Psychonomic Bulletin & Review
2010, 17 (3), 329-334
doi:10.3758/PBR.17.3.329

Correlation versus causation in multisensory perception

Holger Mitterer and Alexandra Jesse
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Events are often perceived in multiple modalities. The co-occurring proximal visual and auditory stimuli events are mostly also causally linked to the distal event, which makes it difficult to evaluate whether learned correlation or perceived causation guides binding in multisensory perception. Piano tones are an interesting exception: They are associated with the act of the pianist striking keys, an event that is visible to the perceiver, but directly results from hammers hitting strings, an event that typically is not visible to the perceiver. We examined the influence of seeing the hammer or the keystroke on auditory temporal order judgments (TOJs). Participants judged the temporal order of a dog bark and a piano tone, while seeing the piano stroke shifted temporally relative to its audio signal. Visual lead increased "piano-first" responses in auditory TOJ, but more so if the associated keystroke was visible than if the sound-producing hammer was visible, even though both were equally visually salient. This provides evidence for a learning account of audiovisual perception.

Events are often perceived in more than one modality. Observers combine evidence from multiple senses to improve perception. One domain in which multisensory input improves perception is, for example, the perception of speech. In situations in which auditory speech is difficult to understand, an additional benefit of seeing the speaker can be shown (Sumby & Pollack, 1954). Visible speech information is also used when acoustic speech information is unambiguous. This can be shown when auditory and visual signals mismatch due to experimental manipulations (McGurk & MacDonald, 1976). Here, the mismatching visual speech alters the overall percept, even when participants are instructed to ignore visual information (Summerfield & McGrath, 1984).

Theories that explain how listeners come to combine information from the two speech channels can be grouped into two broad classes: one for those that suggest that the effect is a result of experience in associating audio and visual speech signals (Diehl & Kluender, 1989; Massaro, 1998) and another for those that assume that experience is not necessary, because perception is tuned to directly perceive the distal causes of proximal events in these modalities (Fowler & Dekle, 1991). In short, the question is whether the learned correlation of visual and auditory stimuli or their perceived common causation drives binding in multisensory perception.

Learning accounts, such as those that are implemented in the fuzzy logical model of perception (FLMP), assume that listeners learn the auditory and visual correlates of syllables in their native language (Massaro, 1998). According to the FLMP, experience provides the observer with summary descriptions of these correlates stored as prototypes in memory, against which incoming information from each modality is evaluated. Learning accounts, therefore, assume that information about the associated visible event, but not necessarily information about the sound-producing event, is used.

According to the theory of direct realism (Fowler, 1996; Gibson, 1979), audiovisual integration is based on the pick-up of auditory, visual, and other information that specifies the distal event. Learning improves the observer's preparedness to pick up information, but multisensory integration relies not on learning but, rather, on lawfully generated stimulus information that specifies its source in multiple modalities, so that common causation drives integration (Rosenblum, Schmuckler, & Johnson, 1997).

To disentangle whether correlation or causation drives multisensory perception, Fowler and Dekle (1991) examined the relative influence of both associated and unassociated, but causally linked, unimodal stimuli on the perception of auditory syllables from a [ba]–[ga] continuum. In one condition, the auditory syllable was paired with associated written representations. In another condition, it was paired with the experience of feeling, but not seeing, the speaker utter the syllable by placing the perceiver's right hand on the lips of the speaker. This unassociated event was nevertheless causally linked to the speech gesture. If learned associations were crucial for multisensory perception, the associated spelled syllable should have a large effect on perception, whereas the felt syllable should not. Because the opposite result was observed, the authors argued that perceivers directly perceive the environmental cause of the proximal stimulation in multiple sensory channels and combine them naturally.

H. Mitterer, holger.mitterer@mpi.nl

One problem in the reasoning is, however, the assumption that the spelled syllable should be closely associated with the acoustic stimulus. Even though there is cross-talk between phonological and orthographic processing for both auditory and written input (e.g., see Frost & Ziegler, 2007), seeing the spelled English syllable “ba” seldom goes together with simultaneously hearing the syllable /bɑː/. This is especially true if one considers the range of possible phonological forms that go together with the bigram “ba” in English ([bə] in “baboon,” [bei] in “baby,” [bæ] in “back,” [bɔː] in “bald,” [bɑː] in “bar,” [bɛə] in “bare”). The printed syllable “ba” is thus not strongly associated with its auditory variant [bɑː]. A second problem in the reasoning is that it is unclear how the haptic information is processed. Massaro (1998, pp. 352–355) argued that haptic prototypes can be acquired during the task from the stored auditory and visual prototypes. This is especially likely, given the explicit and offline nature of the categorical-perception task.

