Atten Percept Psychophys (2011) 73:209–218
DOI 10.3758/s13414-010-0019-0

No effect of synesthetic congruency on temporal ventriloquism

Mirjam Keetels · Jean Vroomen

Published online: 3 December 2010

Abstract A sound presented in temporal proximity to a light can alter the perceived temporal occurrence of that light (temporal ventriloquism). Recent studies have suggested that pitch–size *synesthetic congruency* (i.e., a natural association between the relative pitch of a sound and the relative size of a visual stimulus) might affect this phenomenon. To reexamine this, participants made temporal order judgements about small- and large-sized visual stimuli while high- or low-pitched tones were presented before the first and after the second light. We replicated a previous study showing that, at large sound–light intervals, sensitivity for visual temporal order was better for synesthetically congruent than for incongruent pairs. However, this congruency effect could not be attributed to temporal ventriloquism, since it disappeared at short sound–light intervals, if compared with a synchronous audiovisual baseline condition that excluded response biases. In addition, synesthetic congruency did not affect temporal ventriloquism even if participants were made explicitly aware of congruency before testing. Our results thus challenge the view that synesthetic congruency affects temporal ventriloquism.

Keywords Temporal ventriloquism · Synesthetic congruency · Intersensory perception · Visual TOJ

The question of how sensory modalities cooperate to form a coherent representation of the environment has been the focus of much behavioral and neuroscientific research (Calvert, Spence, & Stein, 2004). In the literature on intersensory perception, the most commonly held view on this topic has been the *assumption of unity*, which states that the more events from different modalities share (amodal) properties, the more likely it is that they originate from a common object or source (e.g., Bedford, 1989; Bertelson, 1999; Radeau, 1994; Stein & Meredith, 1993; Welch, 1999; Welch & Warren, 1980). It is widely accepted that commonality in space and time are of special importance for intersensory binding (Radeau, 1994; Stein & Meredith, 1993). Much less is known, though, about whether associations between *qualitative aspects* of sensory modalities can serve as a potential "binding" factor between the senses.

Of particular interest is whether *semantic congruency* between modalities affects intersensory binding. Although the term *semantic* is rather loosely defined in the literature, semantic stimuli that are used in intersensory studies typically concern an ecologically meaningful visual stimulus that is paired with a matching auditory counterpart, such as, a facial expression combined with a vocal expression (de Gelder, 2000; de Gelder & Vroomen, 2000; Dolan, Morris, & de Gelder, 2001). Other examples of semantic congruency are those between letters and speech sounds (van Atteveldt, Formisano, Goebel, & Blomert, 2004), images and sounds of common objects (Chen & Spence, 2010; Noppeney, Josephs, Hocking, Price, & Friston, 2008), and speaker identities (Noppeney et al., 2008). A task that has been particularly useful for examining congruency effects is the crossmodal temporal order judgement (TOJ) task. In this task, participants judge the relative temporal order of two information streams (e.g., auditory and visual) that are presented at various stimulus onset asynchronies (SOAs). The critical performance measure is the just-noticeable difference (JND), which represents the smallest interval at which the temporal order of the two information streams can still reliably be

M. Keetels · J. Vroomen (✉)
Tilburg University,
Tilburg, Netherlands
e-mail: J.Vroomen@uvt.nl

perceived. For congruent stimulus pairs, the idea is that observers should find it harder to judge temporal order (a large JND) because the information streams of congruent events are strongly bound. The streams will then be perceived as synchronous, so temporal order is lost.

Evidence in support of a role for semantic congruency in temporal perception has come mainly from a study by Vatakis and Spence (2007). They made participants judge the temporal order of audiovisual (AV) speech (an AV TOJ task) that was either matched or mismatched in gender (a female face articulating /pi/ with a sound of either a female or a male /pi/) or phonemic content (a face saying /ba/ with a voice saying /ba/ or /da/). As predicted by the unity assumption, sensitivity for temporal order was worse if the auditory and visual streams were congruent than if they were incongruent in either the gender or the phonemic content. More recently, though, Vatakis, Ghazanfar, and Spence (2008) qualified these findings and reported that the effect may be specific for human speech. In this more recent study, the effect of congruency was examined by comparing speech stimuli with matching or mismatching call types of monkeys (cooing vs. grunts or threat calls). For AV speech, sensitivity of temporal order was again worse for congruent than for incongruent trials, but there was no congruency effect for the monkey calls. In yet another study, Vatakis and Spence (2008) also found no congruency effect for audiovisual music stimuli and object action videos that either matched (e.g., the sight of a note's being played on a piano together with the corresponding sound, or the video of a bouncing ball with a corresponding sound) or mismatched. This made the authors conclude that semantic congruency affects the strength of intersensory binding in AV speech but that congruency does not affect intersensory binding of nonspeech stimuli such as music or object events.

