Perceptual synchrony of audiovisual streams for natural and artificial motion sequences

Perceptual synchrony of audiovisual streams for natural and artificial motion sequences

Roberto Arrighi
Istituto Nazionale di Ottica Applicata (INOA), Firenze, Largo E. Fermi 6, Italy

David Alais
Auditory Neuroscience Laboratory, Department of Physiology and Institute for Biomedical Research, School of Medical Science, University of Sydney, Sydney, Australia

David Burr
Istituto di Neuroscienze del CNR, Via Moruzzi 1, Pisa, Italy

We investigated the conditions necessary for perceptual simultaneity of visual and auditory stimuli under natural conditions: video sequences of conga drumming at various rhythms. Under most conditions, the auditory stream needs to be delayed for sight and sound to be perceived simultaneously. The size of delay for maximum perceived simultaneity varied inversely with drumming tempo, from about 100 ms at 1 Hz to 30 ms at 4 Hz. Random drumming motion produced similar results, with higher random tempos requiring less delay. Video sequences of disk stimuli moving along a motion profile matched to the drummer produced near-identical results. When the disks oscillated at constant speed rather than following "biological" speed variations, the delays necessary for perceptual synchrony were systematically less. The results are discussed in terms of real-world constraints for perceptual synchrony and possible neural mechanisms.

Keywords: audiovisual temporal alignments, biological motion

Introduction

The modular theory of perception proposes that different attributes of an object, such as its color, shape, motion, and so on, are processed by different independent areas in the brain. Although this idea, based on David Marr's (1982) "principles of modular design," has been recently challenged by some authors (Burr, 1999; Lennie, 1998), it has also received a good deal of support from anatomical and physiological studies suggesting that distinct areas of the brain are involved in the analysis of different stimulus attributes (Livingstone & Hubel, 1988; Zeki, 1993). The results of these modular processes must then be made available to a higher level "binding mechanism" that groups the perceptual elements into a coherent global percept.

The binding mechanism must piece together perceptual elements that belong to a particular object to produce a coherent conscious representation of that object. This picture of the perceptual process as a framework of independent and parallel channels of analysis raises an interesting issue in the temporal domain. Simultaneous changes of different attributes of a stimulus could be perceived to occur at different times if processing times of the separate modules were to differ significantly. To investigate this issue, many studies have been conducted to examine whether simultaneous changes in stimulus attributes such as form, color, or movement produce synchronous perceived changes (Arnold, Clifford, & Wenderoth, 2001; Aymoz & Viviani, 2004; Moutoussis & Zeki, 1997; Nishida & Johnston, 2002; Viviani & Aymoz, 2001). Most of these studies suggested that perceptual asynchronies could occur despite simultaneous stimulus changes (Arnold et al., 2001; Aymoz & Viviani, 2004; Moutoussis & Zeki, 1997; but see also Nishida & Johnston, 2002; Viviani & Aymoz, 2001).

Despite some disagreements about the quantitative differences of the relative latencies between these stimulus attributes, most of the data show that the perception of motion is slower than both color and form perceptions. This is a surprising result, given the common belief that the visual system processes motion information very quickly (Livingstone & Hubel, 1987) through rapid myelinated pathways to areas specialized for motion processing (Albright, 1984; Britten, Shadlen, Newsome, & Movshon, 1992). On the other hand, there is one recent suggestion that motion perception is slower only for artificial stimuli and not for natural human movements. In an elegant series of experiments, Aymoz and Viviani (2004) demonstrated a delay in perceiving motion changes relative to changes in color and form, but they also showed that this delay vanished if the motion of stimulus was produced by the action of a human agent. This claim is consistent with a large literature showing that biological motion has particular properties that distinguishes it from other forms of motion (Johansson, 1973, 1976, 1977; Nakayama, 1985).

In this paper, we investigate whether biological motion can also affect temporal perception of visual and auditory cross-modal stimuli. We recorded movies of a professional drummer playing conga drums at various tempos, from

doi: 10.1167/6.3.6 Received June 6, 2005; published March 16, 2006 ISSN 1534-7362

which we created three different classes of stimuli. The first class was the "natural" stimuli, which consist of movies of the drummer playing the conga drums. In the second class of stimulus, two red disks replaced the drummer's hands and moved with motion profiles matching the tip of the drummer's middle finger, hitting a bar representing the congas. In the last class of stimulus, the movement of the two red disks was altered to follow a triangular function with the same period as the biological hand motion. In all conditions, we varied the visual–auditory synchrony and asked participants to judge whether the visual and auditory streams appeared to be synchronous.

Methods

Participants

Two of the authors (R.A. and D.A.) and one naive female participant (mean age 30 years), all with normal hearing and normal or corrected visual acuity, served as participants. All gave informed consent to participate in the study that was conducted in accordance with the guidelines of the University of Sydney and the University of Florence. The tasks were performed in a dimly lit, sound-attenuated room.

Apparatus and stimuli

All visual stimuli were presented on an Asus LH 3500 laptop computer with an LCD screen resolution of 1024 × 768 pixels, 32-bit color depth, refresh rate of 60 Hz, and mean luminance of 17.2 cd/m^2. Auditory stimuli were digitized at a rate of 80 kHz and presented through two high-quality loudspeakers (Creative MMS 30) flanking the computer screen and lying in the same plane 60 cm from the participant. Speaker separation was 70 cm and stimuli intensity was 90 dB at the sound source.

