e., events that have a greater internal temporal coherence and temporally varying structure) may promote more enhanced multisensory integration than stimuli of "low" informational content such as briefly presented lights and tones, where the only time-varying information consists of the onset and offset transitions (Jack & Thurlow, 1973; Jones & Jarick, 2006). This form of multisensory integration, driven by the coherence or correlation between two sensory signals, can be thought of as a bottom-up form of integration (Armel & Ramachandran, 2003; Bermant & Welch, 1976; Radeau & Bertelson, 1987; Welch, 1999). It is, however, actually quite difficult to distinguish between the bottom-up and top-down modulation of multisensory integration in many situations (see the General Discussion on this point).

The possibility that "informationally rich" stimuli may give rise to enhanced multisensory integration has been supported by research reported by Warren et al. (1981). In their spatial ventriloquism study, the dynamic face of a speaker or a static visual stimulus (consisting of a 1 × 2 cm piece of tape placed at the location on the screen where the speaker's mouth would have been) was presented, together with an auditory speech signal under various degrees of spatial discrepancy. Warren et al. reported that the visual bias of perceived auditory localization was significantly larger when the visual stimulus consisted of the dynamic face associated with the speaker's voice than when it consisted of the simple static visual "spot." The authors argued that the "informationally richer" auditory and visual signals—the speaker's face and the matching voice—were more compelling than the visual "spot" and speech signal which, when combined, had lower informational content. As a consequence, more audiovisual integration was observed in the former case than in the latter. It is possible, however, that Warren et al.'s findings reflect nothing more than a temporal coherence difference that was present in the conditions utilized—that is, with the visual "spot" plus speech signal, which obviously lacked any crossmodal temporal coherence (whereas there would have been a high degree of temporal coherence between the matching auditory and visual speech stimuli).

The last few years have seen researchers extending their ideas on sensory dominance from the domain of

spatial ventriloquism to that of temporal ventriloquism (Bertelson & Aschersleben, 2003). In particular, several researchers have argued that auditory cues may dominate visual cues in the temporal domain (Fendrich & Corballis, 2001; Morein-Zamir et al., 2003; Scheier et al., 1999; Vroomen & Keetels, 2006). For example, Morein-Zamir et al. reported a series of experiments in which their participants were presented with pairs of brief visual stimuli, one above and the other below a central fixation point. The stimulus onset asynchrony (SOA) between the onset of the two light emitting diodes (LEDs) was varied using the method of constant stimuli (Spence, Shore, & Klein, 2001). Brief uninformative tones were also presented from a loudspeaker cone situated directly behind the fixation light. One sound was presented before the first light and the other was presented after the second light. Even though the sounds provided no useful information about which light (the upper or lower one) had been presented first, the presence of the sounds significantly improved the sensitivity of participants' temporal order judgment (TOJ) performance over a condition in which the sounds were presented in synchrony with the lights, or a condition of no sounds at all. On the basis of this and other experimental results, Morein-Zamir et al. argued that the first sound temporally ventriloquized the perceived onset of the first light earlier, thereby increasing the apparent temporal separation between the first and second light and improving TOJ performance. Results such as these have therefore been taken to suggest that whereas visual stimuli typically lead to the spatial ventriloquism of auditory stimuli, auditory stimuli may ventriloquize the perceived time at which visual stimuli seem to occur.

As yet, however, no one has attempted to address the question whether or not the "unity assumption" also affects temporal perceptual phenomena such as temporal ventriloquism; therefore, in order to address the question of what effect, if any, the "unity assumption" has on temporal perception in humans, we presented pairs of auditory and visual speech stimuli that were either matched (referring to the same complex underlying perceptual event) or mismatched (by combining the auditory and visual streams from different perceptual events). We matched the informational content of the stimuli across conditions by using exactly the same stimuli in both the matched and mismatched speech conditions (cf. Warren et al., 1981). In our first experiment, we used audiovisual speech stimuli (syllables) uttered by a female and a male speaker. According to the "unity assumption," participants should find it harder to determine whether the visual lip movements or the auditory speech were presented first when the two stimuli refer to the same underlying perceptual event than when they do not. Such an outcome would provide the first empirical demonstration that the "unity assumption" can facilitate the crossmodal binding of audiovisual speech stimuli at a perceptual level, while ruling out the decisional level confound that has hampered the interpretation of all previous studies on this topic.

Although the idea of investigating the perceptual consequences of combining a man's voice with a woman's face (or vice versa) is by no means new, previous research in this area has primarily focused on addressing questions related to speech identification performance, with a particular focus on the McGurk effect (McGurk & MacDonald, 1976). Moreover, the results of these previous studies of audiovisual speech perception are inconclusive as to whether the magnitude of the McGurk effect (i.e., the visual influence on the perception of audiovisual speech) is affected by whether or not the speaker's face and voice are gender-matched (Easton & Basala, 1982; Green & Gerdeman, 1995; Green, Kuhl, Meltzoff, & Stevens, 1991; Walker, Bruce, & O'Malley, 1995).

In the present article, we used this gender match/mismatching design, together with a TOJ task, to address what impact the "unity assumption" has on temporal perception of audiovisual speech stimuli, while avoiding the potential response bias confound encountered in previous research. In our TOJ task, the participants had to decide on each trial whether the auditory or the visual speech stimulus had been presented first. The use of the TOJ task allowed us to obtain two performance measures: the just noticeable difference (JND) and the point of subjective simultaneity (PSS). The JND provides a standardized measure of the sensitivity with which participants could judge the temporal order of the auditory and visual stimuli. The PSS provides an estimate of how long the speech event in one sensory modality had to occur before the speech event in the other modality in order for synchrony to be perceived (or rather, for the "vision first" and "sound first" responses to be chosen equally often). Our use of a TOJ task allowed us to overcome the response bias account that has hampered the interpretation of all previous studies of the "unity assumption" in the spatial domain because the presentation of the matched-versus-mismatched video clips should not differentially affect the likelihood of participants making a "sound first" as opposed to a "vision first" response (i.e., judgments of temporal order were orthogonal to judgments of unity in our experimental design).

EXPERIMENT 1

Method

Participants. Nineteen participants (7 men and 12 women) 18–32 years of age (mean age, 23 years) were given a £5 Sterling gift voucher or course credit in return for taking part in the experiment. All of the participants were naive as to the purpose of the study, and all reported having normal hearing and normal or corrected-to-normal visual acuity. The experiment lasted for approximately 40 min.

Apparatus and Materials. The experiment was conducted in a completely dark sound-attenuated testing booth. The participants sat at a small table, facing straight ahead. The visual stimuli were presented on a 17-in. (43.18-cm) TFT color LCD monitor (SXGA 1,240 × 1,024 pixel resolution; 60-Hz refresh rate), placed at eye level approximately 68 cm from the participant. The auditory stimuli were presented by means of two Packard Bell Flat Panel 050 PC loudspeakers, each 25.4 cm on either side of the center of the monitor; in other words, the auditory and visual speech stimuli were presented from the *same* spatial location. Throughout the experiment, white noise was presented continuously at 55 dB (A; as measured from the participants' head positions) from a loudspeaker cone positioned directly above the monitor (22.3 cm from its center) in order to mask any sounds made by the participants. The audiovisual stimuli consisted of eight black-and-white video clips presented on a black background using the Presentation programming software

(Neurobehavioral Systems Inc., CA). The video clips (300 × 280-pixel, Cinepak Codec video compression, 16-bit audio sample size, 24-bit video sample size, 25 frames/sec) were processed using the Adobe Premiere 6.0 software package. The clips consisted of the following: (1) close-up views of the faces of a British woman and of a British man (see Figure 1A) uttering the syllables (phonemes) /biː/ and /o/; (2) the same lip movements, but with the auditory channels reversed, so that the female face was now paired with the male voice uttering the same syllable and the male face was paired with the female voice. Both clips had a duration of 385 msec; that is, the duration of the event was equal to that of the entire clip, with both the auditory and visual signal being dynamic over the same time period. All speech stimuli were recorded under the same conditions, with the mouth starting and ending in a closed position. The articulation of /biː/ and /o/ was salient enough without having the speaker make the stimuli unnatural (through exaggerated enunciation). In order to achieve accurate synchronization of the dubbed video clips, each original clip was reencoded using XviD codec (single pass, quality mode of 100%). Using the multitrack setting in Premiere 6.0, the visual and auditory components of the to-be-dubbed videos were aligned based on the peak auditory signals of the two videos. The matched audiovisual signals (both lip movements and auditory speech signals) were subsequently aligned with the mismatched auditory or visual signal. A final frame-by-frame inspection of the video clips was performed in order to ensure the correct alignment of the auditory and visual signals.