In the light of even more recent findings, though, this notion needs to be qualified, because *synesthetic congruency* has also been found to affect intersensory binding (Parise & Spence, 2008, 2009). Synesthesia is a condition that has been described as a *mixing* of the senses (Cytowic, 1989). More specifically, in synesthetics, the stimulation of one sense organ leads to an additional perceptual experience in another sensory modality that has not been stimulated at that moment. The most well-known type is that of grapheme–color synesthesia. in which letters or numbers are typically perceived to be colored (Hubbard & Ramachandran, 2005; Rich, Bradshaw, & Mattingley, 2005; Rich & Mattingley, 2002). Although the prevalence of this kind of strong synesthesia is only sparse, there is growing support that neurocognitively normal individuals experience some form of synesthetic associations as well. Synesthetic associations have, for instance, been demonstrated between the visual dimensions of brightness, lightness, size, and shape and the auditory dimensions of pitch and loudness (Evans & Treisman, 2010; Gallace & Spence, 2006; Makovac & Gerbino, 2010; Marks, 1987; Walker et al., 2010). In line with the idea that synesthetic congruency affects intersensory binding, Parise and Spence (2009) reported that sensitivity for the AV temporal order of synesthetically congruent pairs (i.e., a high-pitched tone with a small-sized visual stimulus or a low-pitched tone with a large-sized visual stimulus) was worse than that for incongruent pairs (a low-pitched tone with a small-sized visual stimulus, or a high-pitched tone with a large-sized visual stimulus). In an additional experiment, the effect of synesthetic congruency on judgments of AV *spatial* conflicts was tested, and in line with the findings in the temporal domain, it was shown that participants were less sensitive to AV spatial conflicts whenever stimulus pairs were congruent. This led the authors to conclude that the strength of intersensory interactions is affected by synesthetic congruency.

A similar role of synesthetic congruency was also found in a *visual* TOJ task where the temporal order of the two relevant modalities (audition and vision) was not explicitly at stake (Parise & Spence, 2008). The visual TOJ task allows one to measure the effect of sound on vision in an indirect way—namely, via *temporal ventriloquism*, which refers to the phenomenon whereby a transient sound (or tap) in temporal proximity to a light attracts the temporal occurrence of that light (Keetels, Stekelenburg, & Vroomen, 2007; Keetels & Vroomen, 2008; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Vroomen & Keetels, 2006). Typically, participants are presented pairs of lights at various SOAs and are asked to judge which of the two lights appeared first. Temporal ventriloquism manifests itself in that a task-irrelevant sound before the first light (at ~100 ms) and a second sound after the second light (also at ~100 ms) improve sensitivity, *if compared with a baseline condition in which sound onsets are synchronized with visual onsets*. This finding has been taken as a particularly clear demonstration that the two sounds capture the onset of the two lights and effectively pull them further apart in time. The most cited explanation of this phenomenon is that there is a genuine crossmodal attraction of vision toward audition that is driven by a tendency of the brain to dissolve any conflict between the senses about events that should normally yield converging data (see also de Gelder & Bertelson, 2003; Keetels et al., 2007; Keetels & Vroomen, 2007, 2008; Vroomen & de Gelder, 2000, 2004; Vroomen & Keetels, 2009; see Vroomen & Keetels, 2010, for a review). In terms of the present study, it is of crucial importance to note that the exact size of the temporal ventriloquist effect—the size of the perceived shift of vision toward audition—is measured by comparing a condition in which the sounds are *asynchronous* with the lights (~100-ms AV interval) with a baseline condition in which the sounds are *synchronous* (0-ms AV interval). The choice of the baseline is of importance, if only because synchronized sounds usually yield better JNDs than does a visual-only silent condition, possibly because synchronized sounds increase the reliability of the visual signal (Vroomen & Keetels, 2010). The advantage of using synchronized sounds as a baseline, rather than a visual-only condition or no baseline at all, is that possible effects of the sole presentation of the sound are canceled out and, in this way, the pure effect of the *shift* of the visual stimulus toward the sound can be measured (see also the Method sections in Keetels et al., 2007; Keetels & Vroomen, 2008; Morein-Zamir et al., 2003; Vroomen & Keetels, 2006). As will be explained, this is in contrast to the use of a visual-only baseline in previous studies (Parise & Spence, 2008).

Of importance is that Parise and Spence (2008) reported that the temporal ventriloquist effect was bigger for synesthetically congruent than for incongruent pairs. In their task, participants had to judge the order in which a small and a large circle appeared on a screen. The two visual stimuli were preceded and followed (at 150-ms intervals) by a high- and low-pitched sound. The sensitivity in the visual TOJ task was better if the sounds and lights were synesthetically congruent than if they were incongruent, presumably because there was more intersensory binding for congruent pairs. According to the authors, congruent sounds were better able to pull the lights apart than were incongruent sounds.

Of crucial importance, though, the advantage of the congruent AV pairings in the study by Parise and Spence (2008) may stem from a simple response bias, since participants may have a tendency to report *small circle first*" whenever a high-tone came first (or vice versa). Such a simple response bias will, in their specific setup, result in more correct responses and better JNDs for congruent than for incongruent pairings (see Fig. 1). This response bias can, in principle, be subtracted out if it is compared with a 0-ms AV interval baseline condition. If synesthetic congruency indeed genuinely affects temporal ventriloquism, one would expect better JNDs in congruent than in incongruent pairings at AV intervals of 150 ms, rather than 0 ms. Yet another alternative interpretation of the results of Parise and Spence (2008) is that participants were aware of the synesthetic associations between the auditory and visual modalities and that they become confused whenever the pairs were incongruent (see Fig. 1b). Confusion by incongruent pairs, rather than temporal ventriloquism, would then be the mechanism that causes the congruency effects in Parise and Spence (2008).

To check these alternatives, we reexamined the effect of synesthetic congruency on temporal ventriloquism while avoiding the possible confounds mentioned above. Here, we added a baseline condition in which sounds were presented in synchrony with the visual stimuli—thus, at a 0-ms sound–light interval. Relative to this baseline, we expected sensitivity to improve (smaller JNDs) in asynchronous conditions, in which sounds would pull the two lights further apart due to temporal ventriloquism. If temporal ventriloquism is sensitive to synesthetic congruency manipulations, this effect (i.e., the difference between the synchronous and asynchronous AV intervals) should be bigger for congruent than for incongruent pairs. Importantly, possible side effects of confusion, response bias, or other unknown strategic effects will affect both the asynchronous test condition and the synchronous baseline condition equally, and when the size of the temporal ventriloquist effect relative to the synchronous baseline is calculated, these possible—unwanted—side effects will be subtracted.