Several movies of a professional drummer playing a conga drum were recorded using a Sony digital camera (model TRV33E, image resolution 1280 × 768, frame rate 30 Hz) in a rehearsal room at the Percussion Department of the Sydney Conservatorium of Music. The video camera was set close (27 cm) to the drum pointed directly at the drummer; hence, the conga drum was in the foreground with the drummer's torso and arms also visible behind it. We recorded the drummer playing a constant beat at three different tempos (1, 2, and 4 Hz). These were filmed with the camera located at three different elevations: 0, 30, and 60 deg. At 0 deg elevation, the drum skin was not visible (the head of the drum was seen edge-on). At 30 and 60 deg elevation, the drum skin was visible. For all elevations, the distance from the camera lens to the drum skin was constant, and the drummer's hands, while playing, never left the movie frame. The main difference between the elevation conditions was the precision with which the moment of contact with the drum skin could be determined. At 0 deg, this was easily determined, but it was progressively harder at 30 and 60 deg. The movies were imported to the computer where they were converted into sequences of single bitmap images using the software Virtualdub (http://www.virtualdub.org/) and imported into a Matlab script. This allowed us to present the sequences of bitmap images as movies with the Psychtoolbox routines (Brainard, 1997; Pelli, 1997), rather than a movie player that may introduce audio delays, change visual delays, or both. The display frame rate was 30 Hz, and the frame resolution was 320 × 240 pixels (9.4 × 7 deg) with a color depth of 24 bits. The auditory soundtrack from the original movies was also imported into a Matlab script and edited by using the appropriate Psychtoolbox routines to obtain full control about its synchronization with the movies' visual stream. Synchronization accuracy was checked with the use of a photocell and a microphone and was found to be within 2 ms. An example of one of these movies with the drummer playing at 1 Hz from a 0-deg viewpoint elevation can be seen in Movie 1.

In conditions where the visual component of the drumming sequences was not naturalistic, we simulated the drumming as follows. We drew a white rectangle that has the same size as the original movie frames and added a thin, black horizontal line (0.3 deg in height) across the frame in the lower portion of the rectangle (2.4 deg from the bottom of the frame) to represent the head of the conga drum. The tips of the drummer's middle fingers were simulated by two red circular discs 1 deg in diameter. For the condition where the drumming movement was "biological," these red dots oscillated in such a way that they traced the same path as the tips of the drummer's hands as recorded in the original movies. Figure 1 shows examples of these profiles (for the three temporal frequencies investigated) for both the horizontal and vertical components of the drummer's right hand, along with its instantaneous velocity (considering both motion components). In the condition in which the drumming movement was nonbiological, the dots oscillated in a triangular wave (constant speed and abrupt reversal of direction) interposed with static rest periods at frequencies of 1, 2, or 4 Hz. The auditory patterns in these simulated drumming conditions were sequences of simple clicks from square wave pulses (100 ms wide) played at the same temporal period of the original conga drum sound sequences. One hundred milliseconds was chosen to approximate the duration of the maximum energy period of the recorded conga drums.

Procedures

Each trial started with a midgray screen. The corners of a rectangle (320 × 240 pixels) were drawn with black

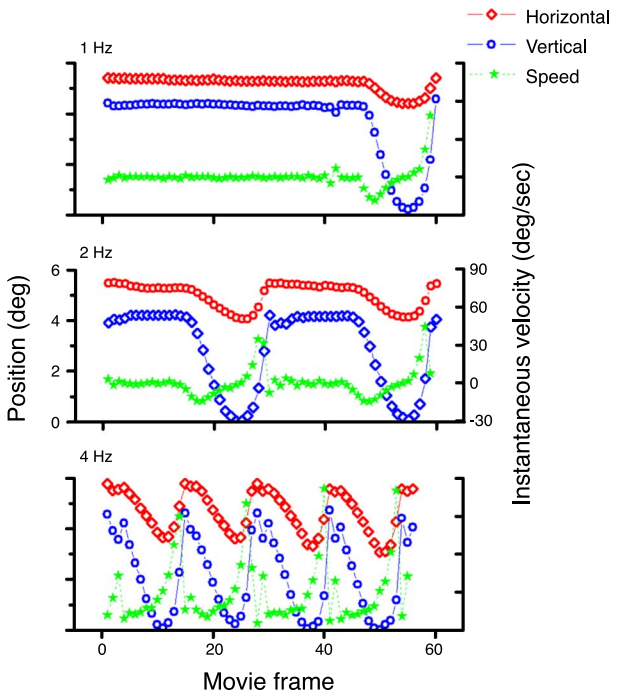

Figure 1. Motion profiles of the tip of the drummer's right middle finger taken from drumming sequences used in Experiment 1 (motion of only one hand is shown). The horizontal component of the motion is shown in red, and the vertical component is shown in blue. In each graph, the abscissa indicates the movie frame number (the frame rate was 30 Hz) and the left ordinate indicates the position of the tip of the drummer's finger relative to the top left corner of the movie frame. In Experiments 2 and 3, these profiles were applied to red disks in the condition where a biological drumming movement was carried by a nonhuman agent. The green symbols and lines indicate the instantaneous velocity of the drummer's hands in each movie frame (right ordinate).