At the beginning and end of each video clip, a still image (extracted from the first and last 33.33 msec of each clip, respectively) and background acoustic noise were presented for a duration equivalent to the SOA (values reported below) in order to avoid cuing the participants as to the nature of the audiovisual delay with which they were being presented. In order to achieve a smooth transition

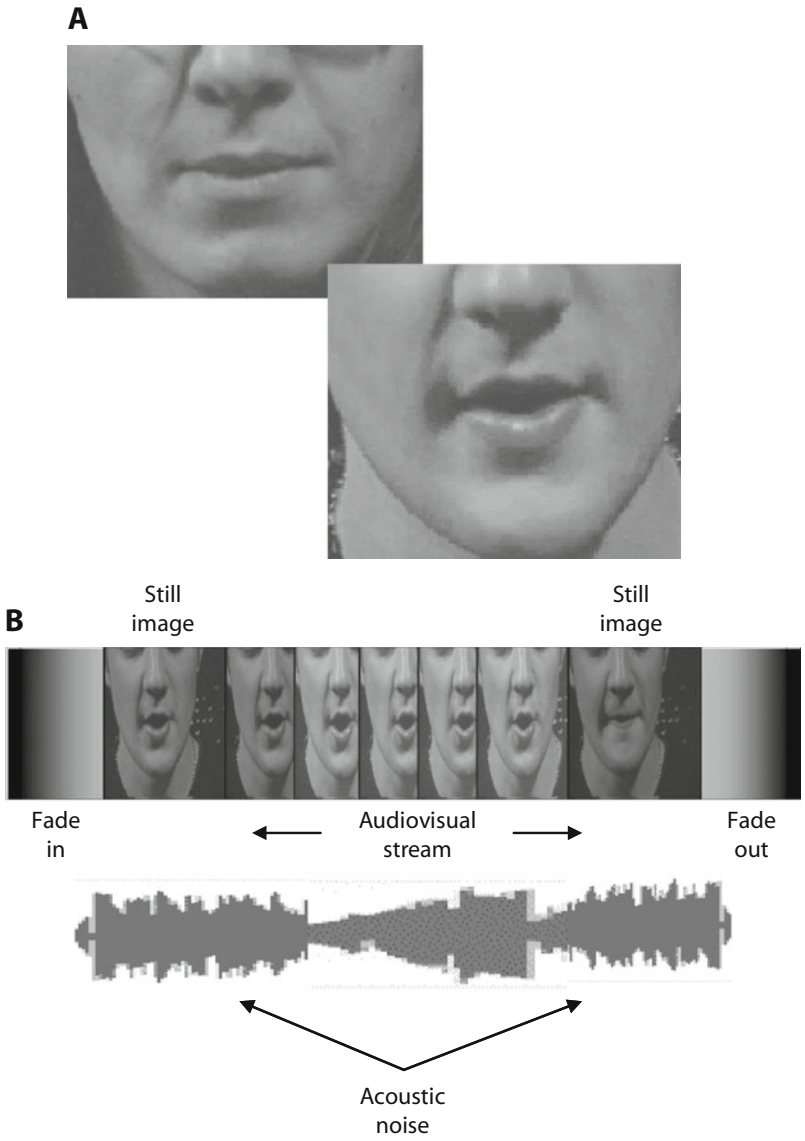

Figure 1. (A) Still images of the speech video clips used. (B) Schematic illustration of the sequence of auditory and visual events presented in each video clip; a still image, acoustic noise signal, and fade in/out were presented at the start and end of each video clip. The still image or the acoustic noise was presented for a duration equivalent to the SOA present in each video clip.

at the start and end of each video clip, a 33.33-msec cross-fade was added between the still image and the video clip. The cross-fading effect initially created a black screen that gradually faded to the still image and subsequently faded to the video clip (a similar fade was introduced at the end of each video clip), creating a smooth start and ending to each video (see Figure 1B). The participants responded by holding a standard computer mouse in both hands, using the right thumb for "vision first" responses and the left thumb for "audition first" responses (or vice versa; the response buttons were counterbalanced across participants).

Design. Nine possible SOAs between the auditory and visual stimuli were used: ±300, ±200, ±133, ±66, and 0 msec. Negative SOAs indicate that the auditory stream was presented first; positive values indicate that the visual stream was presented first. This particular range of SOAs was selected on the basis of previous research showing that people can typically discriminate the temporal order of briefly presented, matched auditory and visual speech stimuli at a 75% correct level (i.e., the conventionally defined JND; Spence et al., 2001) at SOAs of approximately 80 msec (McGrath & Summerfield, 1985; Munhall & Vatikiotis-Bateson, 2004; Vatakis & Spence, 2006). The participants completed one block of 8 practice trials before the main experimental session in order to familiarize themselves with the task and with the video clips. 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 intertrial interval was set at 1,000 msec, and the various SOAs were presented randomly within each block of trials using the method of constant stimuli (see Spence et al., 2001).

Procedure. The participants were informed that they would be presented with a series of matched and mismatched video clips of the faces and voices of a woman and a man. In order to familiarize the participants with the normal appearance of the speakers on the video clips, a clip showing the entire face and playing the original voice of each speaker was presented briefly at the start of the experiment. The participants were informed that on each trial they would have to decide whether or not the auditory or visual speech signal appeared to have been presented first, and that they would sometimes find this task difficult, in which case they should make an informed guess. The participants were also informed that the task was self-paced and that they should respond only when confident of the response. 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 would advance to the next trial.

Results and Discussion

The proportions of "vision first" responses were converted to their equivalent *z* scores under the assumption of a cumulative normal distribution (see Finney, 1964). The data from the 7 intermediate SOAs (±200, ±133, ±66, and 0 msec) were used to calculate best-fitting straight lines for each participant for each condition; in turn, these data were used to derive values for the slope and intercept (matched, /biː/, $r^2 = .91$, $p < .01$; /o/, $r^2 = .95$, $p < .01$; mismatched, /biː/, $r^2 = .95$, $p < .01$; /o/, $r^2 = .92$, $p < .01$); 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; see Figure 2A). The ±300-msec points were excluded from this computation; most participants performed nearly perfectly at this interval, and therefore these points did not provide significant information regarding our experimental manipulations (Spence et al., 2001, took a similar approach). The slope and intercept values were used to

