Exp Brain Res (2010) 203:723–735
DOI 10.1007/s00221-010-2286-3

RESEARCH ARTICLE

Auditory temporal modulation of the visual Ternus effect: the influence of time interval

Zhuanghua Shi · Lihan Chen · Hermann J. Muller

Received: 6 August 2009 / Accepted: 29 April 2010 / Published online: 16 May 2010

Abstract Research on multisensory interactions has shown that the perceived timing of a visual event can be captured by a temporally proximal sound. This effect has been termed 'temporal ventriloquism effect.' Using the Ternus display, we systematically investigated how auditory configurations modulate the visual apparent-motion percepts. The Ternus display involves a multielement stimulus that can induce either of two different percepts of apparent motion: 'element motion' or 'group motion'. We found that two sounds presented in temporal proximity to, or synchronously with, the two visual frames, respectively, can shift the transitional threshold for visual apparent motion (Experiments 1 and 3). However, such effects were not evident with single-sound configurations (Experiment 2). A further experiment (Experiment 4) provided evidence that time interval information is an important factor for crossmodal interaction of audiovisual Ternus effect. The auditory interval was perceived as longer than the same physical visual interval in the sub-second range. Furthermore, the perceived audiovisual interval could be predicted by optimal integration of the visual and auditory intervals.

Z. Shi (✉) · L. Chen · H. J. Müller
Department Psychologie, Ludwig-Maximilians-Universität München, Leopoldstrasse 13, 80802 Munich, Germany
e-mail: shi@psy.lmu.de

L. Chen
Department of Psychology, Peking University, 100871 Beijing, People's Republic of China

H. J. Müller
School of Psychology, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Keywords Time perception · Vision · Audition · Temporal ventriloquism effect · Ternus display

Introduction

Most events we perceive in everyday life consist of simultaneous inputs from different sensory modalities. For example, when knocking at the door, we immediately perceive a sound coupled with a touch feedback. By integrating different sources of sensory information, our brain can achieve accurate representations of the environment (Calvert et al. 2004). However, when different modalities convey inconsistent information, perception is usually biased toward one modality—this has been referred to as modality dominance. For example, a sound source is often localized (or 'captured') toward the location of a light flash which is presented simultaneously at some distance to the sound. This effect is known as 'ventriloquism effect' (Bermant and Welch 1976; Bertelson and Radeau 1981; Howard and Templeton 1966; Radeau and Bertelson 1987). A wealth of literature on multisensory integration has demonstrated phenomena of modality dominance in the spatial domain, with vision being dominant in most cases (Pick et al. 1969; Posner et al. 1976; Soto-Faraco and Kingstone 2004; Welch and Warren 1980, 1986). Besides the classical ventriloquism effect of vision 'capturing' sound, the reversed ventriloquism effect has also been observed when the visual location information was poor (Alais and Burr 2004). Recently, probability-based models, such as Bayesian integration models, have been proposed to provide quantitative accounts of the spatial ventriloquism effect (Alais and Burr 2004; Battaglia et al. 2003).

Interestingly, multisensory interactions also occur in the temporal domain. For instance, the perceived occurrence of

a visual event can be biased by an irrelevant and slightly asynchronous auditory stimulus (Fendrich and Corballis 2001; Morein-Zamir et al. 2003; Scheier et al. 1999; Shimojo et al. 2001). For example, Scheier et al. (1999); see also Morein-Zamir et al. 2003) demonstrated that when two sounds were presented, one slightly before a first flash and the other shortly after a second flash, the sounds attracted the temporal occurrence of the lights, thus improving the visual temporal resolution (i.e., the just noticeable difference, JND). In contrast, when two sounds were inserted in between two light flashes, the visual temporal resolution became worse (Scheier et al. 1999). This phenomenon, referred to as 'temporal ventriloquism effect' (Morein-Zamir et al. 2003), has elicited a great deal of interest in multisensory interaction research. It has been proposed that the temporal ventriloquism effect is related to the modality appropriateness or modality precision (Welch and Warren 1980, 1986). On this hypothesis, the sensory modality with the highest acuity may outweigh the others, so that, for example, audition with its high temporal resolution may dominate temporal perception.

The spatial and temporal constraints on the temporal ventriloquism effect were investigated in a number of follow-up studies (Bruns and Getzmann 2008; Freeman and Driver 2008; Getzmann 2007; Jaekl and Harris 2007; Keetels et al. 2007; Vroomen and Keetels 2006). As there was little influence of the relative spatial positions of the auditory and visual stimuli on the temporal ventriloquism effect (Bruns and Getzmann 2008; Vroomen and Keetels 2006), other studies turned to examining the influence of the temporal configuration on the effect. For example, Morein-Zamir et al. (2003) found that the temporal relationship of audiovisual events was important: with two sounds presented before and, respectively, after the two visual stimuli, the temporal ventriloquism effect occurred only when the second sound was trailing the second light within a range of 200 ms; and with two sounds presented in between the two visual stimuli, a temporal ventriloquism effect was observed only when the sounds were separated by 16 ms. Furthermore, several studies have shown that a single sound leaves temporal-order judgment (TOJ) performance uninfluenced (Morein-Zamir et al. 2003; Scheier et al. 1999; Shimojo et al. 2001). This has been taken to suggest that two sounds are required, one paired directly with each visual event, for the audiovisual stimuli to be perceived as a unitary event (Morein-Zamir et al. 2003; Welch 1999). Similar results have also been reported for the audiovisual apparent-motion paradigm (Bruns and Getzmann 2008; Freeman and Driver 2008; Getzmann 2007). For example, a temporal ventriloquism effect has been observed in an ambiguous bidirectional visual apparent-motion stream by introducing auditory beeps slightly lagging or leading the flashes (Freeman and Driver 2008). More recently, intramodal perceptual grouping has been shown to be an important factor in crossmodal integration (Sanabria et al. 2004; Vroomen and de Gelder 2004; Keetels et al. 2007; see reviews, Spence et al. 2007; Spence 2007). Intramodal grouping may segregate different sensory events from one another, thus weakening the crossmodal interaction. For example, using a sequence of sounds, Keetels et al. (2007) found the temporal ventriloquism effect to be diminished when the two target sounds (paired with visual targets) had the same frequency or rhythm as other flanker sounds.