lines to indicate the area as to where the movies will appear. To prevent participants from using the stimulus onset as a temporal reference point, there was a random delay of 0.5–1.5 s after each key press. The auditory and the visual streams of the movie were smoothly introduced by multiplying the visual contrast and audio intensity with a temporal ramp described by the rising profile of a Gaussian function (σ = 300 ms). The movie presentation began at a randomly chosen phase, lasted 3 s, and was followed by a Gaussian off-ramp. The relative phase of the soundtrack and visual sequences varied systematically. Participants were required to indicate whether the visual and the auditory streams of the movie were in phase or not and encouraged to strictly maintain a constant criterion. On each trial, the visual–auditory asynchrony randomly varied between + $\lambda/2$ and − $\lambda/2$, where λ is the period of the drumming rate being played by the drummer. For example, for a rhythm of 2 Hz, each cycle lasted 500 ms; thus, the value of $\lambda/2$ is 250 ms. However, in the 1-Hz condition, the asynchrony range was limited to $\pm\lambda/4$, as the task was already trivially easy with a quarter-cycle asynchrony. For each asynchrony, 25 trials were collected.

Results

Experiment 1: Temporal alignments of visual and auditory natural stimuli

In this experiment, we delayed the auditory stream of the movie by variable amounts and measured the frequency that observers judged the visual and auditory streams of a drummer to be synchronous. Sample data for three different drumming tempos (1, 2, and 4 Hz) at the 30-deg viewpoint elevation are shown in Figure 2.

The dashed lines show the best-fitting Gaussian curves to the data (all with r^2 values ranging from .83 to .98). The peaks of these Gaussians were taken as an estimate of the point of subjective simultaneity (PSS). The red, blue, and

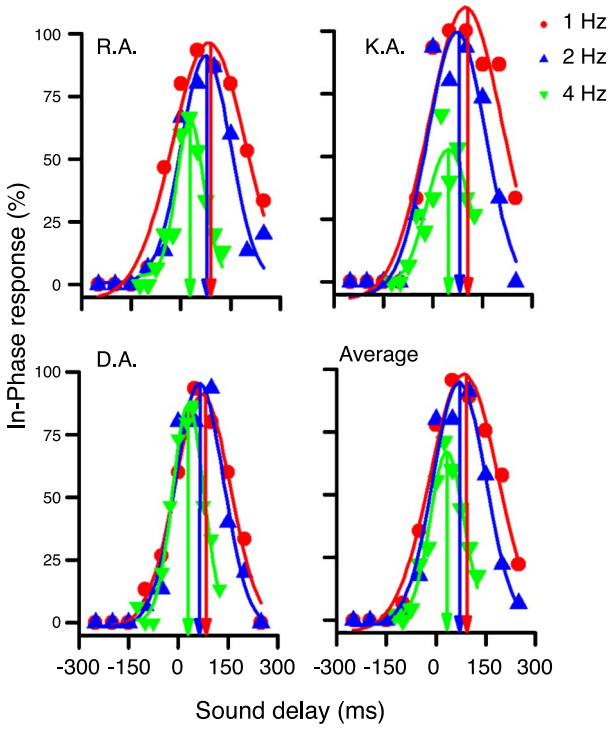

Figure 2. Data from Experiment 1 showing percentage of "in-phase" responses as a function of delay of the auditory stream (positive values indicate that the auditory stream was delayed relative to vision). Red, blue, and green symbols represent drum tempos of 1, 2, and 4 Hz, respectively. The dashed lines are the best-fitting Gaussian curves to the data (25 trials/data point). The peaks of these curves provide an estimate of the PSS. All data were collected with 30 deg viewpoint elevation.

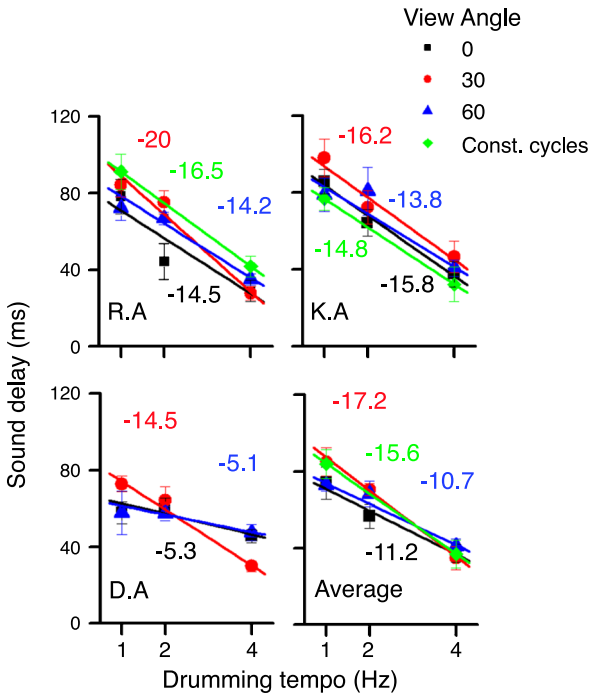

Figure 3. Delays to produce best synchrony, calculated from the best-fitting Gaussians to the data such as those shown in Figure 2, as a function of drum tempo. The dashed lines are the best linear fits to the data, with their slopes. The results for all viewpoint elevations were similar, all producing a negative dependency on tempo of 10–17 ms/Hz.