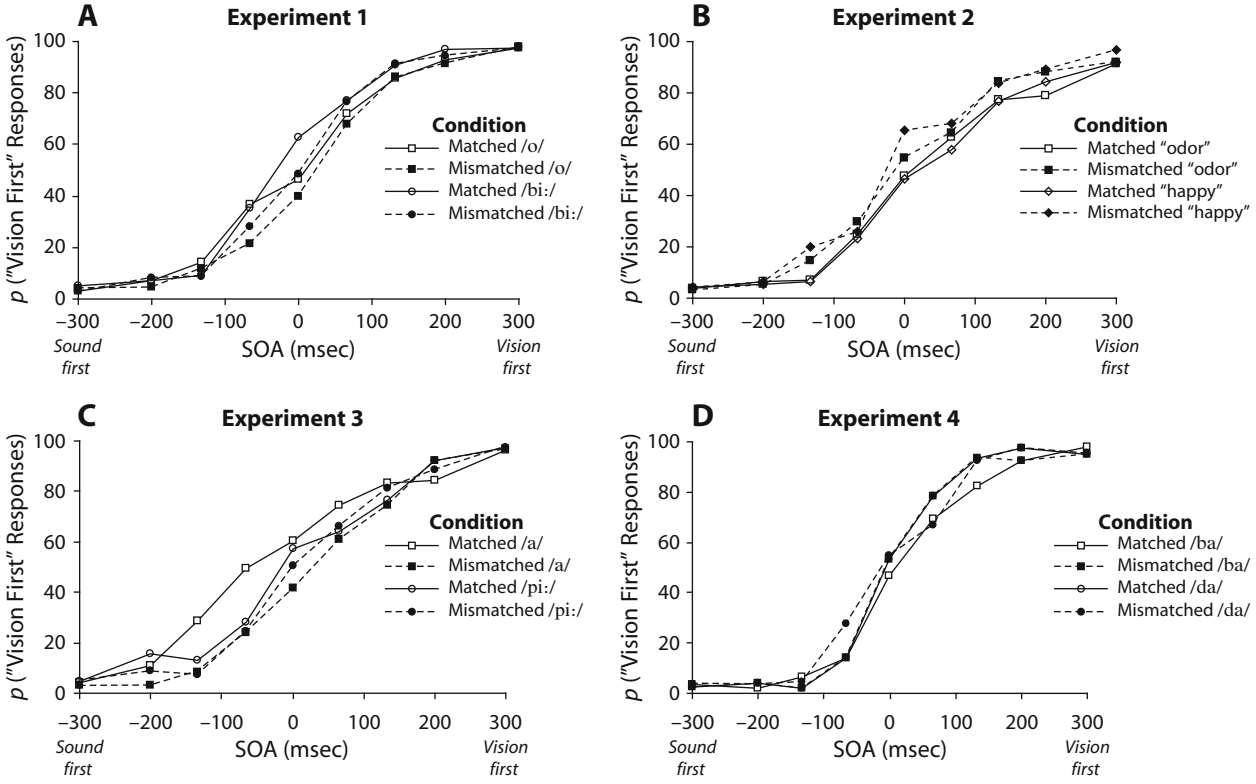

Figure 2. Mean percentage of "vision first" responses plotted as a function of stimulus onset asynchrony (SOA) for each of the phoneme and word conditions used in Experiments 1–4.

calculate the JND (JND = 0.675/slope; because ±0.675 represents the 75% and 25% points on the cumulative normal distribution) and PSS (PSS = −intercept/slope) values (see Coren, Ward, & Enns, 2004, for further details).

For all of the analyses reported here, Bonferroni-corrected *t* tests (where $p < .05$ prior to correction) were used for all post hoc comparisons. The JND and PSS data for each of the matched and mismatched speech stimuli were analyzed using a repeated-measures ANOVA with the factor of face–voice match (matched vs. mismatched). Preliminary analysis of the data with the factors of face–voice match and speech token revealed no main effect [$F(1,18) = 3.96$, $p = .17$] or interaction involving the speech token factor [$F(1,18) = 2.71$, $p = .20$]; and so we combined the data from the two different speech tokens used in our subsequent analysis.

Analysis of the JND data revealed a significant main effect of face–voice match [$F(1,18) = 6.92$, $p = .01$], with participants finding it significantly more difficult[3] to judge the temporal order of the auditory and visual speech stimuli correctly when face and voice matched—that is, when they referred to the same underlying multisensory speech event ($M = 70$ msec)—than when they were mismatched—that is, when they came from speakers of a different gender ($M = 59$ msec; see Figure 3A).[4] The analysis of the PSS data also revealed a significant main effect of face–voice match [$F(1,18) = 4.33$, $p < .05$] (see Figure 3B), with the matched audiovisual speech stimuli requiring an auditory lead for the PSS to be reached ($M = 7$ msec), whereas a visual lead was required for the mismatched speech ($M = 11$ msec). Similar variations in the PSS have been reported in other recent TOJ studies using speech stimuli, and can presumably be attributed to the different "weighting" of visual and auditory signals involved in the production of various different speech sounds (Abry, Cathiard, Robert-Ribes, & Schwartz,