Despite these criticisms, the basic reasoning of Fowler and Dekle (1991) is correct: To distinguish a learning account from a gestural account, correlation and causation have to be dissociated. In speech, correlation and causation usually go together. For example, the opening and closing of the lips not only are associated with the sound of the phoneme /b/, but also are its underlying cause. Piano tones, however, dissociate correlation and causation: The sound of a piano is associated with keystrokes but is produced by hammers that hit strings. This action is (mostly) hidden from observation.

One may argue that the keystroke is still the ultimate cause of the piano sound, because the hammer is caused to move by the keystroke. But given the assumptions of direct realism, it is the sound-producing, causal event that is crucial. Multisensory integration is supposed to be achieved if visual and acoustic stimulation is “lawfully generated by, and therefore fully specificational to, its source events” (Rosenblum et al., 1997, p. 355). Impact events, such as a hammer hitting a string, are assumed to give information about the impacting and impacted material and about the force of the impact (Gaver, 1993). Accordingly, the piano sound should specify a hammer hitting metal strings. But the piano sound cannot specify a keystroke, because keystrokes on organs and harpsichords produce completely different sounds. The dissociation is especially strong for sustained piano sounds, which were used in the present study. A sustained sound indicates a transitory impact, because the movement of the strings is not damped. This is in line with the transitory contact between string and hammer, but not with a sustained keystroke. (Similarly, the sound of a door knock is longer if the knuckles do not stay in contact with the door.) The correlated visual event is thus an unlikely cause of the sound.

We hence tested the relative strength of sound-producing and associated visual events in audiovisual binding. We measured audiovisual binding in an auditory temporal-order judgment (TOJ) task with a concurrent visual stimulus. Participants had to indicate whether a piano sound (the octave A4–A5) or an aperiodic dog bark was earlier. We chose a dog bark because it is acoustically distinct from the piano sound (see Goswami et al., 2002, for a similar periodic/aperiodic stimulus pair) and is not associated with keystrokes. The visual piano stroke led or lagged the piano sound. Due to this asynchrony, audiovisual binding should have the following consequences: If the visual piano stroke leads the piano sound, the perceived onset of the piano sound should be earlier, so that there should be more “piano-first” responses given in auditory TOJs. Experiment 1 established that this is the case when observers see both the hammer and the keystroke.

Experiment 2 assessed whether the influence of seeing the associated visual event (i.e., the fingers hitting the keys) or the sound-producing visual event (i.e., the hammers hitting the strings) leads to stronger audiovisual binding in the auditory TOJ. A learning account predicts that the associated visual event should influence auditory TOJ more strongly than would the sound-producing not-associated event. A direct perception account, however, predicts the opposite result. In analogy to the reasoning in Fowler and Dekle (1991), perceiving the underlying sound-producing event, but not the associated visual event, should influence auditory TOJ.

EXPERIMENT 1

Method

Participants. Twelve members (age, 18–25 years) of the participant pool of the Max Planck Institute with normal hearing and vision participated for pay.

Stimuli and Apparatus. A pianist was videotaped (video: 720 × 576 pixels, 25 fps; audio: 16 bits, 48 kHz) simultaneously playing the octave A4 and A5 on an upright piano, using the first and the fifth digits of his right hand and with minimal wrist movement. The octave was sustained for about 1.5 sec before the fingers moved up again. The upper front board of the piano was removed to make the hammers visible (see Figure 1). We used Adobe Premiere 6.5 to extract a 2.32-sec video (i.e., 58 frames) from the videotape. To create a fade in and out, we added 400-msec transitions from and to black to the first and last frames of the video, respectively.