Despite the empirical evidence reviewed above, the influence of sound structure on the temporal ventriloquism effect has thus far received only little attention in the literature. Arguably, however, to achieve an understanding of the mechanisms underlying the temporal ventriloquism effect, it is critical to examine crossmodal time perception—in particular, the perceived onset/offset time versus the perceived time interval of the events in the target and distractor modalities. Onset/offset time and time interval relate to the two fundamental concepts in time perception: succession and duration (Fraisse 1984). In most of the aforementioned studies, the temporal ventriloquism effects were implicitly assumed to be due to the visual events being captured by the accompanying auditory events at marked points (e.g., onsets or offsets) in time (temporal-marker hypothesis). To date, it is still unknown how the time interval of events in the distractor modality influences the temporal ventriloquism effect. If the onset or offset of a sound is the major factor determining the crossmodal interaction, a temporal ventriloquism effect should also be observable under single-sound conditions. Alternatively, as

one common dot at the center. When the spatial configuration is fixed, observers typically report two distinct percepts dependent on the inter-stimulus onset interval (ISOI): ‘element motion’ and ‘group motion’ (Fig. 1). Short ISOIs usually give rise to the percept of element motion, that is: the outer dots are perceived as moving, while the center dot appears to remain static or flashing. By contrast, long ISOIs give rise to the perception of group motion: the two dots are perceived to move together as a group (Kramer and Yantis 1997; Pantle and Petersik 1980; Pantle and Picciano 1976). Numerous studies have shown that element and group motion are never perceived simultaneously, that is, the two types of percept are mutually exclusive (for a review, see Petersik and Rice 2006).

In Experiment 1, we combined dual sounds with the visual Ternus display, introducing three different audiovisual temporal arrangements. As temporal ventriloquism has been found in both temporal-order judgments (TOJ) and classical apparent-motion tasks, a similar interaction was expected to be observed with visual Ternus apparent motion. In Experiment 2, we examined the effects of single-sound audiovisual arrangements, with a sound presented in temporal proximity to either the first or the second visual frame. If sound onset captures the onset of visual stimuli, one would expect to obtain a similar ventriloquism effect in single-sound conditions as in dual-sound conditions. However, if other factors, such as inter-sound interval or auditory grouping, are important for producing a temporal modulation, the temporal ventriloquism effect with single sounds may be weak, if at all present. In Experiments 3 and 4, we further examined the importance of the interval for the audiovisual temporal interaction. If a given (physical) time interval is perceived differently in different modalities and the auditory interval influences the visual interval, one would expect a temporal interaction to occur even when the onsets of the visual and auditory events coincide in time. Additionally, the variances of the interval estimates for the auditory, the visual, and the combined audiovisual events were examined further to quantitatively describe the interactions between the auditory and visual intervals.

Experiment 1

Two auditory clicks presented close in time to the visual events have been found to influence the sensitivity of (visual) temporal-order judgments (Morein-Zamir et al. 2003; Scheier et al. 1999) as well as the (visual) percept of apparent motion (Freeman and Driver 2008; Getzmann 2007). In Experiment 1, we examined whether a similar audiovisual temporal capture effect would also be found in the visual Ternus paradigm. In particular, we were interested in any temporal capture effect using an audiovisual interval of 30 ms, which is generally within the range of the audiovisual simultaneity window(ity)--

were controlled via the monitor's vertical synchronization pulse.

Design and procedure

Prior to the experiment, participants were shown demos of 'element motion' and 'group motion' and then performed a block of trials for practice. A trial started with a fixation cross presented at display center for 300 ms. Next, a blank display was shown for a random duration of 500–700 ms. This was then followed by the first visual stimulus frame which was presented for 30 ms. After a variable ISOI (80, 110, 140, 170, 200, 230, 260, 290, or 320 ms), the second visual stimulus frame was presented, also for 30 ms. The two visual frames shared one common element, located at the center of the display. The location of the other, outer element of the first frame, either to the left or to the right of the shared element, was always opposite to the outer element of the second frame (Fig. 2a). Each visual frame was accompanied by one brief, 30-ms sound. There were three different conditions of audiovisual interval (Fig. 2b): In condition 1, the first sound preceded the first visual frame and the second sound trailed the second visual frame by 30 ms (hereafter, we will refer to these as 'outer sounds'). In condition 2, the two sounds were presented simultaneously with two visual frames ('simultaneous sounds'). And in condition 3, the first sound trailed the first visual frame and the second sound preceded the second visual frame by 30 ms ('inner sounds'). On each trial, stimuli of the same audiovisual configuration were repeated once more after a 1,000-ms blank interval, so as to provide participants with a strong-enough impression for a response decision. The display then disappeared, and after a random duration of 300–500 ms, participants were presented with a question mark which prompted them to make a two-forced choice (mouse button) response indicating whether they had perceived element or group motion. For half of the participants, a left mouse click corresponded to 'group motion' and a right click to 'element motion,' and vice versa for the other half. Thus, Experiment 1 implemented a 3 (audiovisual intervals) × 9 (ISOIs) within-subjects design. Each configuration was presented 24 times, with the directions of apparent motion balanced across the 24 trial instances.