We made an effort to reproduce the results of Parise and Spence (2008) as much as possible. We therefore manipulated pitch–size congruency in a visual TOJ paradigm and included the sound–light interval of 150 ms. Furthermore, we added a sound–light interval of 0 ms, because that served as a baseline, and an interval of 75 ms, because it is halfway between the two and is close to where we observed, in previous studies, the maximum temporal ventriloquist effect to occur (~100 ms). In an attempt to further increase the possible effects of synesthetic congruency, we explicitly asked about half of our participants to discriminate congruent from incongruent trials in a session preceding the experiment proper. It is important to note that we did not intend this group to learn a *new* synesthetic association; we intended only to make them *aware* of the naturally existing association.The other group of participants received no training at all, as in the study by Parise and Spence (2008).

Method

Participants Thirty students from Tilburg University participated in return for course credit. Sixteen of them received an explicit pretraining in pitch/size congruency before testing proper started. All reported normal hearing and normal or corrected-to-normal vision. They were tested individually and were unaware of the purpose of the experiment.

Stimuli Visual stimuli were presented on a 17-in. CRT screen (refresh rate, 60 Hz) with the participant's head resting on a chinrest at 55 cm. The visual stimuli consisted of small-sized and large-sized gray circles (3 cm/3.1°, and 5 cm/5.2° diameters, respectively) presented against a dark background. The centers of the circles were at 5 cm/5.2° to the left and right of a small white fixation cross at the center

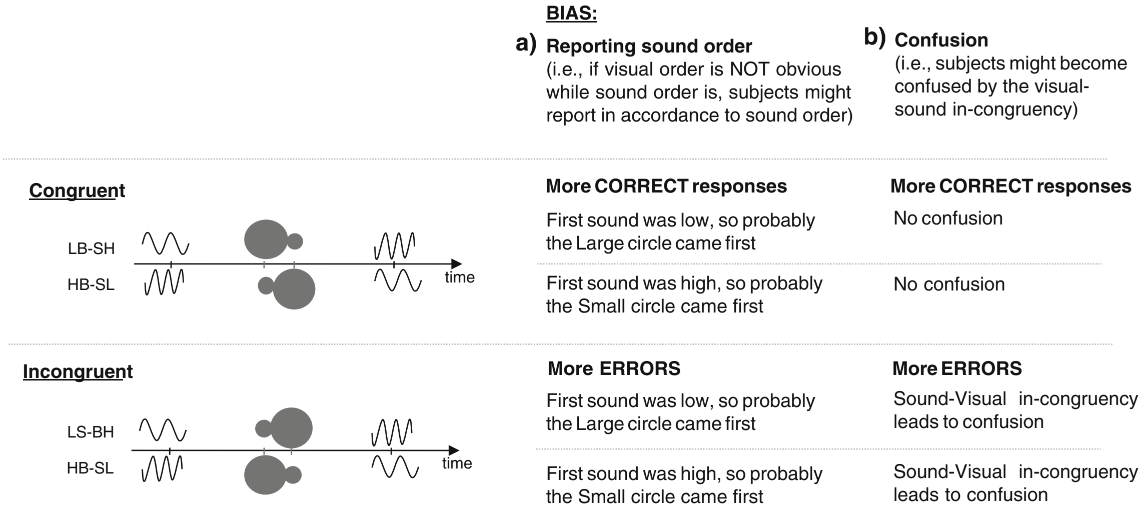

BIAS:
a) Reporting sound order
(i.e., if visual order is NOT obvious while sound order is, subjects might report in accordance to sound order)
b) Confusion
(i.e., subjects might become confused by the visual-sound in-congruency)
Congruent
LB-SH
HB-SL
time
More CORRECT responses
First sound was low, so probably the Large circle came first
First sound was high, so probably the Small circle came first
More CORRECT responses
No confusion
No confusion
Incongruent
LS-BH
HB-SL
time
More ERRORS
First sound was low, so probably the Large circle came first
First sound was high, so probably the Small circle came first
More ERRORS
Sound-Visual in-congruency leads to confusion
Sound-Visual in-congruency leads to confusion