green curves indicate, respectively, drumming tempos of 1, 2, and 4 Hz. For all drumming tempos, best temporal alignment of the visual and auditory stimuli occurred when the movie soundtrack was delayed relative to the visual pattern. However, the size of the best auditory delay depended heavily on drumming tempo, it being larger for lower frequencies. The red curves in Figure 2 (1 Hz) are shifted furthest to the right, followed by blue (2 Hz) and green (4 Hz).

Figure 3 plots the PSS against drum tempo for the three different viewpoints. Viewpoint had very little effect. In all cases, perceived simultaneity for 1, 2, and 4 Hz tempo occurred when the auditory pattern was delayed by about 80, 60, and 40 ms, respectively. The main effect for frequency, on the other hand, was robust and monotonic. The trend is well described by a linear relationship between delay and frequency, with slopes of the best linear fits (dashed lines in Figure 3) ranging from −10 to −17 ms/Hz. A two-way ANOVA test (Factor 1: drumming tempo; Factor 2: viewpoint elevation) was performed to test the statistical significance of our observations. Indeed, it was confirmed that whereas the effect of the drumming tempo is clearly significant, $F(2,2) = 32.223$, $p < .001$, the viewpoint elevation did not produce any significant effect, $F(2,2) = 1.145$, $p = .340$, nor was there any statistically significant interaction between these two factors, $F(2,4) = 1.205$, $p = .343$. For participants R.A. and K.A., we also measured perceptual synchrony with six cycles of drumming in all conditions rather than varying the number of cycles with temporal frequency. The results are essentially the same, suggesting that the task does not depend heavily on the number of cycles.

Experiment 2: Temporal perception of artificial stimuli undergoing biological and nonbiological motion

In this experiment, we investigated the importance of the stimuli being natural and biological for perception of temporal alignment, repeating Experiment 1 with simplified "artificial" stimuli (for more details, see Methods). Examples of stimuli in which the oscillating visual tokens followed biological drumming movements can be seen in Movie 2, whereas examples in which the oscillating visual tokens followed nonbiological (triangular wave) drumming movements can be seen in Movie 3. PSSs for these stimuli were calculated by using the same procedure used in Experiment 1.

The results (Figure 4) show that the optimal delays to perceive audiovisual synchrony were very similar for

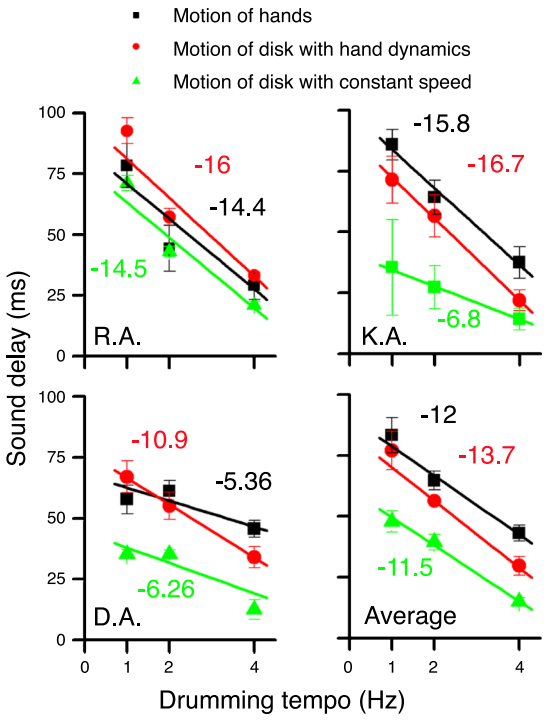

Figure 4. Delay that best supports perceptual synchrony for three different types of motion: disk stimuli of constant speed (green symbols), disk stimuli with human motion dynamics (red symbols), and real human motion (black symbols). The results for real and artificial motion with very similar dynamics were very similar, whereas those for constant speed oscillations showed systematically less asynchrony.

the three conditions. Red symbols show results for stimuli where the oscillating disk followed the same trajectory as the real motion, whereas green symbols refer to triangular wave (constant speed) motion. Black symbols show real drummer motion, taken from Figure 3. The results were similar but not identical for the three conditions: Best perceived simultaneity was always obtained by delaying the auditory pattern relative to the visual pattern, and the size of the optimal delay decreased as the drumming tempo increased. This relationship was characterized quantitatively by the best linear fits whose slopes range on average from 11.5 to 13.5 ms/Hz. However, although the slopes were similar, the triangular oscillation required lower absolute values of delay for subjective simultaneity than the other two conditions [statistically significant on a repeated measures one-way ANOVA, $F(2,8) = 18.8$, $p < .01$, followed by a multiple-comparison procedure, Holm–Sidak method]. For the two conditions where the motion followed the natural drumming motion, the slopes and absolute positions of the curves were similar (no statistically significant difference).