Results and discussion

The proportion of 'group motion' reports was plotted as a function of the ISOI and fitted by logistic regression for each participant. For each condition, the transition threshold between 'element motion' and 'group motion'—that is, the point at which 'group motion' and 'element motion' were reported equally frequently, which is also referred to as the point of subjective equality (PSE)—was calculated by estimating the 50% performance point on the (fitted) logistic function (Treutwein and Strasburger 1999).

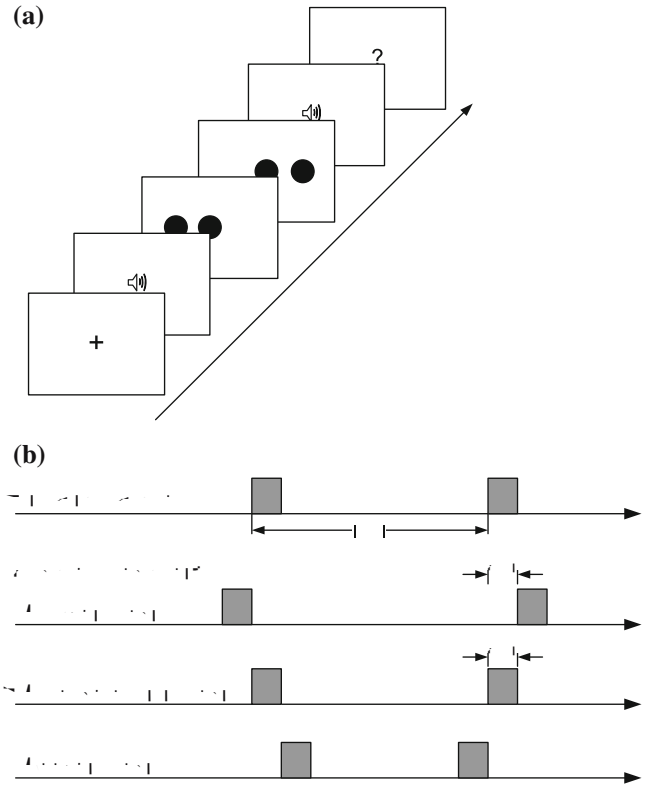

Fig. 2 **a** Schematic illustration of the events presented on one trial. The example illustrates the condition of 'outer sounds,' with the first sound presented before the first visual frame and the second sound after the second visual frame. **b** Illustration of the temporal succession of events for the three different conditions. In the 'outer-sounds' condition, the first sound was presented 30 ms before the onset of the first visual frame and the second sound 30 ms after the onset of the second visual frame. In the 'synchronous-sounds' condition, two sounds were presented simultaneously with the visual frames. In the 'inner-sounds' condition, the first sound was presented 30 ms after the onset of the first visual frame and the second sound 30 ms before the onset of the second visual frame

Figure 3 presents the results of Experiment 1. The mean PSEs are given in Table 1. A repeated-measures analysis of variance (ANOVA) of the estimated PSEs, with audiovisual interval (outer sounds, synchronous sounds, inner sounds) as factor, revealed the main effect to be significant, $F(2, 18) = 91.2$, $P < 0.001$—that is, the temporal configuration of two sounds influenced the visual Ternus effect (see Fig. 3b). Separate paired *t*-tests (with Bonferroni correction) confirmed the PSE to be significantly smaller for outer sounds compared to synchronous sounds, $P < 0.01$, and to be significantly larger for inner sounds compared to synchronous sounds, $P < 0.01$. In other words, participants were more likely to make a 'group motion' judgment with outer sounds compared to

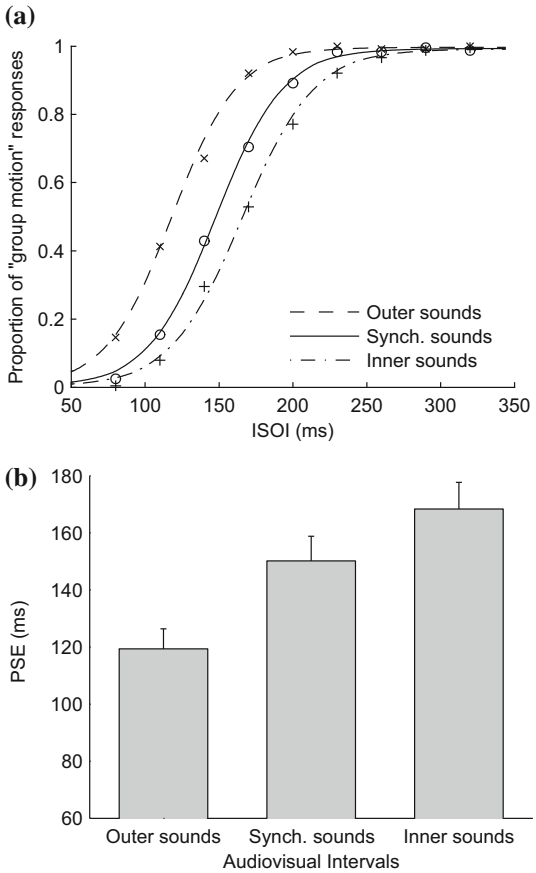

Fig. 3 **a** Psychometric curves fitted for the data from all participants in Experiment 1. The *solid curve* and *circles* represent the 'synchronous-sounds' condition, the *dashed curve* and crosses the 'outer-sounds' condition, and the *dash-dotted curve* and pluses the 'inner-sounds' condition. **b** Mean PSEs (and associated standard errors) for three conditions of audiovisual intervals

synchronous sounds and less likely with inner sounds compared to synchronous sounds. Both (outer and inner sounds) conditions produced the classical temporal ventriloquism effect (TVE) with a Ternus display: compared against the synchronous-sound baseline, the effect sizes were −30.8 (SE: 3.53) ms and 18.1 (SE: 3.05) ms for outer and inner sounds, respectively.