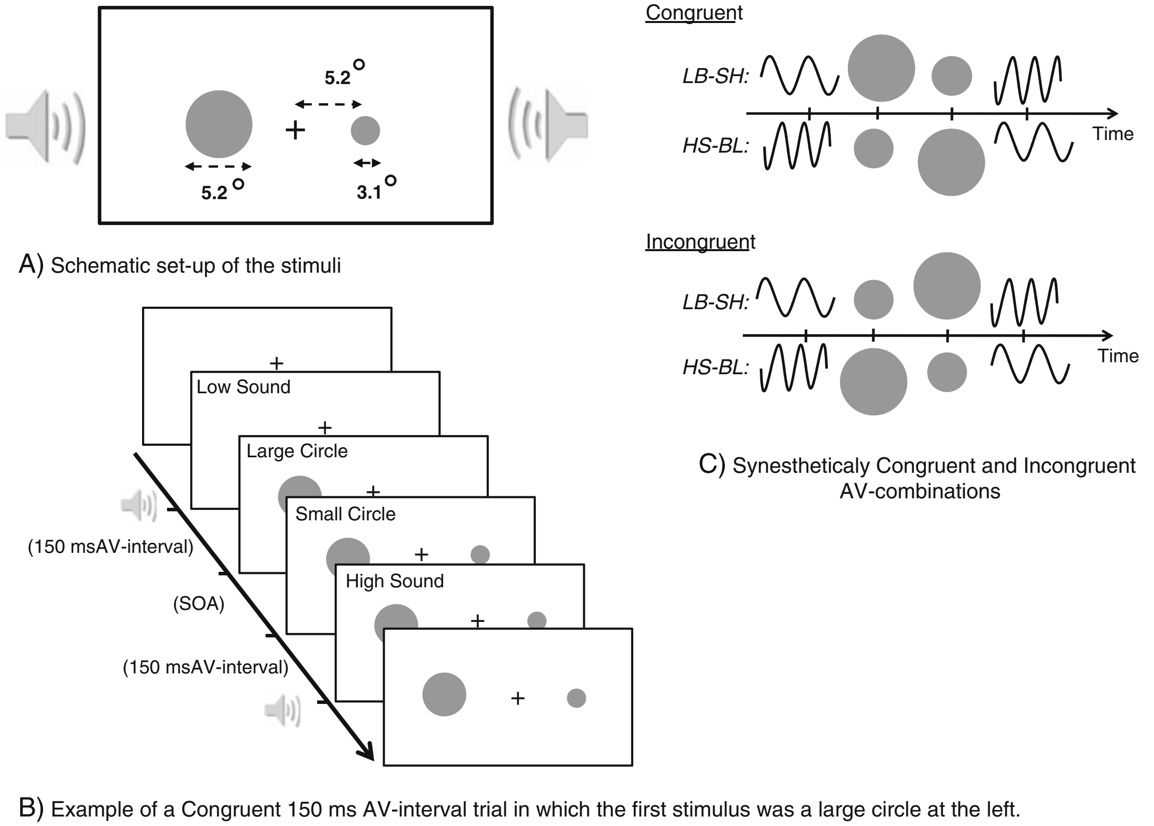
Fig. 1 Schematic representation of the stimuli used in Parise and Spence (2008). A small and a large circle were sequentially presented and were accompanied by a high-tone and a low-tone sound in a synesthetically congruent or incongruent fashion. Participants might be biased to report "large circle first" whenever a low-tone sound came first (or vice versa for small circles). Such a response bias would improve the just-noticeable differences (JNDs) for congruent trials and worsen it for incongruent trials. Alternatively, incongruent trials might lead to confusion, worsening the JND on these trials. To measure temporal ventriloquism in a genuine way, a control condition is required that excludes these alternatives

of the screen (see Fig. 2a). Auditory stimuli were presented via two loudspeakers that were placed directly at the left and right sides of the monitor. The sounds consisted of a low-pitched and high-pitched pure sine-wave tones (frequency of 300 and 4500 Hz, respectively) of 5 ms presented at approximately 75 dB(A) and emanating from the center.

Design There were three within-subjects factors: the AV intervals between the first sound and first light and between the second light and second sound (0, 75, or 150 ms), synesthetic congruency (congruent pairs, a high-pitched tone paired with a small-sized circle and a low-pitched tone paired with a large-sized circle; Incongruent pairs, a high-pitched tone paired with a large-sized circle and a low-pitched tone paired with a small-sized circle) and SOA between the two circles (±83, ±67, ±50, ±33, or ±17 ms; with negative values indicating that the larger visual stimulus was presented first). On half of the trials the small circle was presented first, and on the other half, the large came first. Also, on half of the trials, the small circle appeared on the right, and on the other half, it appeared on the left. A visual-only condition was added as an extra control to confirm that sounds were of help, rather than that they caused interference. Together, these factors yielded 280 unique equiprobable conditions. To equate total testing time, each unique trial was presented 4 times to participants who did not get an explicit training in congruency (four blocks of 280 random trials each) and 3 times to participants who did receive the training (three blocks of 280 random trials each). All the participants also received practice in the TOJ task, in which 28 trials with the largest SOAs (±83 ms) were presented. Participants received feedback (correct/wrong) during this part.

Pitch/size congruency training Participants who received training in pitch/size congruency were presented the same stimuli as those in the main task. They were explicitly told about and given examples of congruent and incongruent sound–light pairs (i.e., for simplicity, the terms *matched* and *mismatched* pairs were used). Whenever they clearly understood and perceived the crucial difference, the main session started, in which participants had to indicate whether the sound–light pairs were synesthetically congruent or incongruent. The session consisted of 96 trials on which circles were presented at the largest SOAs (2 SOAs at ±83 ms × 2 congruency × 3 AVinterval × 2 locations of the first visual stimulus × 4 repetitions). After each trial, participants received feedback about their performance (correct or wrong). Following this training, practice in the visual TOJ task started.

Procedure Participants sat at a table in a dimly lit and soundproof booth. The fixation cross was presented at the beginning of the experiment, and participants were instructed to maintain fixation on this cross during testing. At the beginning of a trial, two gray circles appeared with a variable SOA; one on the left and the other on the right of fixation (see Fig. 2b). The participant's task was to judge which stimulus came first (left or right). Responses were given by pressing one of two corresponding buttons on a response box. Both circles remained visible until a response was



A) Schematic set-up of the stimuli

B) Example of a Congruent 150 ms AV-interval trial in which the first stimulus was a large circle at the left.