Thus, the perceptual asynchronies for real biological motion were the same as that for artificial biological motion of matched trajectories. Changing the trajectories to constant speed triangular waveforms produced less, not more, perceptual asynchrony. This result differs from that of Aymoz and Viviani (2004) for asynchronies between the visual attributes of motion and color.

Ranges of perceived audiovisual synchrony

An index having the size of the temporal window where stimuli were perceived to be in synchrony is given by the width of the best-fitting Gaussians curves (for an example of these curves, see Figure 2). Figure 5 shows these ranges of apparent synchrony (±1 *SD*) from the averaged data of the condition with natural stimuli (viewpoint elevation of 0 deg) as well as those of the two conditions with artificial stimuli.

For all three conditions, the ranges of delays supporting perceptual synchrony were mainly positive. Indeed, audiovisual synchrony was best obtained when the auditory stimuli were delayed relative to vision by as much as 150 ms for the 1-Hz condition but seldom when vision was delayed relative to audition. This indicates that synchrony perception can tolerate auditory delays but not visual delays. The range of delays supporting apparent synchrony depended on the drumming tempo, which is quite broad at 200 ms for 1 Hz and narrows to about 100 ms at 4 Hz. It is worth noting that it was the upper limit of the audiovisual synchrony range that reduced as drumming tempo increased. That the synchrony range is anchored at its negative extent is in line with previous data in the literature with both speech and nonspeech stimuli, showing more tolerance for perceptual integration for a late-arriving than for an early-arriving sound (Dixon & Spitz, 1980; Jaskowski, 1996; McGrath & Summerfield, 1985).

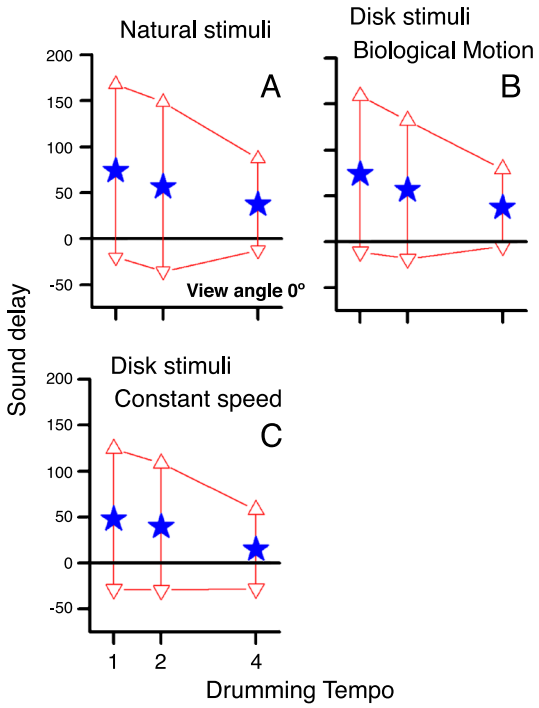

Figure 5. Data from Experiments 1 and 2 showing the ranges of perceived synchrony (averaged across participants) for three different conditions: (A) natural stimuli, (B) artificial stimuli following the biological motion profile, and (C) artificial stimuli moving at constant speed. The blue stars indicate the PSS, whereas the red triangles delimit the ranges of perceived cross-modal synchrony. For all conditions, the visual and auditory streams were more likely perceived to be in phase when the sounds were delayed relative to the visual patterns. The range of auditory delays supporting perceptual synchrony reduced with drumming tempo and did so unilaterally with the upper bound decreasing but the lower bound remaining stable (see "real-world constraints" in Discussion).

However, we show that the range of asynchronies supporting audiovisual perceptual synchrony narrowed for natural and artificial stimuli as drumming speed increases (regardless of the kind of motion). Indeed, for all three conditions, the range of delays supporting perceptual synchrony dropped from about 150–200 ms for drumming at 1 Hz down to just 80–90 ms for drumming at 4 Hz. This means that the resolution of the participants' temporal discriminations increased by about 45% when the drumming frequency of the oscillating visual stimuli increased from 1 to 4 Hz.

Experiment 3: Visual–auditory asynchronies with random motion

The previous experiments show that both the delay needed for audiovisual simultaneity and range of delays decrease with increasing drum tempos, both for natural and artificial stimuli and for biological and constant speed motion. To investigate further the dependency on tempo,

we devised a stimulus in which the drumming tempo within a given trial was not constant but variable. Single cycles of various tempos were randomly intermingled so that a given cycle of oscillation was unpredictable. Using stimuli similar to those of Experiment 2, we presented a single red disk that oscillated vertically at constant speed but with variable periodicity. In one condition (the "slow series"), the drumming rhythm randomly ranged from 1 to 4 Hz (2 octave range in half-octave steps), whereas in the other condition (the "fast series"), the drumming tempos ranged from 4 to 11.2 Hz (1.5 octaves in half-octave steps). Following the same procedures of the previous experiments, we measured the PSS by calculating the percentage of in-phase responses for each visual–auditory delay (ranging from −200 to 200 ms in steps of 50 ms), then obtained the best-fitting Gaussian curves to these data and calculating the delay at its peak. The distributions of in-phase responses for both experimental conditions (slow and fast series) together with the best-fitting Gaussian curves for the three participants are shown in Figure 6.