To our knowledge, with the Ternus display, this is the first demonstration of opposite TVEs on apparent motion by outer and inner sounds, respectively. In this regard, the present results go beyond previous studies (Getzmann 2007; Morein-Zamir et al. 2003; Scheier et al. 1999). For example, Getzmann (2007) demonstrated the influence of inner clicks on a continuous apparent-motion percept, but failed to find a significant effect of outer clicks (though the latter showed some tendency toward an effect). In another study (Morein-Zamir et al. 2003), there was also one intervening-sounds condition (with a 40-ms inter-stimulus-interval between sounds) that yielded no effect in a TOJ task. The failure to find opposite effects in previous studies might be due to the short ISOIs used in TOJ tasks and participants' difficulties in classifying the multiple percepts in classical apparent-motion tasks (see above).

Possible accounts for the temporal interaction effect have been put forward in previous studies. One typical explanation is based on the modality precision hypothesis. When two sensory modalities provide discrepant temporal information about an event, this discrepancy is resolved by favoring the modality that is characterized by a higher precision in registering that event. The auditory system has higher temporal resolution than the visual system (Welch 1999; Welch and Warren 1980). Accordingly, auditory events are assigned high weights, and visual events low weights, in audiovisual integration (Morein-Zamir et al. 2003; Welch and Warren 1980, 1986). However, what remains unclear is whether the differential modality weighting is based on points in time (onsets/offsets) or time intervals. Experiment 2 was designed to produce evidence for deciding between these alternatives by examining how single sounds, which do not provide any auditory-interval information, would influence the visual Ternus apparent motion.

Experiment 2

In a previous study (Morein-Zamir et al. 2003), an influence of dual sounds on visual TOJs was observed only in conditions in which the second sound trailed the second visual event by 100 or 200 ms, while there was no temporal capture effect in a single-sound condition. In Experiment 2, we examined for any crossmodal temporal interaction effect of single sounds, which were temporally manipulated (i.e., preceding, synchronous, or trailing) with respect to either the first or the second visual event in visual Ternus display.

Method

The method was the same as in Experiment 1, with the following exceptions.

Participants

The same ten participants who had taken part in Experiment 1 also participated in Experiment 2. Five participants performed Experiment 2 on day 1 and Experiment 1 on day 2, and vice versa for the other five participants—thus counterbalancing potential practice effects across the two experiments.

Design and procedure

Experiment 2 consisted of two separate sessions, hitherto referred to as Experiments 2a and 2b, respectively. Half the participants performed Experiment 2a first and then Experiment 2b, and vice versa for the other half. In Experiment 2a, only one sound was presented, namely, close to the first visual frame. In Experiment 2b, only the second sound was presented. The settings were the same as in Experiment 1, except that either the second sound (Experiment 2a) or the first sound (Experiment 2b) was omitted. The three conditions of audiovisual interval were preceding sound (30 ms before the onset of the respective visual frame), synchronous sound, and trailing sound (30 ms after the onset of the respective visual frame).

Results and discussion

Individual PSEs for the three audiovisual intervals were computed as in Experiment 1. For Experiment 2a, in which the sound accompanied the first visual frame, the mean PSEs are presented in Table 1 (see also Fig. 4a). A repeated-measures ANOVA of the PSEs showed the main effect of audiovisual interval to be significant, $F(2,18) = 3.69$, $P < 0.05$. Separate paired *t*-tests (with the conservative Bonferroni correction) of the PSEs revealed only a (marginally) significant difference between the preceding-sound and trailing-sound conditions (difference of 7.3 ms, $P = 0.09$). However, the classical TVEs calculated relative to the synchronous-sound baseline were far from reaching significance (although they were in the right direction numerically): −3.7 ms, $P = 0.42$, and 3.7 ms, $P = 0.69$, for the preceding-sound and trailing-sound conditions, respectively.

Similar results were obtained in Experiment 2b, where the sound accompanied the second visual frame—see Table 1 for the mean PSEs (see also Fig. 4b). A repeated-measures ANOVA revealed the main effect of audiovisual interval to be significant, $F(2,18) = 5.49$, $P < 0.05$. Follow-on paired *t*-tests (with Bonferroni correction) showed the PSE for the trailing-sound condition to be significantly smaller compared to both the synchronous sound (classical TVE of −5.9 ms, $P < 0.05$) and the preceding sound (difference of −9.4 ms, $P < 0.05$) condition, while the PSEs did not differ reliably between the preceding-sound and synchronous-sound conditions (3.5 ms, $P = 0.9$).