A dog bark was added to the piano audio track with auditory stimulus onset asynchronies (SOAs) on a quadratic time scale (−144, −64, −16, 0, +16, +64, and +144 msec). SOAs were referenced to subjective points of onsets—that is, to the maximum acceleration of the amplitude envelope (cf. Scott, 1998). These audio tracks were shifted by one frame (40 msec) to the left relative to the video track to create a visual lag, or by two frames (80 msec) to the right to create a visual lead. We chose these values on the basis of results from a pilot study in which participants had to indicate for several intervals whether the auditory or the visual piano stimulus was earlier.[1] A visual lag of one frame (62.4% vision-first responses) and a visual lead of two frames (74.3% vision-first responses) were chosen because they yielded small and similar effects, compared with the synchronous presentation of the piano as a baseline (66.5% vision-first responses), but were not obviously asynchronous. (There was an overall bias toward “vision-first” responses in this pretest.) The resulting 14 videos (7 auditory SOAs with visual lead vs. lag) were encoded in MPEG-1 and presented using the software program Presentation (Neurobehavioral Systems Inc.).

Procedure. Experiments were run in a sound-attenuated booth, with participants facing a computer screen. Audio was presented binaurally over headphones. Participants were instructed that they would hear a dog and hear and see a piano. Their task was to judge

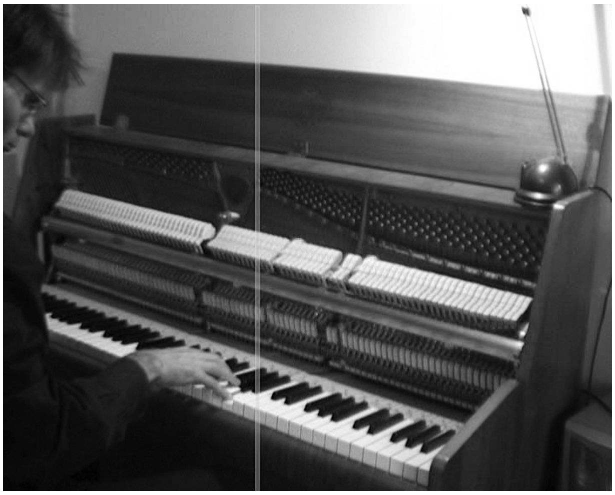

Figure 1. One sample frame of the video stimulus in Experiment 1. The transparent line indicates the parts of the video presented in the different viewing conditions in Experiment 2. For the keys-visible condition, only the part to the left of the line was shown and the other part was covered by a black overlay. For the hammers-visible condition, only the right part was presented, including the point of contact between hammers and strings, and the left part of the video was covered by a black overlay.

whether the dog or the piano sound was earlier by pressing a button on a computer keyboard. Each participant received 20 blocks with randomly permutated orders of the 14 audiovisual stimuli, for a total of 280 trials. Stimulus repetitions at block boundaries were prevented. The experiment lasted about 25 min.

Results and Discussion

Figure 2 shows the results in terms of the "piano-first" responses, which were more likely with a visual lead than with a visual lag, with the exception of the zero auditory SOA. The results were analyzed with a linear mixed-effect regression model, with a logistic linking function because of the categorical dependent variable (cf. Dixon, 2008). Participants were entered as a random factor and with auditory SOA and visual lead/lag as fixed factors. Given the nonlinearity of the visual effect over the auditory SOA range, auditory SOA was treated as a categorical variable. The auditory SOA of -144 msec and the visual-lag condition were mapped on the intercept. The model showed that, compared with the -144-msec SOA at the intercept, all other SOAs produced more "piano-first" responses (all ps $< .001$; regression weights ranging from 1.01, for the SOA $= -64$ msec, to 3.65, for the SOA $= +144$ msec). Moreover, there were more "piano-first" responses in the visual lead than in the visual lag condition at the intercept of -144-msec SOA ($b = 0.99, p < .001$). This visual effect was attenuated significantly only at the auditory SOAs of -16 msec ($b = -0.77, p < .05$) and 0 msec ($b = 0.91, p < .05$).

These results show that audiovisual binding influences the performance on the auditory TOJ task. This is in line with previous research on visual influences on music perception. Schutz and Lipscomb (2007) showed that observers judge a marimba tone to be longer if the manual gesture of the player is longer, even though the sound duration is not influenced by the duration of the gesture.

EXPERIMENT 2

Experiment 2 differed from Experiment 1 only in that only the left or right part (as indicated in Figure 1) of the video was visible—that is, either the key or the hammer stroke, respectively. This allowed us to assess the relative influence of seeing the associated (keystrokes) or the sound-producing (hammer strokes) event on TOJs. The question was whether the effect of visual lead versus lag on auditory TOJs observed in Experiment 1 differs as a function of what is visible. Learning accounts predict a stronger effect of seeing the associated event than seeing the sound-producing event on TOJs. A direct-perception account predicts the opposite.