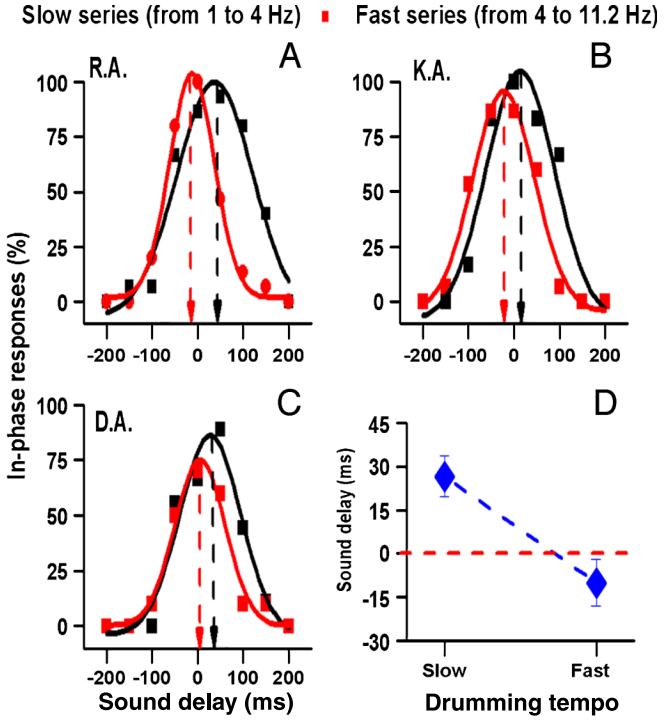

Figure 6. (A–C) Data from drumming at randomly varying tempos showing percentage of "in-phase" responses as a function of delay of the auditory stream (negative values indicate that the auditory stream was anticipated). Black symbols refer to the condition where the drum rhythm randomly varied between 1 and 4 Hz (slow series), and red symbols refer to the condition where the rhythm varied randomly from 4 to 11 Hz (fast series). The best-fitting Gaussian curves to the data (dashed lines) fit with r^2 ranging from .81 to .97. (D) Mean delay (with standard errors) for slow and fast random drumming.

The black symbols (with their fitted Gaussians) in Figure 6 refer to the slow random series (1–4 Hz), and the red symbols refer to the fast series (4–11 Hz). For the slow series, temporal alignment was perceived best when the sound was delayed relative to vision, as with the previous experiments. However, for the fast series, the opposite occurred, and vision had to be delayed slightly to produce perceptual synchrony (best brought out in Figure 6D). On average, a delay of about 30 and −10 ms (i.e., a delay of vision) was needed for perceptual synchrony of the slow and fast series, respectively.

Discussion

The experiments of this study report a number of consistent findings. Firstly, the perceptual asynchronies between visual and auditory signals in natural video sequences of conga drumming are very similar to those in simulated drumming sequences with matching dynamic structure, much more so than those measured with artificial stimuli of constant speed. Secondly, whereas sound must in general be delayed relative to vision to produce audiovisual synchrony, the size of the delay producing the best perceived synchrony is negatively correlated with drumming tempo: Higher drumming frequencies require less sound delay to be perceived as simultaneous. Thirdly, the range of audio delays supporting audiovisual simultaneity decreases as drumming tempo increases and does so in an asymmetrical manner.

One of the aims of this paper was to investigate whether temporal synchrony between senses depends on motion being real or biological, as it seems for purely visual stimuli (Aymoz & Viviani, 2004). Our data revealed no reduction in perceptual asynchronies for biological drumming: Indeed, they were actually larger as compared with artificial stimuli of constant speed. It is not clear at this stage why we did not confirm the results of Aymoz and Viviani (2004). It is possible that their proposal does not apply to cross-modal situations, but more experimental data are clearly needed before drawing firm conclusions. The experiments with random drumming tempos (Experiment 3) agree with the previous two experiments in showing that the auditory delay needed for perceptual synchrony decreases with drumming tempo.

Although the highest drumming tempo we investigated in Experiments 1 and 2 was 4 Hz, a temporal frequency that has been previously indicated as being near the limit for audiovisual synchrony perception (Fujisaki & Nishida, 2005; Recanzone, 2003; Shipley, 1964), extrapolation of the curves shown in Figures 3 and 4 suggests that the auditory delay required for perceptual synchrony should drop to zero for drumming tempos of 6–8 Hz and could become negative for higher frequencies. Consistent with this speculation are the results we obtained in the fast

random drumming (4–11 Hz) conditions investigated in Experiment 3. Indeed, these results not only suggest that audiovisual temporal alignment is still obtainable under some circumstances for frequency rates higher than 4 Hz (data distribution were indeed optimally fitted by Gaussian curves as shown by red curves in Figure 6) but also confirm that with high temporal frequency (on average above 6 Hz), negative auditory delays are required for perceptual synchrony. The results cannot be attributed to the fact that the higher drumming frequencies contained more cycles (hence, more opportunities to make the temporal comparison), as keeping number of cycles rather than duration constant had little effect on the results (green symbols in Figure 3).