C) Synestheticaly Congruent and Incongruent AV-combinations

Fig. 2 A Schematic layout of the stimuli. Participants viewed a fixation cross while a small and a large circle were presented on the left and right of fixation. B The timing diagram of a synesthetically congruent trial at an audiovisual interval of 150 ms. C Schematic representation of the congruent and incongruent audiovisual combinations. L, low-tone sound; H, high-tone sound; B, big circle; S, small circle

given, after which the screen turned black. The next trial started after a random interval between 1,000 and 2,000 ms.

Results

Performance at training in synesthetic congruency was almost flawless, so participants were well able to discriminate congruent from incongruent trials. Trials in the TOJ practice session were excluded from the analyses. Responses were pooled according to whether the first visual stimulus was small or large or was presented on the left or right. The individual proportion of *large-circle-first* responses was calculated for each SOA and for each of the seven conditions (proportions were based on 16 and 12 repetitions for participants without and with training in synesthetic congruency, respectively). These proportions were then converted into equivalent *Z* scores assuming a cumulative normal distribution (cf. Finney, 1964). The best-fitting straight line was calculated over the 10 SOAs. These lines' slopes and intercepts were used to determine the JND (0.675/slope) and the point of subjective simultaneity(PSS). The JND represents the smallest interval between the onsets of the two visual stimuli needed for participants to correctly judge which stimulus had been presented first on 75% of the trials. The PSS represents the average interval at which the participant is maximally uncertain about the order of the visual stimuli. This is conventionally taken to be the interval at which perception of simultaneity is maximal (a positive PSS represents a preference for small-circle-first

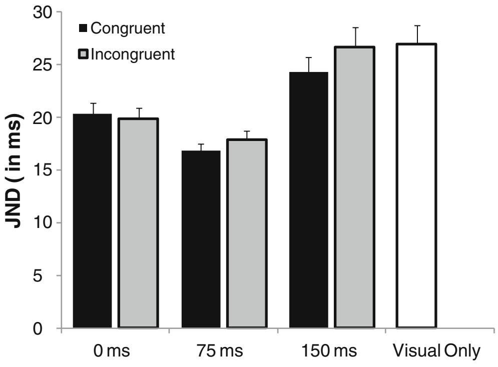


Fig. 3 Mean just-noticeable differences (JNDs) for congruent and incongruent trials (i.e., pooled over with or without training in synesthetic congruency). Error bars represent 1 *SEM*

Table 1 Mean Just-Noticeable Differences (JNDs) and Points of Subjective Simultaneity (PSSs; in Parentheses) for Synesthetically Congruent and Incongruent Audiovisual (AV)Stimulus Pairs at 0-, 75-, and 150-ms AV Intervals, the Upper Part for Participants who Did Not Receive Prior Training in Congruency, the Lower Part for Participants Who Did Receive Prior Training. Right Column: JNDs and PSSs for the Visual-Only Condition

	AV Interval	Congruent JND (PSS)	Incongruent JND (PSS)	Visual Only JND (PSS)
No training	0 ms	22.3 (14.6)	20.7 (15.7)	28.9 (21.7)
	75 ms	18.2 (11.0)	19.8 (14.1)	
	150 ms	26.3 (21.2)	27.5 (17.9)	
With training	0 ms	18.6 (17.3)	19.1 (13.6)	25.3 (22.7)
	75 ms	15.7 (9.7)	16.1 (10.9)	
	150 ms	22.5 (18.0)	25.9 (22.1)	

responses). Fig. 3 and Table 1 give an overview of the mean JNDs and PSSs.

The JNDs were submitted to a 2 (with or without training in synesthetic congruency) × 2 (synesthetic congruency) × 3 (AV interval) overall ANOVA. There were significant effects of AV interval, $F(2, 56) = 40.23$, $p < .0001$, since the group-averaged JNDs varied between 20.2, 17.4, and 25.6 ms for AV intervals of 0, 75, and 150 ms, respectively. The interaction between congruency and AV interval was significant, $F(2, 56) = 3.53$, $p=.036$, but none of the other effects were, all ps $> .14$.

In order to further test these data, contrast analyses were run in which the effect of the 75- and 150-ms AV intervals were contrasted against baseline (i.e., 0-ms AV interval). The JNDs at the 75-ms AV-intervals were significantly *lower,* as compared with the synchronous baseline, thus indicating temporal ventriloquism, $F(1, 28) = 23.37$, $p < .0001$. However, at the 150-ms AV interval, JNDs were found to be *higher,* as compared with the synchronous baseline, $F(1, 28) = 25.48$, $p < .0001$. So, at the 150-ms AV-interval, sounds did not improve but hampered performance, when compared with baseline. Furthermore, this disturbing effect at the 150-ms AV interval was larger for incongruent than for congruent stimulus pairs (a 6.8-ms vs. 3.9-ms effect, respectively), $F(1, 28) = 6.70$, $p = .015$. The effect of training and all other effects were nonsignificant, all ps $> .12$.

We also checked whether we could replicate the original congruency effect reported by Parise and Spence (2008) by directly comparing the JNDs of the congruent and incongruent conditions at the 150-ms AV interval. The results indeed showed slightly lower JNDs for congruent (24 ms) than for incongruent (26 ms) pairs, $t(29) = 2.16$, $p = .039$. These results closely match those of Parise and Spence (2008), who reported JNDs of 21 and 25 ms for congruent and incongruent pairs, respectively.

As a final control, we checked whether sounds presented at appropriate AV intervals were indeed of help, when compared with the silent visual-only condition. All JNDs were indeed lower than those in the visual-only condition (all ps $< .001$ after Bonferroni correction), except when sounds were presented at AV intervals of 150 ms [congruent, $t(29) = 1.97$, $p = .06$, incongruent, $t(29) = 0.26$, $p = .80$].