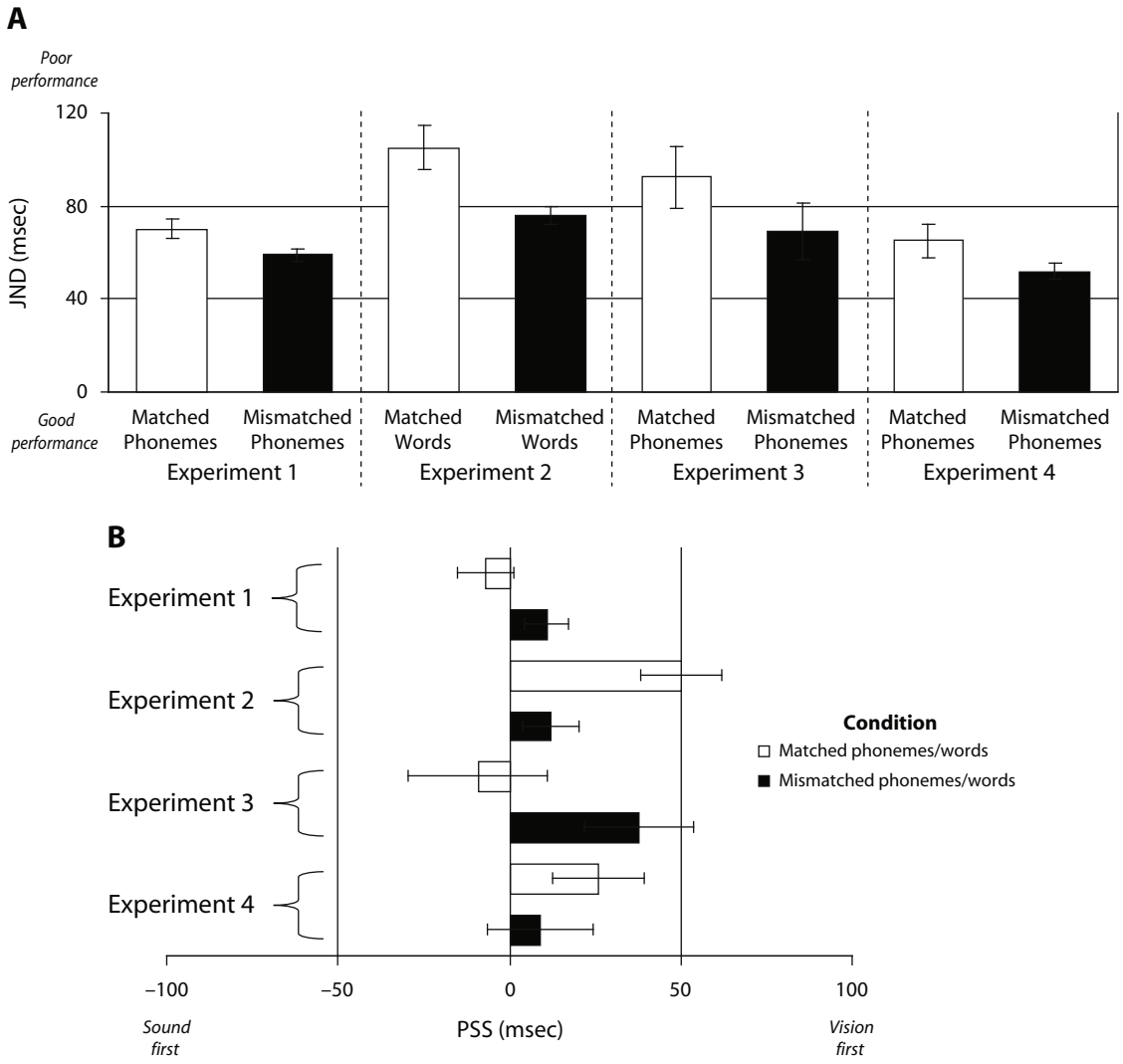

Figure 3. (A) Average JNDs for the matched and mismatched audiovisual speech stimuli (phonemes and words) presented in Experiments 1–4. (B) PSSs for the matched and mismatched phonemes and words. The error bars represent the standard errors of the mean.

1994; van Wassenhove, Grant, & Poeppel, 2005; Vatakis & Spence, 2006). Note that we had no specific predictions of what effect, if any, the presentation of matched-versus-mismatched video clips would have on the PSS values reported here, because researchers have not, as yet, specifically predicted the effect of the "unity assumption" on this particular performance measure. The PSS data are therefore reported in the present study primarily for the sake of completeness.

Overall, the results of Experiment 1 indicate that participants found it easier to discriminate the temporal order of the auditory and visual speech stimuli when these were mismatched than when they were matched. This outcome provides the first robust psychophysical evidence that the "unity assumption" can facilitate the crossmodal binding of multisensory information in speech stimuli at a perceptual level. This conclusion is warranted by the fact that any bias to assume that the matched auditory and visual stimuli should be bound together could not have biased participants toward either a "vision first" or "sound first" response as would have been the case had we used a simultaneity judgment task, in which any assumption of unity might also have led to a decisional bias toward participants making the simultaneous response.

We worried that if participants had somehow "seen through" the experiment, they might intentionally have performed more accurately on those TOJ trials in which the stimuli were mismatched; however, it is important to note that participants could not simply have performed more accurately, since it would have been unclear to them at many of the intermediate SOA values tested what the correct/appropriate response should have been (i.e., either audition or vision first). This is one of the principal advantages of the TOJ task over the simultaneity judgment task when attempting to test the "unity assumption." One might also wonder, however, whether participants might intentionally have "thrown" trials—that is, made incorrect responses on those trials in which they felt certain of the correct response, in order to please the experimenter. One argument against this possibility is the fact that performance in the matched and mismatched conditions did not differ significantly at the longest SOAs, at which participants were most confident of their responses, and presumably the point at which they would have been most likely to respond incorrectly. Note also that if participants had adopted a conscious strategy of deliberately responding incorrectly at these longer SOAs, a slowing of response latencies at the longer intervals, as compared with the shorter intervals, could have been expected (cf. Shore, Spence, & Klein, 2001). However, subsequent analysis of the response time (RT) data from Experiment 1 did not provide any evidence to support this account. In particular, there was no main effect of SOA in the RT data [$F(1,18) = 1.18, p = .29$].

In our Experiment 2, we attempted to replicate our findings on the effect of the "unity assumption" on the temporal perception of speech stimuli, while at the same time examining whether the effect would extend to the perception of spoken words. We used the words *happy* and *odor*; *happy* was chosen because it has an onset phoneme (/h/) that is "poor" in visual cues and information. (According to the 1993 International Phonetic Alphabet, /h/ has a glottal place of articulation, compared with the bilabial /biː/ used in Experiment 1.) We chose to test word stimuli in Experiment 2 because we first wanted to explore whether the unity effect would generalize to influence people's perception of other kinds of speech stimuli. One could argue that the use of only two phonemes (in Experiment 1) might not be sufficient to provide solid evidence for our findings, so we used words that provided another complex stimulus. Second, aware of the great variability in the PSS values sometimes obtained for syllables versus words (Vatakis & Spence, 2006), we wanted to test both kinds of stimuli, in case this PSS variability had any effect on the matching-versus-mismatching effect.

EXPERIMENT 2

Method

Eighteen new participants (5 men and 13 women), 18–42 years of age (mean age, 25 years), took part in this experiment. The apparatus, stimuli, design, and procedure were exactly the same as in Experiment 1, except that the audiovisual speech stimuli now consisted of (1) close-up views of the two faces from Experiment 1 uttering the words *happy* and *odor* (both clips of 800 msec in duration); and (2) movements the same but with the auditory channels reversed, so that the female face was paired with the male voice uttering the same word, and vice versa.

Results and Discussion

The goodness of the data fits was significant for all conditions (matched, *happy*, $r^2 = .93$, $p < .01$; *odor*, $r^2 = .95$, $p < .01$; mismatched, *happy*, $r^2 = .94$, $p < .01$; *odor*, $r^2 = .95$, $p < .01$; see Figure 2B). Analysis of the JND data once again revealed a significant main effect of face–voice match [$F(1,17) = 11.12$, $p < .01$], indicating that participants found it significantly harder to judge the temporal order of the auditory and visual speech stimuli when face and voice matched ($M = 105$ msec) than when they were mismatched ($M = 76$ msec; see Figure 3A), just as in Experiment 1. The analysis of the PSS data (see Figure 3B) also revealed a significant main effect of face–voice match [$F(1,17) = 17.93$, $p < .01$], with the matched audiovisual speech stimuli requiring a greater visual lead ($M = 50$ msec), than did the mismatched stimuli ($M = 12$ msec). Note here that the trend in the PSS data was in the reverse direction to that reported in Experiment 1 (where the matched stimuli required an auditory lead at the PSS).

The results of Experiment 2 demonstrate that participants found it significantly easier to discriminate the temporal order of the auditory and visual speech stimuli when the latter were mismatched than when they were matched. These results therefore provide further evidence that the "unity assumption" can facilitate the crossmodal perceptual binding of multisensory speech stimuli—in this case, audiovisually presented words—as indexed here by the increased JND observed for matched as opposed to mismatched speech stimuli reported in both Experiments 1 and 2.

In the two experiments reported so far, however, only the mismatched video clips were dubbed. The matched

video clips were not dubbed, because exact synchronization of the auditory and visual streams was assured by the original recordings themselves. We thought it possible that the dubbing of the mismatched video clips might inadvertently have created between the matched and mismatched video clips some kind of physical difference that participants noticed and used to improve their temporal discrimination performance. In Experiment 3, therefore, we attempted to replicate our findings (on the effect of the "unity assumption" on the temporal perception of speech stimuli) by using both dubbed matched and dubbed mismatched versions of syllables that belonged to the same phonetic categories as those used in Experiment 1. We used the syllables /pi:/ and /a/ (i.e., a bilabial and a back vowel, respectively, just as in Experiment 1). If the pattern of results reported in Experiment 3 was similar to that reported in the two previous experiments, we could rule out any account of the JNDs reported between the matched and mismatched speech stimuli in terms of artifacts introduced by the dubbing procedure.