We added a visual TOJ task as a control condition, in which participants saw, as in the main experiment, videos with either the key or the hammer stroke. A small red

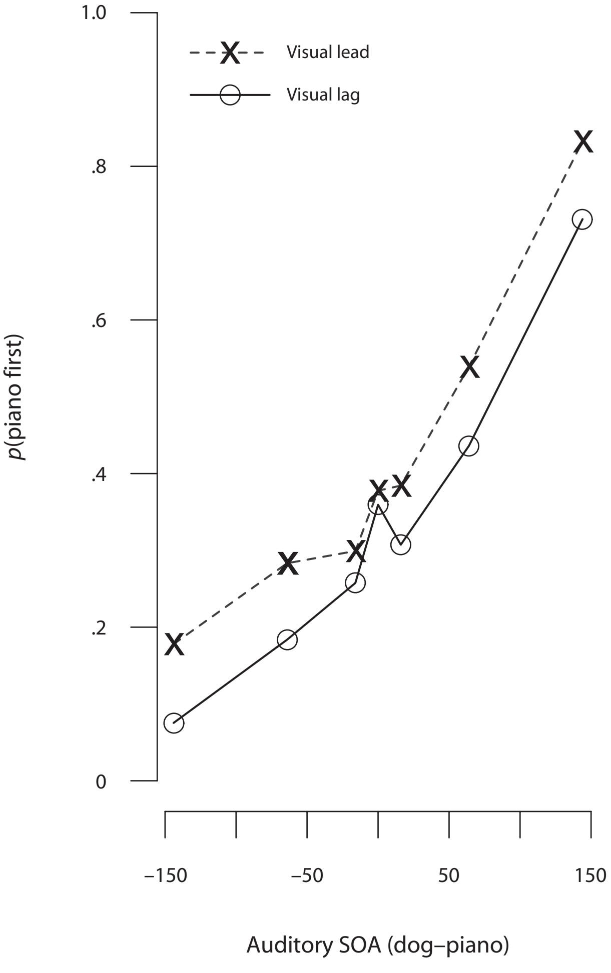

Figure 2. Proportion of "piano-first" responses in the auditory temporal order judgment (TOJ) task in Experiment 1 as a function of the SOA of the auditory piano stimulus and the dog bark and the visual lead/lag of the visual and auditory piano stimulus.

square was added in proximity to the stroke at different SOAs. (The sound was muted in this control condition.) The participants had to decide whether the red square appeared before or after the stroke. This allowed us to test whether either of the two strokes (hammer or key) is more salient than the other and, therefore, might have had a stronger impact on the auditory TOJ task.

Method

Participants. We tested 34 new participants (age, 18–28 years) from the same population as in Experiment 1; 16 participated in the main experiment, and 18 participated in the control condition.

Stimuli and Procedure. The stimuli and procedure for the main experiment were similar to those of Experiment 1. As an additional independent variable, either the left or the right part of the video was shown, so that only the keystroke or only the hammer action was visible (see Figure 1). To keep the number of trials at an acceptable level for participants, only the two endpoints and two intermediate SOAs (±16, ±144 msec) were used. Experiment 1 had shown that there is a visual influence on the auditory TOJ at these SOAs.

Each participant judged each of the 16 stimuli (4 [auditory SOAs] × 2 [hammer vs. keys visible] × 2 [visual lead/lag]) 20 times—that is, each participant received a total of 320 trials. The visibility condition (hammer vs. keys) was blocked and order was balanced across participants. Within each block, participants were presented with 20 subblocks of 8 stimuli, with either the key stroke or hammer stroke visible, randomly permutated and without repetitions at subblock transitions. After the experiment, participants were asked whether they played piano and how often they had seen the hammer action before (in four logarithmic categories: *never*, *1–10*, *10–100*, *more than 100*).

In the control condition, participants were instructed to judge whether the piano or the square appeared first. The SOAs between piano and square ranged from −200 to 200 msec in 40-msec steps. Each of the 11 stimuli was presented 30 times to each participant.