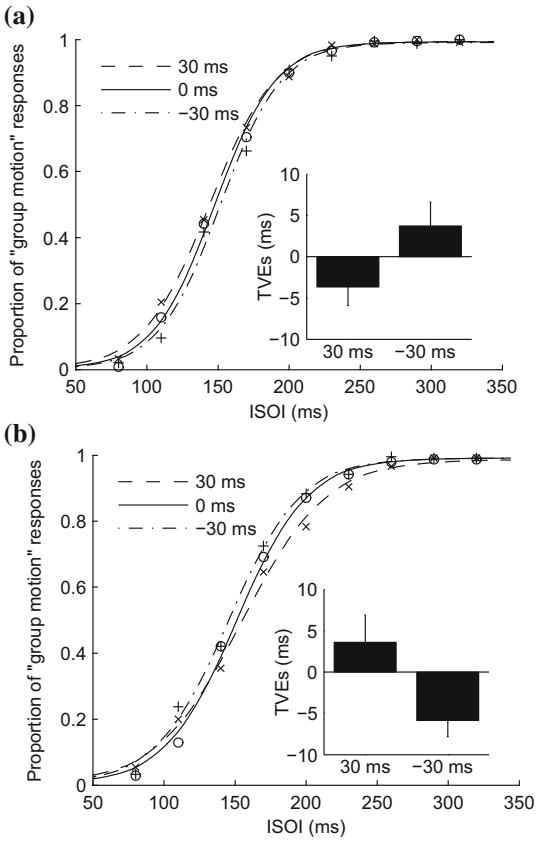

Fig. 4 **a** Psychometric curves fitted for the data from all participants in Experiment 2a. The *solid curve* and *circles* represent the 'synchronous-sound' condition (audiovisual interval 0 ms), the *dashed curve* and crosses the 'preceding-sound' condition (audiovisual interval 30 ms), and the *dash-dotted curve* and pluses the 'trailing-sound' condition (audiovisual interval −30 ms). The magnitude of the TVEs, calculated against the 'synchronous-sound' baseline, is presented in a subplot for the 'preceding-sound' (30 ms) and 'trailing-sound' conditions (−30 ms). **b** Psychometric curves fitted for the data from all participants in Experiment 2b. The magnitude of the TVEs, calculated against the 'synchronous-sound' baseline, is presented in a subplot for the 'preceding-sound' (30 ms) and 'trailing-sound' conditions (−30 ms)

The fact that there was no significant TVE with a sound preceding or trailing the first visual frame or a sound preceding the second visual frame in Experiment 2 is consistent with Bruns and Getzmann (2008), who reported a similar pattern with temporally proximal single sounds. However, a small TVE was found for trailing sound associated with the second visual frame in Experiment 2b. The asymmetry between preceding and trailing sounds is similar to the results obtained in a previous TOJ study with a dual-sound modulation (Morein-Zamir et al. 2003). The underlying mechanism is still unclear, though it has been suggested that this pattern relates to an asymmetry in audiovisual simultaneity (Dixon and Spitz 1980; Morein-Zamir et al. 2003).

In any case, compared to the TVEs obtained with the two-sound configurations in Experiment 1, TVEs by single sounds are relatively weak and significant only with a sound trailing the second visual frame. Even comparing the transition thresholds between the preceding- and trailing-sound conditions, the differences are merely 7.3 and 9.4 ms for Experiments 2a and 2b, respectively. Note that, compared with Experiment 1, all settings were the same except that one sound was removed. In addition, in Experiment 2, the additive (classical TVE) effect of the sound preceding the first visual frame plus that trailing the second visual frame was only −9.6 ms, which is much smaller than the TVE (of −30.8 ms) in the outer-sounds condition of Experiment 1. Similarly, the additive effect of two individual 'inner sounds' was 7.2 ms, which is also smaller than the TVE (of 18.1 ms) in the inner-sounds condition of Experiment 1. This suggests that the effects obtained with single sounds in Experiment 2 cannot fully explain the temporal ventriloquism effect, based only on the influence of the sounds' onsets (i.e., temporal markers). Previous attempts to account for the weak or absent TVE with single-sound configurations have been in terms of a violation of the fundamental 'assumption of unity' (Morein-Zamir et al. 2003; Welch 1999). On this assumption, two separate multisensory events which share physical properties (e.g., both involving a visual component associated with an auditory component) are more likely considered as originating from the same source. Alternatively, if two sensory modalities are experienced by the observers as signaling disparate events (e.g., if one of the two events involves only one modality), the crossmodal interaction is less likely to happen (Welch 1999). Consequently, in single-sound conditions, two events with different physical properties (one involving audiovisual and the other visual stimuli only) weaken the attribution of a common cause and, therefore, the crossmodal temporal interaction. However, besides a violation of the 'assumption of unity,' an alternative account for the weak TVE in Experiment 2 could be the lack of an inter-sound interval (with single-sound configurations).

Experiment 3

One possible way to disentangle the role of temporal markers from that of time interval without violating the 'assumption of unity' is to keep the onsets of the audiovisual events physically the same, while making the perceived intervals different. To realize such a situation, in Experiment 3, synchronous audiovisual events (with sound onset occurring simultaneously with visual frame onset) were used. The rational for this is as follows. There is ample evidence from studies of audiovisual time judgments that the interval between two auditory stimuli appears longer than that between two visual stimuli, even if the auditory and visual stimuli are presented simultaneously (Goldstone and Lhamon 1974; Walker and Scott 1981; Wearden et al. 1998). If the inter-sound interval plays a (critical) role in audiovisual temporal interactions, one would expect that the transition threshold (from element to group motion) in visual Ternus displays would be lowered by the inter-sound interval in the synchronous audiovisual condition, compared with the unimodal (visual-display-only) condition. By contrast, the hypothesis of temporal-marker capture would predict that the transition threshold would remain the same for both conditions, as the auditory events are presented simultaneously with the visual events.

Method

The method was the same as in Experiment 2, with the following exceptions.

Participants

Ten participants (5 females, mean age of 26.5) took part in Experiment 3, all with normal or corrected-to-normal vision and normal hearing. Participants were naïve as to the purpose of the experiment. Informed consent was obtained before the start of testing. Participants were paid 8 Euros for their service.

Design and procedure

The settings were the same as in Experiment 2, except for the following differences. As most participants had made nearly 100% group motion judgments for long ISOIs in the previous experiments, the range of ISOIs was adjusted to 80–260 ms, with increasing step sizes of 30 ms. There were two experimental conditions: the audiovisual synchronous condition, in which the two sounds were presented synchronously with the two visual frame; and the unimodal condition, in which the visual stimuli were presented without sounds. In order to avoid inter-trial effects due to the presentation of sounds, the two conditions were presented blockwise, with block order counterbalanced across participants.