EXPERIMENT 3

Method

Fourteen new participants (6 men and 8 women), 19–24 years of age (mean age, 21 years), took part in this experiment. The apparatus, stimuli, design, and procedure were exactly the same as for Experiment 1, with the sole exception that the audiovisual speech stimuli were dubbed for both the matched and mismatched conditions and consisted of (1) clips 385 msec long, with close-up views of the two faces from Experiment 1 which were now uttering the syllables /pi:/ and /a/; and (2) the same lip movements, but with the auditory channels reversed, so that the female face was paired with the male voice uttering the same word, and vice versa. In order to accurately synchronize the video clips, we used the multitrack setting in Premiere 6.0 and aligned the visual and auditory components of the to-be-dubbed videos on the basis of the peak signals of the two videos. A final frame-by-frame inspection of the video clips was performed in order to ensure the correct alignment of the auditory and visual signals.

Results and Discussion

The goodness of the data fits was significant for all conditions (matched, /pi:/, $r^2 = .95$, $p < .01$; /a/, $r^2 = .96$, $p < .01$; mismatched, /pi:/, $r^2 = .94$, $p < .01$; /a/, $r^2 = .95$, $p < .01$; see Figure 2C). Analysis of the JND data revealed a significant main effect of face–voice match [$F(1,13) = 11.42$, $p < .01$], with participants finding it significantly harder to judge the temporal order of the auditory and visual speech stimuli correctly when face and voice matched (i.e., when they referred to the same underlying multisensory speech event; $M = 92$ msec) than when they were mismatched (i.e., when they came from different gender speakers; $M = 69$ msec; see Figure 3A). Analysis of the PSS data also revealed a significant main effect of face–voice match [$F(1,13) = 17.87$, $p < .01$] (see Figure 3B), with the matched audiovisual speech stimuli requiring an auditory lead for the PSS to be achieved ($M = 9$ msec), just as in Experiment 1, whereas a visual lead was required for the mismatched speech ($M = 38$ msec).

Overall, the results of Experiment 3, like those of Experiments 1 and 2, indicate that participants found it easier to discriminate the temporal order of the auditory and visual speech stimuli when the latter were mismatched than when they were matched. Given these findings, and the fact that the same dubbing procedure was used to create all of the video clips (both matched and mismatched) in our third experiment, we can safely rule out an artifact of the particular dubbing procedure utilized in the creation of the mismatched video clips as being the principal cause of the difference in performance between the matched and mismatched audiovisual speech stimuli. Rather, these results provide additional support for our claim that the effect of the "unity assumption" seen in Experiments 1 and 2 represents a robust empirical phenomenon.

In the matched conditions of the three experiments reported so far, the auditory and visual components were derived from exactly the same utterance and speaker, whereas the mismatched conditions were not (see also Green et al., 1991, for a similar stimulus manipulation used in previous studies). One could therefore argue that the mismatched conditions, in comparison with the matched conditions, were gender-mismatched and were composed of auditory and visual components derived from different speech utterances. In order to determine whether the unity effect obtained in the three experiments reported so far relied on gender-based differences (i.e., involving amplitude and/or timbre differences between female and male voices) that may have been present in our mismatched speech stimuli, we conducted a fourth (and final) experiment. In Experiment 4, we attempted to replicate our findings with regard to the effect of the "unity assumption" on the temporal perception of speech stimuli but now using the matched versus mismatched presentations of the syllables /ba/ and /da/ derived from the same female speaker. If the pattern of results in our final experiment was found to be similar to that reported in our three previous experiments (where different speakers were utilized in the mismatched conditions), then this would show that the obvious physical gender mismatch present in the mismatched video clips was not required in order to obtain the unity effect reported in Experiments 1–3.

EXPERIMENT 4

Method

Twelve new participants (3 men and 9 women), 21–33 years of age (mean age, 25 years), took part in this experiment. The apparatus, stimuli, design, and procedure were exactly the same as for Experiment 1 with the following exceptions in the audiovisual speech stimuli: (1) close-up views of the face of just one woman (different from the one used in the previous experiments) uttering the syllables /ba/ and /da/ (both clips of 385 msec in duration); and (2) the same lip movements but with the auditory channels reversed so that the female face uttering /ba/ was now paired with the same female voice uttering /da/ and vice versa.

Results and Discussion

The goodness of the data fits was significant for all conditions (matched, /ba/, $r^2 = .96$, $p < .01$; /da/, $r^2 = .96$, $p < .01$; mismatched, /ba/, $r^2 = .95$, $p < .01$; /da/, $r^2 = .95$, $p < .01$; see Figure 2D). Analysis of the JND data revealed a significant main effect of face–voice

match [$F(1,11) = 5.76$, $p = .03$], with the participants finding it significantly harder to judge the temporal order of the auditory and visual speech stimuli correctly when the face and voice matched—that is, when they referred to the same underlying multisensory speech event ($M = 65$ msec)—than when they were mismatched—that is, when they came from different utterances ($M = 52$ msec; see Figure 3A). The analysis of the PSS data also revealed a significant main effect of face–voice match [$F(1,11) = 4.92$, $p < .05$] (see Figure 3B), with the matched audiovisual speech stimuli requiring a visual lead for the PSS to be achieved ($M = 26$ msec), whereas a smaller visual lead was required for the mismatched speech ($M = 9$ msec). Note here that whereas the trend in the PSS data toward a visual lead is similar to that reported in Experiment 2, it is in the reverse direction to that reported in Experiments 1 and 3; where the matched stimuli required an auditory lead at the PSS, thus suggesting that the "unity assumption" does not have a particularly reliable influence on the PSS. Instead, it seems more likely that the shifts reported in the PSS are driven by differences between the relative visual salience and onset time of the auditory components of different speech sounds (van Wassenhove et al., 2005).

Overall, the results of Experiment 4, in which the same speaker was utilized in both the matched and mismatched video clips, replicate the findings of all three previous experiments; that is, participants found it significantly easier to discriminate the temporal order of the auditory and visual speech stimuli in the mismatched than in the matched conditions, even when the mismatch occurred between different speech events uttered by the same speaker. Thus, our findings cannot be accounted for by any possible gender-based differences in the matched and mismatched conditions of the three previous experiments. The results of our final experiment therefore provide additional support for the view that the "unity assumption" can influence audiovisual speech perception.

GENERAL DISCUSSION

The most important finding to emerge from the four experiments reported in the present article was that participants found it significantly easier to discriminate the temporal order of the auditory and visual speech stimuli when they were mismatched (i.e., when the female voice was paired with a male face, or vice versa, in Experiments 1–3) than when they were matched (see Figure 3A). Taken together, these results therefore provide the first psychophysical evidence that the "unity assumption" can modulate the crossmodal binding of multisensory information at a perceptual level, at least in the case of the temporal perception of audiovisual speech stimuli as studied here. The enhanced multisensory binding taking place in the matched speech condition resulted in poorer temporal discrimination performance when the participants were presented with matched audiovisual speech events (i.e., originating from the same perceptual event) than that seen in the mismatched speech condition (i.e., when the stimuli came from different speech events). We were able to equate the informational content of the stimuli by using precisely the same stimuli to generate both the matched and mismatched video clips (cf. Warren et al., 1981). Moreover, the response bias account that has confounded the interpretation of all previous research in this area (see Bertelson & de Gelder, 2004, for a review) cannot account for our results, because any response bias—that is, any bias to assume that the vocal and facial signals either were or were not matched—should not have influenced participants preferentially toward making either an "auditory first" or a "visual first" response.