One of the interesting findings of this study was that the range of delays that support audiovisual synchrony could be very broad, especially for low drumming tempos. However, it usually never extended far into the negative range (where sound precedes vision). This may just be a consequence of two independent factors, a shift in the mean and a decrease in the width of the synchrony functions with temporal frequency. Alternatively, it may mean that the perceptual system is more willing to integrate late-arriving sounds into a coherent audiovisual percept than a late-arriving visual signal. This was a consistent result across observers and across biological and nonbiological motion. Other researchers have found a similar tendency for both speech and nonspeech signals (Dixon & Spitz, 1980; Jaskowski, 1996; McGrath & Summerfield, 1985). Indeed, this constancy suggests a possible alternative explanation for our results: that irrespective of the stimulus frequency, participants are equally sensitive to negative audiovisual synchronies (where audio precedes visuals) that are very unlikely to occur in drumming events. If we further assume that the synchrony range is broader at low temporal frequencies, this could result in the shift in the peak of the distribution without necessarily implying that the best timing of subjective audiovisual synchrony also shifts. In fact, for the peak percentage of in-phase response at the peak of the 4-Hz functions, the percentage of in-phase responses is not higher for 4 Hz, as compared with the lower ones (Figure 2).

This asymmetry may reflect real-world constraints. There is a fixed lower limit to how much sooner a sound can arrive than vision, as vision involves a relatively slow transduction process compared with audition. Acoustic transduction between the outer and inner ears is a direct and fast mechanical process, taking 1 ms or less (Corey & Hudspeth, 1979; King & Palmer, 1985). Retinal phototransduction is a relatively slow photochemical process followed by several cascading neurochemical stages, lasting around 50 ms (Bolz, Rosner, & Wassle, 1982; Lamb & Pugh, 1992; Lennie, 1981; Rodieck, 1998; Schnapf, Kraft, & Baylor, 1987). Studies of audiovisual temporal alignment have generally found that an auditory stimulus needs to be delayed to be perceptually aligned with a visual stimulus (Bald, Berrien, Price, & Sprague, 1942; Bushara, Grafman, & Hallett, 2001; Hamlin, 1895; Hirsh & Sherrick, 1961; Lewkowicz, 1996; Rutschmann & Link, 1964). Thus, for near-field presentations, where auditory travel time is negligible, sounds will become perceptually available before visual stimuli by several tens of milliseconds. However, as sound source distance increases, this asynchrony decreases, and for distances of 10–15 m, acoustic and visual signals probably become perceptually available at approximately the same time. However, for distances beyond this point, sounds will always become available perceptually later than vision, and there is no upper limit to this. Thus, the broad positive extent of the delay range could reflect this real-world constraint.

The question remains, however, as to why both the PSS and the tolerance range should decrease with drumming tempo. The explanation is not clear, although it is noteworthy that similar observations have been made before. In a visual–temporal perception study, Clifford, Arnold, and Pearson (2003) investigated perceived synchrony between square wave oscillations in color and orientation. When observers had to judge whether the oscillating colors and orientations were "predominantly aligned" or not (in effect, a phase-based task), the asynchrony that was evident at low orientation frequencies (1 Hz) reduced progressively as orientation frequency increased, approaching zero and even reversing for the highest frequency (10 Hz). Paradoxically, this finding was dependent on the observer's task. When asked to judge if the changes were simultaneous or not (in effect, judging the instantaneous changes in the square wave oscillation), they found that asynchronies were more or less constant.

The task in our experiment was similar to Clifford et al.'s (2003) phase alignment task because observers tracked several cycles of drumming and then judged whether the sound sequence was synchronized with the vision sequence. Although the sound sequence was a series of impulses (rather than cycling oscillations), the task could not be based on instantaneous changes because the visual sequence was smoothly cyclical (like a series of half-sine waves; see Figure 1). Moreover, because the video display was a series of static frames that sampled the half-sine oscillations of the hands, there may not necessarily have been a frame that showed the actual point of contact between hand and drum. Thus, our task was necessarily a phase alignment task, and therefore our results could extend those of Clifford et al. from the visual domain into the audiovisual domain in showing that audiovisual judgments of temporal phase become less asynchronous as oscillation frequency increases. This indicates a better temporal discrimination under conditions involving faster drumming frequencies.

If observers were making a relative phase judgment, this would also explain why there was no effect of viewpoint elevation. We had thought that elevating the viewpoint would make it harder to determine the moment of contact between hand and drum because only in the 0-deg

condition is the varying distance between hand and drum clearly visible. With the precise moment of contact harder to determine at elevated viewing positions, we expected that the range of asynchronies supporting perceptual simultaneity would broaden. As Figure 5 shows, this was not the case, and elevation had virtually no effect. However, this is precisely what would be expected if judgments of synchrony were made based on phase discrimination because the cyclic nature of the visual signal is approximately equally evident at all viewpoint elevations.