Results and discussion

Individual PSEs for two conditions were computed as in Experiment 1. The mean PSEs (and associated standard errors) were 164 (±7.46) ms and 153 (±6.78) ms for the unimodal (visual-only) and the audiovisual synchronous condition, respectively (see Fig. 5). A repeated-measures ANOVA of the TVEs revealed the main effect of condition to be significant, $F(1,9) = 10.67$, $P < 0.01$. That is,

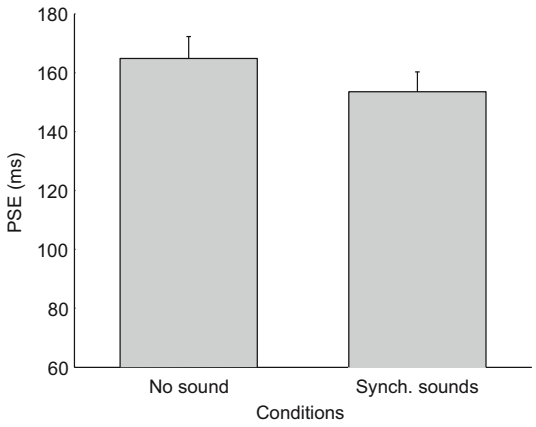

Fig. 5 Mean PSEs (and associated standard errors) for the visual-only (i.e., no sound) condition and the audiovisual synchronous condition

perceptual reports were biased toward group motion in the audiovisual synchronous condition.

This finding cannot be explained by auditory temporal-marker capture, since the temporal onsets of the auditory and the visual stimuli were the same. One potentially critical factor for this finding in the audiovisual synchronous condition was the subjective auditory interval, which may have influenced the subjective visual interval. Studies of time perception have shown that auditory intervals or durations in the range of seconds are perceived as longer than the same intervals or durations in the visual modality (Goldstone and Lhamon 1974; Walker and Scott 1981). This may also be true for sub-second interval perception. Furthermore, auditory intervals may be assigned greater weights, so that they dominate the perception of audiovisual intervals (Goldstone and Goldfarb 1964a, b; Goldstone and Lhamon 1974; Wearden et al. 1998). To corroborate this hypothesis, Experiment 4 was designed to permit a direct comparison of the subjective visual intervals, auditory intervals, and audiovisual intervals in the visual Ternus paradigm.

Experiment 4

In Experiment 4, we estimated and compared the subjective intervals (near the thresholds obtained in the preceding experiments) between two Ternus display frames with and without synchronous auditory stimuli (Experiment 4a). In addition, we examined the variances of the interval estimates for the auditory, the visual, and the audiovisual intervals, so as to be able to quantitatively examine the influence of the auditory interval on the perception of the audiovisual interval (Experiment 4b). If audiovisual temporal integration, like the spatial integration, operates by optimally integrating the independent interval estimates from each modality, the audiovisual interval would be predictable by Bayesian maximum-likelihood estimation (MLE) (Alais and Burr 2004; Battaglia et al. 2003; Burr et al. 2009; Ernst and Banks 2002). On the MLE model, the 'optimal' weighting scheme for crossmodal integration is based on the variability of the unimodal estimates:

$$\widehat{I}_{av} \quad {}_a\widehat{I}_a \quad {}_v\widehat{I}_v, \qquad 1$$

where $\widehat{I}_a$, $\widehat{I}_v$, and $\widehat{I}_{av}$ are the perceived auditory, visual, and audiovisual intervals; ${}_a$ and ${}_v$ are the weights of unimodal auditory and visual intervals, respectively. The weights are inversely proportional to the variances ${}^2_a$ and ${}^2_v$ of the auditory and visual intervals:

$${}_a \quad \frac{1/\ {}^2_a}{1/\ {}^2_a \quad 1/\ {}^2_v} \quad \text{and} \quad {}_v \quad 1 \quad {}_a \qquad 2$$

The estimate of $\widehat{I}_{av}$ is optimal since it has the greatest reliability (Alais and Burr 2004; Burr et al. 2009; Ernst and Banks 2002).

Method

The method was the same as in Experiment 1, with the following exceptions:

Participants

Ten participants took part in Experiment 4a (5 females, mean age of 25.4) and ten in Experiment 4b (6 females, mean age of 25.9), all with normal or corrected-to-normal vision and normal hearing. Participants were naïve as to the purpose of the study. Payment for participating was 8 Euros.

Design and procedure

In Experiment 4a, the visual Ternus display was adopted for visual interval presentation. Prior to the experiment, participants performed one block of trials for practice. A trial started with a fixation cross presented at the center of the display for 300–500 ms. After that, a blank display was shown for a random duration of 300–500 ms. This was then followed by the successive presentation of two types of interval (with a random gap of 1–1.2 s between the presentations): (1) One was the standard interval in between the two successive visual frames (presented without auditory stimuli). The inter-stimulus interval (ISI) was fixed at 130 ms, and both frames were presented for 30 ms. (2) The other was the comparison interval which was created by two 30-ms 'frames' of audiovisual synchronous stimuli. The ISI between the frames was

selected at random from the set [40, 70, 100, 130, 160, 190, 220] ms. The order of the two interval presentations (standard, audiovisual) was determined randomly and counterbalanced across trials. Following the presentation of both intervals, participants were prompted to make a judgment of which of the two intervals was longer. Participants pressed the left or right buttons to indicate the first or, respectively, the second interval was the longer one. Overall, the (within-subject design) experiment consisted of 336 trials: comparison interval (7 levels) × 48 trials with counterbalanced order of interval presentation (2 levels) and direction of (apparent) motion (2 levels).