Unity and the McGurk Effect

The results of previous studies of audiovisual speech perception have provided somewhat inconsistent evidence concerning the question of whether or not the magnitude of the McGurk effect is affected by gender-based matching versus mismatching of speech stimuli. Our results, demonstrating the existence of significant differences between gender-matched and gender-mismatched stimuli, with audiovisual temporal discrimination performance being poorer in the former case, would appear to conflict with the results of a study by Green et al. (1991), which reported that the magnitude of the McGurk effect—the influence of visual speech on the perception of audiovisual speech—was unaffected by whether or not the lip movements and auditory speech signals were gender-matched. There are, however, two important points to note here. First, it has been argued by several researchers that the McGurk effect may dissociate from other audiovisual tasks, due to the possibly distinct perceptual processes underlying the former effect (Rosenblum & Saldaña, 1992; but see Bedford, 2001). Second, two other published studies (Easton & Basala, 1982; Walker et al., 1995) have in fact reported a significant effect of gender mismatching on audiovisual speech perception (i.e., providing results that also conflict with those of Green et al., 1991).

Easton and Basala (1982) reported two experiments in which they assessed the ability of participants to lip-read (monosyllabic and/or compound words) under conditions of unimodal visual presentation rather than under conditions of discrepant visual-auditory presentation (i.e., by varying the degree of discordance of the initial and final phonetic positions of words that were presented). Easton and Basala reported that their participants' ability to recognize (i.e., lip-read) visual speech was substantially impaired by the presence of discrepant auditory information. The magnitude of this crossmodal interference effect was related to the "compellingness," the degree to which the two events appeared to "go together," of the visual-auditory signals, in terms of the extent to which they were perceived as constituting a unified perceptual event (cf. Warren et al., 1981). Walker et al. (1995) also observed an interaction between gender discrepancy and familiarity in their study of the McGurk effect. In particular, a smaller McGurk effect was observed with gender-incongruent auditory and visual speech stimuli, but only when participants were familiar with either the speaker's face or voice. Thus, the results of these two studies would appear to suggest that the "unity assumption" *can* affect the audiovisual integration of speech stimuli, at least under cer-

tain conditions, a finding that would be in agreement with the findings concerning audiovisual temporal perception reported here.

Temporal Ventriloquism and the Unity Effect

The unification of events presented in different sensory modalities (an auditory and visual signal, in our case) should lead to the merging of the sensory signals into a single multisensory percept. This multisensory stimulus/event should presumably have a unique temporal onset. Consequently, any asynchrony that happened to be present between the auditory and visual signals prior to their integration would be expected to be reduced, if not eliminated completely, by the multisensory integration of the component unisensory events. Hence, participants should find it harder to judge whether the visual lip movements or the auditory speech came first, thus resulting in a higher JND. The specific mechanisms that govern this reduced temporal resolution are still unknown, but one possible explanation emerges from recent work on the phenomenon of temporal ventriloquism: Over the last few years, a number of studies have shown that auditory signals can ventriloquize the perceived time of occurrence of slightly asynchronous visual signals (Fendrich & Corballis, 2001; Morein-Zamir et al., 2003; Scheier et al., 1999; Vroomen & Keetels, 2006). Hence, in the present study, the greater degree of multisensory integration seen in the matched conditions (attributable to the "unity assumption") may have led to a reduction in the onset latency differences between the auditory and visual stimuli and therefore to worse temporal discrimination performance than in the mismatched conditions, where less integration (temporal ventriloquism) should presumably have taken place. Uncovering the specific mechanism underlying the unity effect will clearly represent an important issue for future research.

The analysis of the PSS data in the present article reveals that the matched phoneme video clips used in Experiments 1 and 3 required small auditory leads, whereas the mismatched clips required visual leads for the PSS to be reached. This was not true for the word video clips used in Experiment 2, or for the syllable clips used in Experiment 4, where visual leads were required instead for both the matched and mismatched video clips. One possible explanation for this difference might be that the phonetic and physical properties involved in the production of different speech sounds vary as a function of the particular speech sound being uttered (e.g., /biː/ in Experiment 1; /ba/ in Experiment 4). For example, for speech stimuli having a high visibility speech contrast (e.g., for bilabial stimuli such as /ba/), the human information processing system requires less of a visual lead for the PSS to be achieved than when in the presence of ambiguous (i.e., less informative) visual signals (e.g., Vatakis & Spence, 2006).

This view has received support from a recent neuroimaging study by van Wassenhove et al. (2005), in which it was shown that salient visual inputs, as in the phoneme /pa/, affected auditory speech processing at very early stages of processing (i.e., within 50–100 msec of stimulus onset) by enabling the perceptual system to make a prediction concerning the about-to-be-presented auditory input. The authors argued that this difference in the nature of the production of different speech sounds might explain the variations reported previously in the auditory or visual delays and/or leads required for the successful perception of synchrony for audiovisual speech (Abry et al., 1994; van Wassenhove et al., 2005; Vatakis & Spence, 2006). Given that auditory stimuli may dominate for certain audiovisual speech stimuli, whereas visual cues dominate for other speech stimuli, the crossing of the auditory and visual stimuli (of putatively differing "dominance" values) in the present study may help to explain why the PSS values in the mismatched videos were sometimes different from those reported for the matched conditions despite the fact that, overall, the same stimuli were presented in both cases. Resolving the factors contributing to the PSS shifts observed in the experiments reported in the present study represents a pertinent area for future research.

Top-Down and Bottom-Up Factors Contributing to Multisensory Integration

It is important to note that multisensory integration can be modulated by both top-down and bottom-up factors (see also Welch, 1999). Thus far, we have assumed that the effect of the "unity assumption" on performance reported in the present study was attributable to top-down factors (i.e., cognitive factors that affect the decision about whether or not two signals go together, or refer to the same event; Radeau & Bertelson, 1977). However, it is important to note that stimulus-driven (or structural) factors, such as any fine-timescale temporal correspondence or correlation between the stimuli occurring in the two streams (see Armel & Ramachandran, 2003; Bermant & Welch, 1976; Jones & Jarick, 2006; Radeau & Bertelson, 1977, 1987; Welch, 1999) can also facilitate multisensory integration in a purely bottom-up manner. We attempted to minimize any such bottom-up differences in the integration of the auditory and visual speech stimuli in the present study by carefully matching the timing of the visual and auditory events used to make the matched and mismatched videos. We cannot, however, unequivocally rule out the possibility that any subtle residual differences in the degree of matching may also have resulted in slightly different levels of bottom-up multisensory integration. Therefore, at the present time, it would seem most prudent to assume that our findings on the influence of the "unity assumption" on multisensory temporal perception reflect some unknown combination of both top-down and bottom-up factors influencing multisensory integration. It is worth noting, however, that both top-down and bottom-up factors presumably operate at the same time to facilitate the appropriate multisensory integration of environmental events under the majority of naturalistic conditions (see Radeau & Bertelson, 1977; Welch, 1999).

It would appear likely that the processes underlying the unification of sensory signals are more attentionally demanding when driven in a top-down manner than when driven in a bottom-up way; it would, therefore, be interesting for future research to investigate any such attentional differences that might help to discriminate between

top-down and bottom-up driven integration. One possibility here involves manipulating the perceptual load of a concurrently presented secondary task (Lavie, 2005). Loading a participant's attention might be expected to impair any top-down (i.e., attentionally demanding) process of unification, while leaving any bottom-up integration relatively intact (though see Alsius, Navarra, Campbell, & Soto-Faraco, 2005).

Suggestive evidence concerning the interplay between top-down and bottom-up effects on the unification of auditory and visual signals has been reported in a recent audiovisual spatial ventriloquism study by Hairston et al. (2003). Specifically, these authors demonstrated that the accuracy of their participants' auditory localization performance was dependent on both the low-level characteristics of the sensory environment (i.e., structural factors) and on higher order (i.e., cognitive) factors (Welch, 1999). The participants in Hairston et al.'s study not only had to judge the perceived location of the auditory stimuli (as in traditional spatial ventriloquism studies), but also had to report whether or not the auditory and visual stimuli appeared to have been spatially "unified" on each trial (i.e., originating from the same location). The crossmodal visual bias of auditory localization was found to be more pronounced on those trials where participants reported the auditory and visual stimuli to be unified than when they did not.