Results and Discussion

Figure 3 shows the proportion of "piano first" responses in the main experiment. The data were analyzed with a linear mixed-effect model with participant as random factor, and with auditory SOA (as a categorical factor), visual lead/lag, and hammers versus keys visible as fixed factors. Nonsignificant interactions were removed from the model. The final model showed that "piano-first" responses were

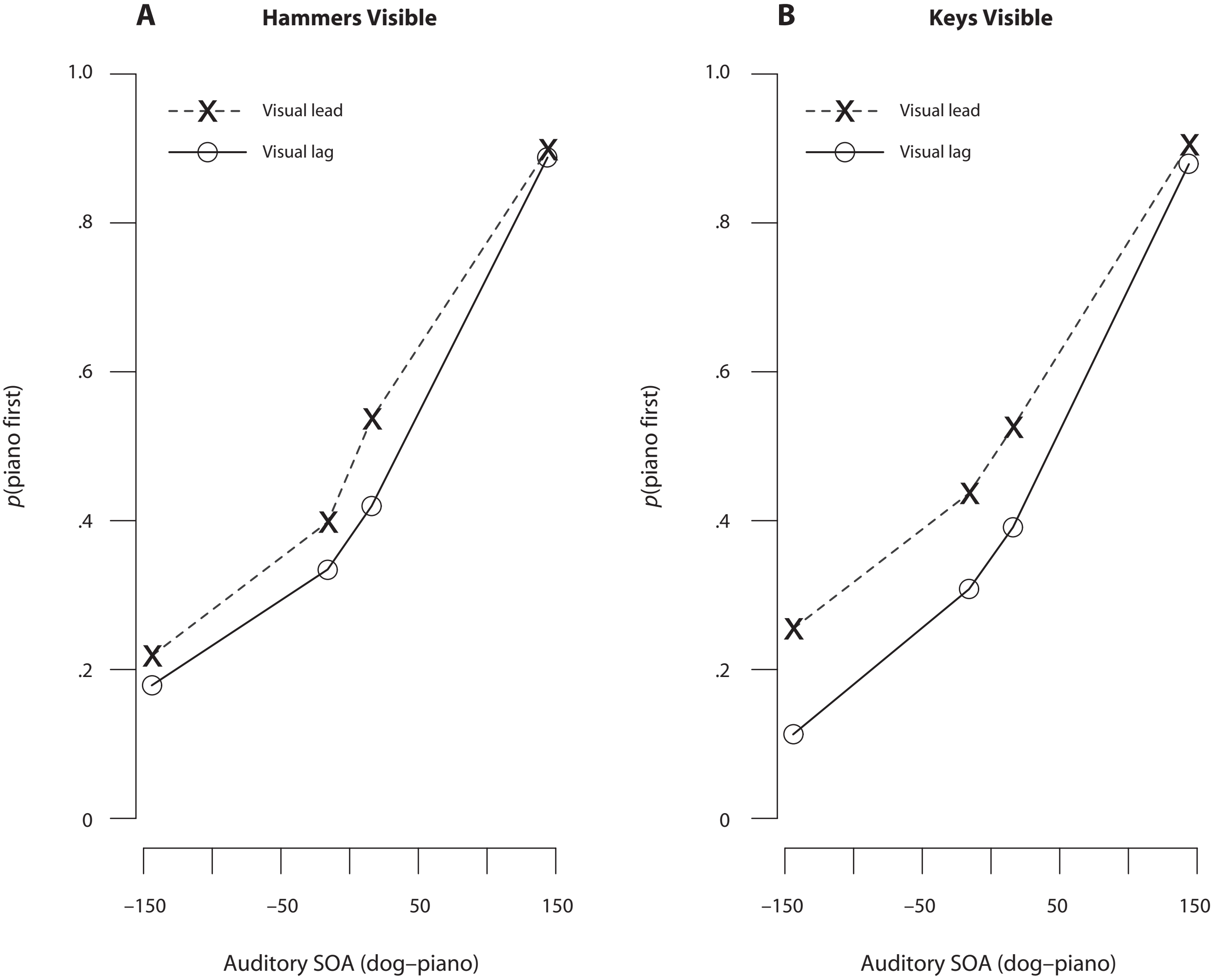

Figure 3. Proportion of "piano-first" responses in the auditory TOJ task in Experiment 2 as a function of the stimulus onset asynchrony (SOA) of the auditory piano stimulus and the dog bark and the visual lead/lag of the visual and auditory piano stimulus. Panel A shows the data obtained when only the hammers were visible, and panel B shows the data obtained when only the keys were visible.