In Experiment 4b, the variances associated with three different interval estimates were examined: (1) visual interval (V), realized in terms of the classical visual Ternus display (without sounds; see Experiment 4a above); (2) auditory interval (A), created by two 500-Hz tones of 30-ms duration; (3) audiovisual interval (AV), formed by the audiovisual Ternus display (see Experiment 4a above). Unlike Experiment 4a, for each condition, the same type of interval (i.e., visual or auditory or audiovisual) was used for both the standard stimulus and the comparison stimulus. Similar to Experiment 4a, in the comparison stimulus, a random ISI of 40, 70, 100, 130, 160, 190, or 220 ms was inserted between two visual frames (V condition), two sounds (A condition), and two audiovisual synchronous stimuli (AV condition), respectively. The standard stimuli for the auditory and visual conditions involved the same interval, of 130 ms. In the audiovisual condition, there were three standard audiovisual stimuli which had a similar temporal structure to those illustrated in Fig. 2b for Experiment 1: two brief sounds flanking (outer sounds, inner sounds) or synchronizing with the two visual frames. The ISI between visual frames was fixed to 130 ms, but the auditory intervals were 170, 130, and 90 ms for the outer-sound, synchronous-sound, and inner-sound conditions, respectively. These conditions will hitherto be abbreviated as $AV_{+40,}$ $AV_{0,}$ AV_{-40} (where the number subscript gives the extension/contraction of the interval time relative to the standard interval of 130 ms). In all other respects, the procedure was the same as in Experiment 4a. Experiment 4b implemented a within-subject design: the five types of interval (i.e., A, V, $AV_{+40,}$ $AV_{0,}$ AV_{-40}) were presented in separate trial blocks (with randomized block order), with 7 levels of comparison interval presented randomly within blocks, with 24 repetitions for each condition.

Results and discussion

In Experiment 4a, the PSEs were calculated individually from the estimated psychometric curves. The mean PSE (±SE) was 103 (±10.9) ms, which means that an interval of 103 ms in the audiovisual synchronous condition was perceived as equivalent to the visual interval of 130 ms. A *t*-test comparing the mean PSE to 130 ms revealed the difference to be significant, $t(9) = -2.59$, $P < 0.05$. Thus, with audiovisual synchronous presentation, the interval was perceived as longer than the same physical visual-only interval, due to the auditory interval expanding the (synchronous) visual interval. This result is consistent with previous comparisons of audiovisual intervals (Goldstone and Lhamon 1974; Walker and Scott 1981; Wearden et al. 1998).

For Experiment 4b, a psychometric function based on the cumulative normal distribution was fitted for each interval condition (i.e., A, V, $AV_{+40,}$ $AV_{0,}$ AV_{-40}) within each participant, and the PSEs and the standard deviations of the distribution ( ) were calculated from the thresholds of 50% and 84.1% (see details in Alais and Burr 2004; Battaglia et al. 2003; Ernst and Banks 2002; Witten and Knudsen 2005). The overall psychometric curves are presented in Fig. 6. The mean PSEs and mean standard deviations of the distributions for the five different intervals are listed in Table 2.

Separate t-tests comparing the PSEs relative to standard interval of 130 ms revealed the PSEs to differ significantly from 130 ms only in the AV_{-40} and the AV_{+40} conditions, $t(9) = -20.67$, $P < 0.01$, and $t(9) = 16.72$, $P < 0.01$, respectively. This means that the audiovisual subjective interval was influenced by the auditory interval. To further quantify this interaction, the weights associated with the auditory interval (modality) were calculated according to Equation (2), and the PSEs predicted by the MLE model were calculated according to Equation (1). These values

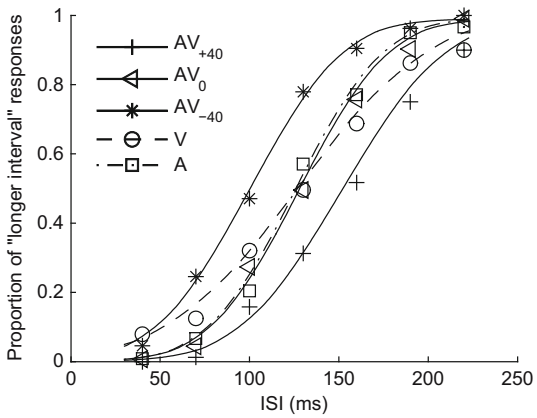

Fig. 6 Psychometric curves fitted for the data from all participants in Experiment 4b. The *solid curves* with *stars*, *triangles*, and *pluses* represent the AV_{-40}, AV_0, and AV_{+40} audiovisual interval conditions, respectively; the *dashed curve* and *circles* the visual interval condition (V); and the *dash-dotted curve* and *squares* the auditory-interval condition (A)

are listed in Table 2. The mean weight of the auditory interval ($_a$) was 0.717, which was substantially higher than that for the visual interval ($_v = 0.283$). More interestingly, the predicted PSEs were very close to the PSEs estimated from the empirical data.

The results of Experiment 4b go beyond the merely qualitative modality precision hypothesis: using the Bayesian model, the subjective audiovisual interval could be quantitatively predicted from the weighted integration of the auditory and visual intervals. Taken together with Experiment 3, the results provide further support for the hypothesis that synchronous auditory stimuli expand the (estimated) interval between visual stimuli. Consequently, one can conclude that the audiovisual synchronous presentation shifted the visual apparent-motion percept (in Experiment 3) toward group motion, thus decreasing the