It is, however, important to note here that Hairston et al.'s (2003) use of a pointing response to localize the auditory targets means that the crossmodal bias reported in their study might unfortunately reflect nothing more than a response bias, as discussed earlier (see Bertelson & Aschersleben, 1998; Bertelson & de Gelder, 2004; Caclin et al., 2002; Welch, 1999). Moreover, it must at present remain unclear whether the top-down unification of the sensory signals by participants resulted in more multisensory integration (as indexed by a larger ventriloquism effect), or whether instead any trial-by-trial differences in the strength of the ventriloquism effect resulted in any variation in the degree of unification of the stimuli that the participants subsequently experienced; perhaps the unification judgments were driven solely by how close in space the auditory and visual stimuli appeared to have been presented in any given trial. Furthermore, the failure to monitor the participants' adherence to the central fixation instructions also means that an overt orienting (eye-movement) account of Hairston et al.'s findings cannot be ruled out either (Thurlow & Jack, 1973; Weerts & Thurlow, 1971).

It is perhaps worth pointing out that recent research has also suggested a potential role for purely top-down factors in determining whether distance constancy is observed for the case of audiovisual temporal judgments (see Arnold, Johnston, & Nishida, 2005, pp. 1281–1283; Sugita & Suzuki, 2003). In particular, Arnold et al. conducted an experiment in which they compared the psychophysical performance of their participants on an audiovisual TOJ task as a function of whether or not they were told to imagine the sound and visual stimuli as having originated from a common environmental location. Certain of Arnold et al.'s results suggest that this purely cognitive (top-down) manipulation concerning the origin (same vs. different) of the auditory and visual stimuli did indeed influence the pattern of results obtained.

There appear to be specific mechanisms in the human perceptual system involved in the binding of spatially and temporally aligned sensory stimuli. At the same time, the perceptual system also appears to exhibit a high degree of selectivity in terms of its ability to separate highly concordant events from events that meet the spatial and temporal coincidence criteria, but which do not necessarily "belong together." This has been demonstrated recently by Laurienti et al. (2004), who presented semantically congruent and semantically incongruent audiovisual stimuli to participants in a feature discrimination task. The participants in their study had to make speeded discrimination responses regarding whether a "red" or "blue" target had been presented on each trial. The target stimuli consisted of either a red or blue circle and/or the word *red* or *blue* spoken over headphones, respectively. On trials where a bimodal target was presented, it was always congruent (i.e., a red circle together with the word *red*). Distractor stimuli, which participants had to try to ignore, consisting of a green circle and/or the word *green*, were also presented on many of the trials. The participants responded more rapidly and accurately on the congruent bimodal target trials than on unimodal target trials, with the worst performance being reported on those trials where incongruent distractor stimuli were presented at the same time as the target. These results therefore show that the semantic content of multisensory stimuli, together with their temporal and spatial attributes, can also play an important role in multisensory perceptual processing.

When taken together with the results of the present study, Laurienti et al.'s (2004) findings add to a growing body of evidence suggesting a significant role for more cognitive factors in modulating multisensory integration (see also Arnold et al., 2005; Miller, 1972; Welch, 1999). Finally, it should be noted that the existence of such top-down influences on multisensory integration is inconsistent with previous claims (e.g., Radeau, 1994) that multisensory integration might reflect an encapsulated (i.e., cognitively impenetrable) process.

AUTHOR NOTE

A.V. was supported by a Newton Abraham Studentship from the Medical Sciences Division, University of Oxford. We thank the three anonymous reviewers for their very detailed and helpful comments on the manuscript. Correspondence regarding this article should be addressed to A.Vatakis at the Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England (e-mail: argiro.vatakis@gmail.com).

REFERENCES

ABRY, C., CATHIARD, M. A., ROBERT-RIBES, J., & SCHWARTZ, J. L. (1994). The coherence of speech in audio-visual integration. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **13**, 52-59.

ALAIS, D., & BURR, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, **14**, 257-262.

ALSIUS, A., NAVARRA, J., CAMPBELL, R., & SOTO-FARACO, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, **15**, 1-5.

ARMEL, K. C., & RAMACHANDRAN, V. S. (2003). Projecting sensations to

external objects: Evidence from skin conductance response. *Proceedings of the Royal Society of London: Series B*, **270**, 1499-1506.

ARNOLD, D. H., JOHNSTON, A., & NISHIDA, S. (2005). Timing sight and sound. *Vision Research*, **45**, 1275-1284.

BATTAGLIA, P. W., JACOBS, R. A., & ASLIN, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, **20**, 1391-1397.

BEDFORD, F. L. (1994). A pair of paradoxes and the perceptual pairing processes. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **13**, 60-68.

BEDFORD, F. L. (2001). Towards a general law of numerical/object identity. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **20**, 113-175.

BERMANT, R. I., & WELCH, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual & Motor Skills*, **42**, 487-493.

BERTELSON, P., & ASCHERSLEBEN, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, **5**, 482-489.

BERTELSON, P., & ASCHERSLEBEN, G. (2003). Temporal ventriloquism: Crossmodal interaction on the time dimension. 1. Evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology*, **50**, 147-155.

BERTELSON, P., & DE GELDER, B. (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 141-177). Oxford: Oxford University Press.

CACLIN, A., SOTO-FARACO, S., KINGSTONE, A., & SPENCE, C. (2002). Tactile "capture" of audition. *Perception & Psychophysics*, **64**, 616-630.

CALVERT, G. A., SPENCE, C., & STEIN, B. E. (EDS.). (2004). *The handbook of multisensory processing*. Cambridge, MA: MIT Press.

CORBALLIS, M. C. (1994). Do you need module any more? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **13**, 81-83.

COREN, S., WARD, L. M., & ENNS, J. T. (2004). *Sensation and perception* (6th ed.). Fort Worth, TX: Harcourt Brace.

EASTON, R. D., & BASALA, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics*, **32**, 562-570.

EPSTEIN, W. (1975). Recalibration by pairing: A process of perceptual learning. *Perception*, **4**, 59-72.

ERNST, M. O., & BANKS, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429-433.

FENDRICH, R., & CORBALLIS, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics*, **63**, 719-725.

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

FISHER, B. D., & PYLYSHYN, Z. W. (1994). The cognitive architecture of bimodal event perception: A commentary and addendum to Radeau (1994). *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **13**, 92-96.

GREEN, K. P., & GERDEMAN, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 1409-1426.

GREEN, K. P., KUHL, P. K., MELTZOFF, A. N., & STEVENS, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.

GUSKI, R., & TROJE, N. F. (2003). Audiovisual phenomenal causality. *Perception & Psychophysics*, **65**, 789-800.

HAIRSTON, W. D., WALLACE, M. T., VAUGHAN, J. W., STEIN, B. E., NORRIS, J. L., & SCHIRILLO, J. A. (2003). Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience*, **15**, 20-29.

HERON, J., WHITAKER, D., & MCGRAW, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research*, **44**, 2875-2884.

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

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.

LAURIENTI, P. J., KRAFT, R. A., MALDJIAN, J. A., BURDETTE, J. H., & WALLACE, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, **158**, 405-414.

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

LYONS, G., SANABRIA, D., VATAKIS, A., & SPENCE, C. (2006). The modulation of crossmodal integration by unimodal perceptual grouping: A visuo-tactile apparent motion study. *Experimental Brain Research*, **174**, 510-516.

MACDONALD, J., & MCGURK, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253-257.