The question of why the PSS would reduce with drumming tempo may also be related to recent electrophysiological recordings (Bair & Movshon, 2004) in monkey MT, an area specialized for visual motion. Responses in MT depend on both temporal and spatial frequencies of motion. At low temporal frequencies, responses are broadly distributed in time, with a long latency to peak response (approximately 80 ms). At high temporal frequencies, the response is narrower in time with a shorter latency to peak response (30–40 ms). Our results may therefore be explained by a delayed neural response from motion centers for lower drum tempos. Interestingly, if this explanation were correct, then it would account for the variations in audiovisual asynchronies purely in terms of a peripheral unimodal process rather than a central bimodal process.

Acknowledgments

Commercial relationships: none.
Corresponding author: Roberto Arrighi.
Email: arrighi@inoa.it.
Address: Istituto Nazionale di Ottica Applicata (INOA), Firenze, Largo E. Fermi 6, Italy.

References

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology, 52,* 1106–1130. [PubMed]

Arnold, D. H., Clifford, C. W., & Wenderoth, P. (2001). Asynchronous processing in vision: Color leads motion. *Current Biology, 11,* 596–600. [PubMed] [Article]

Aymoz, C., & Viviani, P. (2004). Perceptual asynchronies for biological and non-biological visual events. *Vision Research, 44,* 1547–1563. [PubMed]

Bair, W., & Movshon, J. A. (2004). Adaptive temporal integration of motion in direction-selective neurons in macaque visual cortex. *The Journal of Neuroscience, 24,* 7305–7323. [PubMed] [Article]

Bald, L., Berrien, F. K., Price, J. B., & Sprague, R. O. (1942). Errors in perceiving the temporal order of auditory and visual stimuli. *The Journal of Applied Psychology, 26,* 382–388.

Bolz, J., Rosner, G., & Wassle, H. (1982). Response latency of brisk-sustained (X) and brisk-transient (Y) cells in the cat retina. *The Journal of Physiology, 328,* 171–190. [PubMed]

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience, 12,* 4745–4765. [PubMed] [Article]

Burr, D. C. (1999). Vision: Modular analysis—Or not? *Current Biology, 9,* R90–R92.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *The Journal of Neuroscience, 21,* 300–304. [PubMed] [Article]

Clifford, C. W., Arnold, D. H., & Pearson, J. (2003). A paradox of temporal perception revealed by a stimulus oscillating in colour and orientation. *Vision Research, 43,* 2245–2253. [PubMed]

Corey, D. P., & Hudspeth, A. J. (1979). Response latency of vertebrate hair cells. *Biophysical Journal, 26,* 499–506. [PubMed] [Article]

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719–721. [PubMed]

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166,* 455–464. [PubMed]

Hamlin, A. J. (1895). On the least observable interval between stimuli addressed to disparate senses and to different organs of the same sense. *The American Journal of Psychology, 6,* 564–575.

Hirsh, I. J., & Sherrick, C. E., Jr. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology, 62,* 423–432. [PubMed]

Jaskowski, P. (1996). Simple reaction time and perception of temporal order: Dissociations and hypotheses. *Perceptual and Motor Skills, 82,* 707–730. [PubMed]

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. An experimental and theoretical analysis of calculus-like functions in visual data processing. *Psychological Research, 38,* 379–393. [PubMed]

Johansson, G. (1977). Studies on visual perception of locomotion. *Perception, 6,* 365–376. [PubMed]

King, A. J., & Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60,* 492–500. [PubMed]

Lamb, T. D., & Pugh, E. N., Jr. (1992). A quantitative account of the activation steps involved in photo-transduction in amphibian photoreceptors. *The Journal of Physiology, 449,* 719–758. [PubMed]

Lennie, P. (1981). The physiological basis of variations in visual latency. *Vision Research, 21,* 815–824. [PubMed]

Lennie, P. (1998). Single units and visual cortical organization. *Perception, 27,* 889–935. [PubMed]

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

Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal for Neuroscience, 7,* 3416–3468. [PubMed] [Article]

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science, 240,* 740–749. [PubMed]

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77,* 678–685. [PubMed]

Moutoussis, K., & Zeki, S. (1997). A direct demonstration of perceptual asynchrony in vision. *Proceedings: Biological Sciences / The Royal Society, 264,* 393–399. [PubMed]

Nakayama, K. (1985). Biological image motion processing: A review. *Vision Research, 25,* 625–660. [PubMed]

Nishida, S., & Johnston, A. (2002). Marker correspondence, not processing latency, determines temporal binding of visual attributes. *Current Biology, 12,* 359–368. [PubMed] [Article]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89,* 1078–1093. [PubMed] [Article]

Rodieck, R. W. (1998). *The first steps in seeing*. Sunderland MA.

Rutschmann, J., & Link, R. (1964). Perception of temporal order of stimuli differing in sense mode and simple reaction time. *Perceptual and Motor Skills, 18,* 345–352. [PubMed]

Schnapf, J. L., Kraft, T. W., & Baylor, D. A. (1987). Spectral sensitivity of human cone photoreceptors. *Nature, 325,* 439–441. [PubMed]

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145,* 1328–1330. [PubMed]

Viviani, P., & Aymoz, C. (2001). Colour, form, and movement are not perceived simultaneously. *Vision Research, 41,* 2909–2918. [PubMed]

Zeki, S. (1993). *A vision of the brain*. Oxford: Blackwell Scientific.