by providing the same paired (audiovisual) features in all events (fulfilling the 'assumption of unity'), crossmodal temporal integration becomes possible (Getzmann 2007; Morein-Zamir et al. 2003; Welch 1999). However, besides the imbalance in features, the auditory interval is also missing in single-sound conditions. Experiments 3 and 4 revealed that it is the audiovisual time interval, rather than the temporal event markers, that critically determines the audiovisual apparent motion. Since the onsets of the auditory and visual stimuli were the same in the audiovisual synchronous condition, the temporal-marker hypothesis and the notion of audiovisual pairing alone are not sufficient to explain the threshold shift of the apparent motion. Studies of crossmodal time perception have shown that perceived inter-stimulus intervals are not equal in the various modalities. For example, auditory intervals/durations are typically perceived as longer than visual intervals/durations with intervals/durations in the range of seconds (Goldstone and Lhamon 1974; Walker and Scott 1981). A recent study of vibro-tactile and visual asynchronies reported that empty tactile intervals were also perceived as longer than visual intervals (van Erp and Werkhoven 2004). To explain the asymmetric interval perception among sensory modalities, it has been proposed that the respective internal pacemakers run at different speeds (Wearden 2006): the internal clock runs faster in modalities with higher temporal precision, so that there are more 'clock ticks' accumulating for stimuli defined in these modalities compared to others (with the number of clicks determining the lengths of perceived intervals). Experiment 4a revealed a similar result, namely, that the auditory interval is perceived as longer than the same visual interval at the sub-second level. Experiment 4b further suggested that the perceived audiovisual interval is integrated from the optimal weights of the auditory and visual intervals. Quantitative probability-based models have previously been found to provide a good account of crossmodal spatial integration (Alais and Burr 2004; Battaglia et al. 2003; Ernst and Bülthoff 2004; Ernst and Banks 2002). Yet it is less clear whether these models do also apply to crossmodal temporal integration. Requiring participants to make temporal-order judgments (TOJ), Ley et al. (2009) found performance in a bimodal TOJ task, with combined auditory-tactile stimuli, to be well explained by an optimal MLE model. By contrast, using a different, temporal-bisection task, Burr et al. (2009) found that the model of optimal combination fit only roughly with their pattern of results. In contrast to their study, using an apparent-motion paradigm, we provide new evidence that humans may be able to estimate crossmodal intervals in an optimal manner, though the variances of the estimates were not predicted accurately by the Bayesian model. The discrepancies among the relevant studies might arise from the different paradigms used. However, it is also possible that prerequisites of Bayesian models, such as assumption of Gaussian noise, are not fully satisfied in crossmodal temporal judgments, as a result of which the predictions of Bayesian models are not quite accurate (Burr et al. 2009).

Besides the absence of an auditory interval, intramodal grouping was also absent in single-sound conditions (Experiment 2). A number of studies have shown that the

Acknowledgments This research was supported in part by a grant from the German Research Foundation (DFG) within the Collaborative Research Centre SFB 453. We thank Hans Strasburger for stimulating comments and the reviewers for insightful suggestions.

References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14(3):257–262

Allen PG, Kolers PA (1981) Sensory specificity of apparent motion. J Exp Psychol Hum Percept Perform 7(6):1318–1328

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20(7):1391–1397

Bermant RI, Welch RB (1976) Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. Percept Mot Skills 42(43):487–493

Bertelson P, Radeau M (1981) Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Percept Psychophys 29(6):578–584

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10(4):433–436

Bruns P, Getzmann S (2008) Audiovisual influences on the perception of visual apparent motion: exploring the effect of a single sound. Acta Psychol 129(2):273–283

Burr D, Banks MS, Morrone MC (2009) Auditory dominance over vision in the perception of interval duration. Exp Brain Res 198(1):49–57

Calvert G, Spence C, Stein BE (2004) The handbook of multisensory processes. MIT Press, Cambridge

Dixon NF, Spitz L (1980) The detection of auditory visual desynchrony. Perception 9(6):719–721

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415(6870):429–433

Ernst MO, Bülthoff HH (2004) Merging the senses into a robust percept. Trends Cogn Sci 8(4):162–169

Fendrich R, Corballis PM (2001) The temporal cross-capture of audition and vision. Percept Psychophy 63(4):719–725

Fraisse P (1984) Perception and estimation of time. Annu Rev Psychol 35:1–36

Freeman E, Driver J (2008) Direction of visual apparent motion driven by timing of a static sound. Curr Biol 18:1262–1266

Getzmann S (2007) The effect of brief auditory stimuli on visual apparent motion. Perception 36(7):1089–1103

Goldstone S, Goldfarb JL (1964a) Auditory and visual time judgment. J Gen Psychol 70:369–387

Goldstone S, Goldfarb JL (1964b) Direct comparison of auditory and visual durations. J Exp Psychol 67:483–485

Goldstone S, Lhamon WT (1974) Studies of auditory-visual differences in human time judgment. 1. Sounds are judged longer than lights. Percept Mot Skills 39(1):63–82

Harrar V, Harris LR (2007) Multimodal Ternus: visual, tactile, and multisensory grouping in apparent motion. Perception 36:1455–1464

Howard IP, Templeton WB (1966) Human spatial orientation. Wiley, New York

Jaekl PM, Harris LR (2007) Auditory-visual temporal integration measured by shifts in perceived temporal location. Neurosci Lett 417(3):219–224

Keetels M, Stekelenburg J, Vroomen J (2007) Auditory grouping

Wearden JH (2006) When do auditory/visual differences in duration judgments occur? Q J Exp Psychol 59(10):1709–1724

Wearden JH, Edwards H, Fakhri M, Percival A (1998) Why "sounds are judged longer than lights": application of a model of the internal clock in humans. Q J Exp Psychol 51(2):97–120

Welch RB (1999) Meaning, attention, and the 'unity assumption' in the intersensory bias of spatial and temporal perceptions. In: Aschersleben G, Bachmann T, Müsseler J (eds) Cognitive contribution to the perception of spatial and temporal events. Elsevier, Amsterdam, pp 317–387

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88(3):638–667

Welch RB, Warren DH (1986) Intersensory interactions. In: Boff KR, Kaufman L, Thomas JP (eds) Handbook of perception and human performance: sensory processes and perception, vol 1. Wiley, New York, pp 1–36

Witten IB, Knudsen EI (2005) Why seeing is believing: merging auditory and visual worlds. Neuron 48(3):489–496