more likely as the auditory SOA increased (min[b] = 0.93, $p < .001$). "Piano-first" responses were more likely with a visual lead than with a visual lag in the hammer-visible condition ($b = 0.32$, $p < .001$). Seeing the hammer was thus sufficient to influence the auditory TOJ. Importantly, the effect of visual lead/lag was enhanced in the keys-visible condition ($b_{\text{visual lead/lag} \times \text{keys visible}} = 0.29$, $p < .05$). The overall difference in "piano-first" responses between the visual lead and lag conditions was 5.8% when the hammers were visible, but was 10.8% when the keys were visible. An additional analysis tested whether the experience with seeing the hammers—as assessed in the postexperiment questionnaire—influenced the size of the effect in the hammer-visible condition. It did not ($p > .2$). The sample included only 2 participants who play the piano, which made it difficult to test whether piano playing has an effect.

In the control condition (see Figure 4), an analysis of the "piano-first" responses gave no rise to an interaction of visibility condition and SOA (max[b] = -0.88, min[p] = .13). After removing the interaction, the model showed an effect of SOA (min[b] = 0.54, $p < .05$) and no overall difference between the hammer- and key-visible conditions ($b = 0.14$, $p = .33$). This shows that the effect in the main experiment cannot be explained by a difference in visibility between the hammers and the keys.

GENERAL DISCUSSION

We used piano tones to test the role of correlation and direct causation in audiovisual perception. Piano tones are an interesting case, because they are associated with nearly inaudible keystrokes but are caused by hammers hitting strings that are typically hidden from observation. We tested whether audiovisual integration is stronger if the correlated event or the sound-producing event was visible. Experiment 1 established a method for measuring the strength of integration. The auditory TOJ for a dog bark and a piano sound was influenced by whether the visual piano display led or lagged the piano sound. If it led, participants were more likely to perceive the piano sound before the dog bark. In Experiment 2, we tested whether this effect is driven by the correlated event (the keystrokes) or only by the sound-producing event (the hammers hitting the strings). Experiment 2 showed that observing the hammers was sufficient for a visual influence, but that the visual influence was larger if the keystroke was visible than if the hammer stroke was visible.

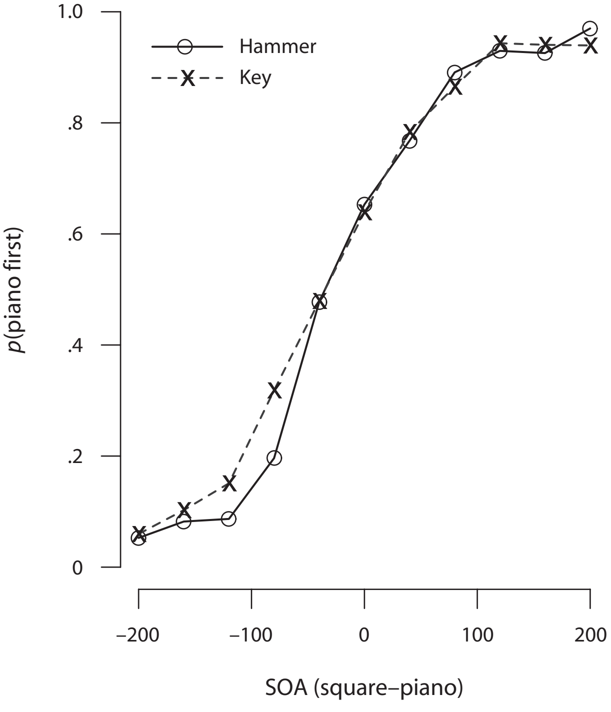

Figure 4. Proportion of "piano-first" responses in a visual TOJ task (control condition of Experiment 2) as a function of the SOA between square and piano.

One question that arises is why seeing the hammers still affects TOJ. This effect was significant, but was only half of the effect observed with the keys. Two accounts for this effect are possible. In a pilot study, participants saw only the hammers and not the piano case. Nevertheless, participants immediately identified the hammer as belonging to a piano. This indicates some experience with recognizing hammers as part of a piano, which may have been sufficient for participants to have learned the association between the hammer stroke and piano tone. Alternatively, participants could have indeed perceived the common cause of piano tone and hammer action, as is predicted by the theory of direct perception. Causation may hence be sufficient for audiovisual binding, but our results show that it is not necessary. In contrast to the prediction of a direct-perception account, our results indicate that learned associations foster multisensory perception, so that binding is stronger for an associated event than for the sound-producing event. This is in line with findings that the reaction of auditory brain areas to visual input is modified by experience. Hasegawa et al. (2004), for instance, found that only well-trained pianists showed activation of the left *planum temporale* when watching a silent video of a piano player. To account for such findings, the theory of direct perception would have to be extended, so that learning is considered to influence not only unimodal information pick-up but also multisensory integration.