PSSs For completeness, we also analyzed the PSSs, although there was no specific prediction. In the 2 (with or without training in synesthetic congruency) × 2 (synesthetic congruency) × 3 (AV interval) overall ANOVA, there was an effect of AV interval, $F(2, 56) = 19.44$, $p < .0001$, but none of the other effects was significant, all ps $> .05$. The average PSSs were 17.9, 11.4, and 22.7 ms for the 0-, 75-, and 150-ms AV intervals, respectively. Separate one-sample t tests revealed that all PSSs were bigger than 0, indicating a preference for *small-first* responses (all Bonferroni corrected ps $< .006$).

Discussion

The primary aim of the present study was to examine whether synesthetic congruency between sound pitch and visual size affects intersensory binding in the temporal domain. To examine this, we measured the size of the temporal ventriloquist effect (a shift in the perceived temporal occurrence of a light by an asynchronous sound) for congruent and incongruent sound–light pairs. Participants were presented small and large circles at various SOAs and judged their relative temporal order. High-pitched and low-pitched sounds were presented before the first and after the second circles in either a synesthetically congruent (high-pitch/small-size and low-pitch/large-size) or incongruent (high-pitch/large-size and low-pitch/small-size) way. At large sound–light intervals (~150 ms), JNDs were lower for congruent than for incongruent pairs, confirming previous reports (Parise & Spence, 2008). However, this effect could not be attributed to a proper temporal ventriloquist effect, because sounds at this long AV interval *deteriorated,* rather than improved, performance, if compared with a baseline in which sounds were synchronized with the lights (at ~0 ms). At an intermediate sound–light interval (~75 ms), there was a regular improvement by asynchronous sounds, indicative of temporal ventriloquism, but, most importantly, no effect of synesthetic congruency was observed here. Moreover, even when participants were made explicitly aware of sound–light congruency during the training session before test, there still was no effect of congruency on temporal ventriloquism. Temporal ventriloquism, as demonstrated at the 75-ms AV interval, thus appears *not* to be affected by synesthetic congruency.

These findings raise the question of whether participants in the present study might simply have been insensitive to our manipulations of synesthetic congruency. To answer this, it should be pointed out that stimulus properties in the present study were comparable to the ones used in Parise and Spence (2008, 2009), in which congruency effects were reported (see below for a more elaborate discussion). Additionally, the fact that congruency effects were not present even when participants were made explicitly aware of congruency before testing makes it very unlikely that participants were insensitive to congruency. Finally, the observation that at large sound-light intervals, sensitivity was better for congruent pairs indicates that congruency was, in some way, noticed.

Assuming that *insensitivity* cannot explain why temporal ventriloquism is unaffected by congruency, how then does this result fit the broader picture where congruency effects have been found? Congruency effects in previous studies have been demonstrated mostly with stimuli that are relatively complex in nature. For example, the effects have been shown to occur for semantically associated meaningful stimuli such as letters and speech sounds (van Atteveldt et al., 2004), images and sounds of common objects (Chen & Spence, 2010; Noppeney et al., 2008), or speaker identity (Noppeney et al., 2008). One might argue, then, that an ecologically meaningful audiovisual association may be necessary for inducing congruency effects. If complexity is indeed crucial, it might become understandable that the relatively low-level physical dimensions of pitch and size did not induce a typical congruency effect.

The notion of *stimulus complexity*, though, does not really clear the picture. For example, Parise and Spence (2009) demonstrated pitch–size congruency effects using noncomplex stimuli, and also several other studies have reported congruency effects using rather simple stimuli such as flashes and beeps (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; Spence, Baddeley, Zampini, James, & Shore, 2003; Zampini, Shore, & Spence, 2003b). Also, others who did use meaningful and complex stimuli, nevertheless failed to observe congruency effects (Koppen, Alsius, & Spence, 2008; Vatakis et al., 2008; Vatakis & Spence, 2006). As an example, Koppen et al. examined semantic congruency using ecologically meaningful stimuli such as animal pictures (cats or dogs) that were either matched or mismatched with the sound of a cat meowing or a dog barking. The effect of semantic congruency was tested on the magnitude of the Colavita effect, an intersensory phenomenon in which participants typically fail to respond to an auditory component of bimodal targets (Colavita, 1974; Colavita, Tomko, & Weisberg, 1976). Their findings showed that semantic congruency did affect the speed and accuracy of the participants' responses (so congruency was noticed), whereas it had no effect on the magnitude of the intersensory Colavita effect. Stimulus complexity in general thus does not appear to be critical for observing congruency effects.

Might it be possible, then, that the effects of meaningful semantic congruency operate on a higher level of perception than do those induced by low-level stimulus properties such as spatial and temporal correspondence that is typically explained in terms of correspondence in the receptive field properties of intersensory neurons (Stein & Meredith, 1993)? Although many studies have focused on the neural processes that are associated with intersensory integration with relatively basic stimuli, the role of multisensory integration in higher order processes is a much less studied topic. In an EEG study by Molholm, Ritter, Javitt, and Foxe (2004), it was shown that the intersensory integration between an animal picture (i.e., a dog, cat, frog, etc.) and a congruent animal sound affected object processing at a relatively early stage in the information-processing stream: a visual modulation of the N1 component when sound was added. Yet, in another EEG study (Stekelenburg & Vroomen, 2007), congruent and incongruent semantic intersensory associations did not affect early components. More specifically, ecologically valid speech and nonspeech audiovisual events (i.e., syllables and human actions such as handclapping or tapping a cup with a spoon) were found to evoke a speeding-up and suppression of the auditory N1 and P2 amplitudes. However, incongruent audiovisual stimulus pairs induced an equivalent N1 modulation, showing that sensory integration does occur at this level of perception, whereas congruency does not affect this process. Importantly for the present discussion, the mid-latency and late interactions (i.e., P2 modulations) were susceptible to informational congruency and, according to the authors, possibly were indicative of multisensory integration at the associative, semantic, or phonetic level (Stekelenburg & Vroomen, 2007). In order to further explore the levels at which congruency effects for different kind of stimuli occur, it seems critical to examine the time course of intersensory effects—for example, via measuring event-related potentials on congruency manipulations in ecologically meaningful versus simple intersensory associations.