MCGRATH, M., & SUMMERFIELD, Q. (1985). Intermodal timing relations and audiovisual speech recognition by normal hearing adults. *Journal of the Acoustical Society of America*, **77**, 678-685.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MICHOTTE, A. (1946). *La Perception de la Causalité* [The perception of causality]. Louvain: Institut Supérieur de Philosophie.

MILLER, E. A. (1972). Interactions of vision and touch in conflict and nonconflict form perception tasks. *Journal of Experimental Psychology*, **96**, 114-123.

MOREIN-ZAMIR, S., SOTO-FARACO, S., & KINGSTONE, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research*, **17**, 154-163.

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. (1994). Auditory-visual spatial interaction and modularity. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **13**, 3-51.

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.

ROACH, N. W., HERON, J., & MCGRAW, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio–visual integration. *Proceedings of the Royal Society of London: Series B*, **273**, 2159-2168.

ROSENBLUM, L. D., & SALDAÑA, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, **52**, 461-473.

SANABRIA, D., SOTO-FARACO, S., CHAN, J. S., & SPENCE, C. (2004). When does visual perceptual grouping affect multisensory integration? *Cognitive, Affective, & Behavioral Neuroscience*, **4**, 218-229.

SCHEIER, C. R., NIJHAWAN, R., & SHIMOJO, S. (1999). Sound alters visual temporal resolution. *Investigative Ophthalmology & Visual Science*, **40**, S792.

SHAW, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 277-296). Hillsdale, NJ: Erlbaum.

SHORE, D. I., SPENCE, C., & KLEIN, R. M. (2001). Visual prior entry. *Psychological Science*, **12**, 205-212.

SLUTSKY, D. A., & RECANZONE, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport*, **12**, 7-10.

SOTO-FARACO, S., KINGSTONE, A., & SPENCE, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia*, **41**, 1847-1862.

SPENCE, C., SANABRIA, D., & SOTO-FARACO, S. (2007). Intersensory Gestalten: Assessing the influence of intramodal perceptual grouping on crossmodal interactions. In K. Noguchi (Ed.), *The psychology of beauty and Kansei: New horizons of Gestalt perception* (pp. 519-579). Tokyo: Fuzanbo International.

SPENCE, C., SHORE, D. I., & KLEIN, R. M. (2001). Multisensory prior entry. *Journal of Experimental Psychology: General*, **130**, 799-832.

STEIN, B. E., & MEREDITH, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

STONE, J. V., HUNKIN, N. M., PORRILL, J., WOOD, R., KEELER, V., BEANLAND, M., ET AL. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society of London: Series B*, **268**, 31-38.

SUGITA, Y., & SUZUKI, Y. (2003). Implicit estimation of sound-arrival time. *Nature*, **421**, 911.

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

THURLOW, W. R., & JACK, C. E. (1973). Certain determinants of the "ventriloquism effect." *Perceptual & Motor Skills*, **36**, 1171-1184.

VAN WASSENHOVE, V., GRANT, K. W., & POEPPEL, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, **102**, 1181-1186.

VATAKIS, A., & SPENCE, C. (2006). Audiovisual synchrony perception for speech and music using a temporal order judgment task. *Neuroscience Letters*, **393**, 40-44.

VROOMEN, J. (1999). Ventriloquism and the nature of the unity assumption. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 389-393). Amsterdam: Elsevier.

VROOMEN, J., & DE GELDER, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 1583-1590.

VROOMEN, J., & KEETELS, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception & Performance*, **32**, 1063-1071.

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.

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.

WEERTS, T. C., & THURLOW, W. R. (1971). The effects of eye position and expectation on sound localization. *Perception & Psychophysics*, **9**, 35-39.

WELCH, R. B. (1972). The effect of experienced limb identity upon adaptation to simulated displacement of the visual field. *Perception & Psychophysics*, **12**, 453-456.

WELCH, R. B. (1999). 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.

WELCH, R. B., & WARREN, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, **88**, 638-667.

WITKIN, H. A., WAPNER, S., & LEVENTHAL, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology*, **43**, 58-67.

NOTES

1. Note that whereas the term "unity assumption" has appeared frequently in the literature on the ventriloquism effect over the years, there appears to be some disagreement whether it refers to a top-down (i.e., cognitive) or a bottom-up (i.e., stimulus-driven) process. Part of the problem is that many of the factors (such as spatial and temporal coincidence) that promote a top-down assumption of unity are also likely to lead to enhanced bottom-up multisensory integration (Stein & Meredith, 1993; Welch, 1999). Several researchers (Bedford, 1994; Epstein, 1975; Fisher & Pylyshyn, 1994; Radeau & Bertelson, 1977; Welch & Warren, 1980) have used the term *pairing* instead to designate essentially the same notion, albeit with fewer cognitive overtones (see Bertelson & Aschersleben, 2003; Welch, 1999).

2. For the purposes of the present article, we define *consistency* operationally, in terms of two or more sensory inputs being highly consistent if they originate in the same environmental event (and hence inconsistent whenever they have different environmental causes). Other researchers have used the term *compellingness* to designate essentially the same notion (Easton & Basala, 1982; Warren, Welch, & McCarthy, 1981). It should, however, be noted that under certain contrived laboratory conditions, stimuli from different sensory modalities can sometimes appear to be highly consistent even when they originate from different environmental events (see, e.g., Welch, 1972, for one such example).

3. It is important to note that lower JND values (estimated from a given psychometric function) do not necessarily equate to a higher percentage of correct responses, especially if they are accompanied by a large vision-first or sound-first bias (i.e., a nonzero PSS, as reported in the experiments documented in the present study). Therefore, we also calculated the accuracy of participants' TOJ responses in all four of our experiments. The results showed that participants responded more accurately to the mismatched stimuli than to matched stimuli in all four experiments. Main effects of matching, SOA, and matching × SOA interaction: for Experiment 1, $F(1,18) = 12.34$, $p < .01$; $F(5,90) = 21.63$, $p < .01$; $F(5,90) = 6.33$, $p < .01$; for Experiment 2, $F(1,17) = 4.50$, $p < .05$; $F(5,85) = 10.20$, $p < .01$; $F(5,85) = 10.63$, $p < .01$; for Experiment 3, $F(1,13) = 17.59$, $p < .01$; $F(5,65) = 13.24$, $p < .01$; $F(5,65) = 8.22$, $p < .01$; for Experiment 4, $F(1,11) = 5.63$, $p < .05$; $F(5,55) = 7.00$, $p < .01$; $F(5,55) = 3.77$, $p < .05$.

4. Because the slope differences were not particularly visible in our psychometric functions, one might wonder whether the JND differences obtained across the various conditions might have reflected an artifact of the particular analysis performed. Specifically, we thought it possible that the shift in the psychometric functions observed between the matched and mismatched conditions could have been driven by changes in performance as a function of the particular range of SOAs used to fit our psychometric functions. In order to address this concern, we repeated the analysis of the data from all four experiments but now utilized the data from only the middle five SOAs (−133 msec to +133 msec). However, this analysis revealed exactly the same pattern of results as those obtained utilizing the 7 SOAs reported in the main analyses in the text (in particular, for the main effect of matching: Experiment 1 [$F(1,18) = 4.93$, $p = .03$], Experiment 2 [$F(1,17) = 5.95$, $p = .02$], Experiment 3 [$F(1,13) = 7.72$, $p = .01$], and Experiment 4 [$F(1,11) = 3.10$, $p = .04$]. The results of these additional analyses therefore show that the main effect of matching reported in all four of our experiments does not simply reflect an artifact of the range of SOAs used to calculate the psychophysical estimates of performance.

(Manuscript received December 6, 2005;
revision accepted for publication November 27, 2006.)