Another question arises as to whether the present results can be generalized to speech perception. From the point of view of both direct-realism and learning accounts of multisensory integration, audiovisual speech perception is simply another incidence of multisensory perception, so that no difference would be expected. Indeed, many multisensory phenomena observed in speech perception have analogues in the nonspeech domain (e.g., Saldaña & Rosenblum, 1993).

Nevertheless, Vatakis and Spence (2007, 2008) found some evidence against such a claim: They showed to participants various audiovisual events (e.g., spoken syllables, piano tones, ice being crushed) and varied their audiovisual asynchrony. Additionally, audiovisual stimuli were matching (e.g., audiovisual piano) or mismatching (e.g., piano audio with visual guitar). Participants were asked to decide whether the auditory or the visual stimulus was earlier. Vatakis and Spence found that, if the audio and visual stimuli were from the same source, only for speech stimuli was the audiovisual asynchrony more difficult to detect. This seems to show that audiovisual perception is different for speech stimuli than for other stimuli.

It should be noted, however, that the task used in those experiments tested explicit audiovisual separation; that is, participants had to say whether the auditory or the visual stimulus was first. Our task measured implicit binding. The difference between explicit separation and implicit binding is evident in audiovisual speech perception. Soto-Faraco and Alsius (2009) showed that an (integrative) McGurk effect arises at SOAs at which participants are able reliably to perceive an onset asynchrony between visual and auditory speech. The results of an audiovisual order-judgment task thus do not speak to whether implicit audiovisual binding also differs between speech and nonspeech events.

In summary, our data show that vision can influence the perceived temporal order of two auditory events: Seeing the associated event had a stronger influence than did seeing the sound-producing event. Although the present data do not rule out that this influence can occur without learning, they at least show that learning is sufficient and is more important than actual causation in audiovisual perception.

AUTHOR NOTE

This research was supported in part by an Innovational Research Incentive Scheme Veni Grant awarded to A.J. by the Netherlands Organization for Scientific Research (NWO). The authors thank Sabrina Jung for help with the preparation of the materials and Lies Cuijpers for help with conducting the experiments. Address correspondence to H. Mitterer, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands (e-mail: holger.mitterer@mpi.nl).

REFERENCES

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, **1**, 121-144.

Dixon, P. (2008). Models of accuracy in repeated-measures design. *Journal of Memory & Language*, **59**, 447-456. doi:10.1016/j.jml.2007.11.004

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, **99**, 1730-1741.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 816-828.

Frost, R., & Ziegler, J. C. (2007). Speech and spelling interaction: The interdependence of visual and auditory word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 107-118). Oxford: Oxford University Press.

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, **5**, 1-29.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences*, **99**, 10911-10916. doi:10.1073/pnas.122368599

Hasegawa, T., Matsuki, K.-I., Ueno, T., Maeda, Y., Matsue, Y., Konishi, Y., & Sadato, N. (2004). Learned audio–visual cross-modal associations in observed piano playing activate the left planum temporale: An fMRI study. *Cognitive Brain Research*, **20**, 510-518. doi:10.1016/j.cogbrainres.2004.04.005

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

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

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, **59**, 347-357.

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

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception*, **36**, 888-897.

Scott, S. K. (1998). The point of P-centres. *Psychological Research*, **61**, 4-11. doi:10.1007/PL00008162

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology: Human Perception & Performance*, **35**, 580-587.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

Summerfield, A. Q., & McGrath, M. (1984). Detection and resolution of audio–visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, **36A**, 51-74.

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.

NOTE

[1] Pilot participants saw the audiovisual piano stimulus with audiovisual asynchronies ranging from −200 to +200 msec in 40-msec steps and had to decide whether the auditory or the visual signal was earlier. Each SOA was presented 20 times.

(Manuscript received June 23, 2009;
revision accepted for publication January 11, 2010.)