Another relevant dimension for observing congruency effects seems to be the nature of the task. This has been most clearly demonstrated in studies that have examined whether *spatial* congruency between a sound and light affects intersensory binding. One may expect a sound and light to be more strongly bound as a unitary event if they are presented from the same spatial location. In accordance with this view, it has indeed been found that judging the relative temporal order of a sound burst and a light flash in an AV TOJ task is more difficult if the auditory and visual stimuli are presented from the same location, rather than from different locations (Bertelson & Aschersleben, 2003; Keetels & Vroomen, 2005; Zampini, Shore, & Spence, 2003a). This effect may occur because co-located sounds and lights are strongly paired, so that their relative order is lost. From that point of view, though, it is surprising that spatial correspondence does not affect temporal ventriloquism (Vroomen & Keetels, 2006). So, whether a sound is presented from the same or a different location of a light, the temporal shift of the light as induced by the sound is equivalent. Why is temporal ventriloquism insensitive to spatial correspondence, whereas the crossmodal TOJ task is? One possibility is that in spatially discordant trials, extra spatial cues are added that enhance performance in the crossmodal TOJ task (*which came first* can be deduced from *where it came first*), while these cues are irrelevant in the visual TOJ task that has been used to demonstrate temporal ventriloquism.

It remains to be answered how one can explain that synesthetic congruency did affect the crossmodal TOJ task (Parise & Spence, 2009), but not proper temporal ventriloquism in a visual TOJ task. We already mentioned that participants may have remembered the pitch order of the sounds and responded accordingly whenever unsure about the order of the flashes (e.g., *small-first* if the first sound was high-pitched). Another possibility is that they noticed the incongruent synesthetic associations between the sound and light, which may be especially obvious at the large 150-ms interval, and became confused by incongruent pairings. As was proposed by one of the reviewers, it might also be that, at short AV intervals (0 and 75 ms), AV stimuli are so well integrated that there is no additional benefit of congruency, whereas when temporal separation between sound–light pairs increases enough to start breaking down integration (at a 150-ms AV interval), effects of (mis)matching congruency can be observed. Effects of synesthetic congruency might thus be observed only at the border. From that perspective, though, it remains to be explained why congruent sounds did not improve JNDs at 150-ms AV intervals, when compared with the 0-ms condition or even silence, so it leaves one wondering what the congruency effect at the 150-ms AV interval was based on, if sounds were not of any help.

An interesting finding is that in many of the previous studies of temporal ventriloquism, the effect has been demonstrated to be at maximum at ~100-ms intervals, but to only gradually decline when sound–light intervals increase, even up to 300 ms (Morein-Zamir et al., 2003). It thus appears that the optimal time window in the present situation was rather small. One possibility is that this is caused by the nature of the auditory and visual stimuli used here. In previous studies, visual stimuli were delivered by two identical and relatively small LEDs, while the sounds consisted mostly of short broadband noise bursts. One might speculate that small visual stimuli are less well defined in time, making them more susceptible to being captured by sounds, whereas noise bursts are more well defined in time than are pure tones (Blauert, 1997), making noise bursts particularly potent to capture visual stimuli (see also Spence & Driver, 2000, where pure tones were found to be more susceptible to spatial ventriloquism than were white noise bursts).

Another interesting finding is that all PSSs were positive, reflecting a bias toward *small-first* responses (see also Parise & Spence, 2008). A few findings might be related to this. One is that participants have an overall preference to see a dimmer stimulus first, when compared with a brighter one (Bachmann, Poder, & Luiga, 2004). Our *small-first* preference might be related to this finding in such a way that the small stimulus is probably also seen as the dimmer one simply because it has a smaller surface (i.e., so less light is emitted from the monitor). It has also been shown that when two stimuli are presented simultaneously at different depth positions, observers perceive the distant stimulus before the nearer one (Ichikawa, 2009). Again, this preference for distance first may play a role here, in that the smaller stimulus might be perceived as farther away because, in general, a smaller surface on the retina represents a farther object. At present, though, these explanations are only speculative, and no strong conclusions can be drawn about the mechanisms underlying these preferences.

For completeness, we also have to consider whether any difference in the stimulus setup between the present study and the one by Parise and Spence (2008) might have caused a different outcome. Whereas Parise and Spence (2008) used gray disks displayed on a white background, we used light-gray circles on a gray background. Also, our small visual stimulus was slightly bigger than the one used by Parise and Spence (2008); 3.1° vs. 2.1°, respectively, and in our study, the different AV- ntervals varied randomly between trials, whereas Parise and Spence (2008) used a fixed 150-ms AV interval. As concerns the size of the stimuli, we have to point out that the correspondence in pitch height and visual size is a relative attribute; What is “low-pitched” or “large-sized” in one setting, may be ”high-pitched” or ”small-sized” in another (see also Parise & Spence, 2009), so there is little reason why this should be of importance. As concerns the random variation in AV interval between trials, we are of the opinion that this is the best possible solution, since varying the interval between trials prevents the occurrence of temporal recalibration, a phenomenon in which participants adjust their perception of AV synchrony to a repeatedly presented AV interval (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Vroomen, Keetels, de Gelder, & Bertelson, 2004; see Vroomen & Keetels, 2010, for a review). To us, it seems rather unlikely that these differences in stimuli and design somewhat blurred the picture or caused an absence of a congruency effect on temporal ventriloquism.

To conclude, here we demonstrated that synesthetic congruency between sound pitch and visual size does *not* affect temporal ventriloquism, when measured against a baseline that excludes response biases. Clearly, though, there are various inconsistencies in the literature on the effect of this specific form of synesthetic congruency on intersensory perception that deserve further research.

Acknowledgment M.K. is supported by NWO-VENI Grant 451-08-020.

References

Bachmann, T., Poder, E., & Luiga, I. (2004). Illusory reversal of temporal order: The bias to report a dimmer stimulus as the first. *Vision Research, 44*, 241–246. doi:S0042698903006886

Bedford, F. L. (1989). Constraints on learning new mappings between perceptual dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 232–248.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachmann, & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347–363). Amsterdam: North-Holland.

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.

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization*. Cambridge, MA: MIT Press.

Calvert, G., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114*, 389–404.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*, 409–412.

Colavita, F. B., Tomko, R., & Weisberg, D. (1976). Visual prepotency and eye orientation. *Bulletin of the Psychonomic Society, 8*, 25–26.

Cytowic, R. E. (1989). *Synesthesia: A union of the senses*. New York: Springer.

de Gelder, B. (2000). Recognizing emotions by ear and by eye. In R. D. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion* (pp. 84–105). New York: Oxford University Press.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*, 460–467.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion, 14*, 289–311.

Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences, 98*, 10006–10010.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1, Art. 6), 1–12.

Finney, D. J. (1964). *Probit analysis*. Cambridge: Cambridge University Press.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*, 773–778.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*, 1191–1203.

Hubbard, E. M., & Ramachandran, V. S. (2005). Neurocognitive mechanisms of synesthesia. *Neuron, 48*, 509–520.

Ichikawa, M. (2009). Illusory temporal order for stimuli at different depth positions. *Attention, Perception, & Psychophysics, 71*, 578–593. doi:10.3758/APP.71.3.578

Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: Evidence from temporal ventriloquism. *Experimental Brain Research, 180*, 449–456.

Keetels, M., & Vroomen, J. (2005). The role of spatial disparity and hemifields in audio-visual temporal order judgments. *Experimental Brain Research, 167*, 635–640.

Keetels, M., & Vroomen, J. (2007). No effect of auditory-visual spatial disparity on temporal recalibration. *Experimental Brain Research, 182*, 559–565.

Keetels, M., & Vroomen, J. (2008). Tactile-visual temporal ventriloquism: No effect of spatial disparity. *Perception & Psychophysics, 70*, 765–771.

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research, 184*, 533–546. doi:10.1007/s00221-007-1120-z

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.

Makovac, E., & Gerbino, W. (2010). Sound-shape congruency affects the multisensory response enhancement. *Visual Cognition, 18*, 133–137.

Marks, L. E. (1987). On cross-modal similarity: Perceiving teomporal patterns by hearing, touch, and vision. *Perception & Psychophysics, 42*, 250–256.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*, 452–465.

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

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex, 18*, 598–609. doi:10.1093/cercor/bhm091

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters, 442*, 257–261. doi:10.1016/j.neulet.2008.07.010

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*, e5664. doi:10.1371/journal.pone.0005664

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

Rich, A. N., Bradshaw, J. L., & Mattingley, J. B. (2005). A systematic, large-scale study of synaesthesia: Implications for the role of early experience in lexical-colour associations. *Cognition, 98*, 53–84.

Rich, A. N., & Mattingley, J. B. (2002). Anomalous perception in synaesthesia: A cognitive neuroscience perspective. *Nature Reviews. Neuroscience, 3*, 43–52.

Spence, C., Baddeley, R., Zampini, M., James, R., & Shore, D. I. (2003). Multisensory temporal order judgments: When two locations are better than one. *Perception & Psychophysics, 65*, 318–328.

Spence, C., & Driver, J. (2000). Attracting attention to the illusory location of a sound: Reflexive crossmodal orienting and ventriloquism. *NeuroReport, 11*, 2057–2061.

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

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*, 271–282. doi:10.1016/j.neuron.2004.06.025S0896627304003964

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, Art. 14), 11-11. doi:10.1167/8.9.14/8/9/14/

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111*, 134–142.

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

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. doi:10.1016/j.actpsy.2006.12.002

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

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 513–518.

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.

Vroomen, J., & Keetels, M. (2009). Sounds change four-dot masking. *Acta Psychologica, 130*, 58–63. doi:10.1016/j.actpsy.2008.10.001

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*, 871–884.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22*, 32–35.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., et al. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*, 21–25.

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.

Zampini, M., Shore, D. I., & Spence, C. (2003a). Audiovisual temporal order judgments. *Experimental Brain Research, 152*, 198–210.

Zampini, M., Shore, D. I., & Spence, C. (2003b). Multisensory temporal order judgments: The role of hemispheric redundancy. *International Journal of Psychophysiology, 50*, 165–180.