

Waka Fujisaki · Shin'ya Nishida

Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals

Received: 20 July 2004 / Accepted: 11 November 2004 / Published online: 20 July 2005
© Springer-Verlag 2005

Abstract Temporal synchrony is a critical condition for integrating information presented in different sensory modalities. To gain insight into the mechanism underlying synchrony perception of audio-visual signals we examined temporal limits for human participants to detect synchronous audio-visual stimuli. Specifically, we measured the percentage correctness of synchrony–asynchrony discrimination as a function of audio-visual lag while changing the temporal frequency and/or modulation waveforms. Audio-visual stimuli were a luminance-modulated Gaussian blob and amplitude-modulated white noise. The results indicated that synchrony–asynchrony discrimination became nearly impossible for periodic pulse trains at temporal frequencies higher than 4 Hz, even when the lag was large enough for discrimination with single pulses (Experiment 1). This temporal limitation cannot be ascribed to peripheral low-pass filters in either vision or audition (Experiment 2), which suggests that the temporal limit reflects a property of a more central mechanism located at or before cross-modal signal comparison. We also found that the functional behaviour of this central mechanism could not be approximated by a linear low-pass filter (Experiment 3). These results are consistent with a hypothesis that the perception of audio-visual synchrony is based on comparison of salient temporal features individuated from within-modal signal streams.

Keywords Psychophysics · Visual perception · Auditory perception · Time perception · Discrimination

Introduction

Temporal synchrony is a critical condition for integrating information presented in different sensory modalities. We have a clear sense of a single audio-visual source while watching movies, but that sense is disrupted if the audio and video tracks are largely misaligned in time. Modulations of the perception of one modality by the other modality (McGurk and MacDonald 1976; Sekuler et al. 1997) are also sensitive to the time lag of audio-visual signals (Munhall et al. 1996; Watanabe and Shimojo 2001).

Despite of its functional significance, however, we do not know much about the mechanism underlying the perception of temporal synchrony between auditory and visual signals. For the brain to judge synchrony between external events, active computation to find event synchrony is necessary (Johnston and Nishida 2001). Particularly, the presence of a neural mechanism specialised for audio-visual synchrony detection is indicated by the recalibration of the audio-visual simultaneity after adaptation to a constant time lag (Fujisaki et al. 2004), which probably reflects a post-adaptation change in the temporal tuning of audio-visual synchrony detectors.

After successful analysis of early sensory mechanisms from temporal frequency responses (e.g. Kelly 1979), this study examined the effects of the stimulus temporal frequency on the perception of audio-visual temporal synchrony.

Previous studies have demonstrated that it is hard to correctly judge temporal synchrony between visual flickers and auditory flutters when these stimuli are presented at rapid rates, and that the visual events are often seen as having the same time course as the auditory ones when there is a conflict between visual and auditory events (Gebhard and Mowbray 1959; Shipley 1964; Welch et al. 1986; Fendrich and Corballis 2001; Recanzone 2003; Morein-Zamir et al. 2003). This phenomenon, often called *auditory driving* (Shipley 1964), could be accounted for by the *modality appropriateness*

W. Fujisaki · S. Nishida (✉)
NTT Communication Science Laboratories,
NTT Corporation, 3-1 Morinosato Wakamiya,
Atsugi Kanagawa, 243-0198, Japan
E-mail: nishida@brl.ntt.co.jp
Tel.: +81-46-240-3594
Fax: +81-46-240-4716

hypothesis (Welch and Warren 1980; Shimojo and Shams 2001; Wada et al. 2003), which predicts the dominance of the sensory modality the more appropriate for the present task—audition dominates vision in temporal judgements because temporal acuity is better for audition.

Whereas these previous studies have focused on auditory driving under the condition where audio-visual correspondence fails, the question addressed in this study is why audio-visual correspondence fails for rapid stimulus changes. A simple account, schematically illustrated in Fig. 1a, is that the audio-visual simultaneity window (a range of time lags for two consecutive audio-visual signals to be perceived as simultaneous) is wide (Dixon and Spitz 1980; Lewkowicz 1996). In previous studies (e.g. Shipley 1964), as the temporal frequency of the modulation waveform became higher, the time lag between audio-visual signals became smaller, and the audio-visual signals are more likely to enter the

same simultaneity window. Alternatively, the failure of audio-visual synchrony detection at high temporal frequency may indicate a limitation of the audio-visual correspondence process in comparing rapid flicker/flutterers. Such a limit, if it exists, would afford a valuable clue for inferring the mechanisms underlying audio-visual synchrony perception.

To evaluate the perceptual accuracy of audio-visual synchrony detection, we measured the participants' performance in discriminating an asynchronous audio-visual stimulus from a physically synchronous one.¹ The results give an objective measure of the accuracy of audio-visual lag judgement, which characterises the perceptual mechanism that underlies the perception of audio-visual synchrony and binding.

In Experiment 1 we compared the performance of synchrony–asynchrony discrimination by systematically changing the temporal frequency and the time lag of audio-visual signals. If the window of simultaneity is the only limiting factor, the discrimination performance should be constant in terms of time lag, irrespective of the temporal frequency. If, on the other hand, the temporal frequency itself is another limiting factor, audio-visual synchrony–asynchrony discrimination should fail for high temporal frequencies even when the time lag is beyond the window of simultaneity. The results followed the latter prediction, suggesting a temporal-frequency limit of audio-visual synchrony detection. In experiments 2 and 3 we investigated the nature of this temporal-frequency limit.

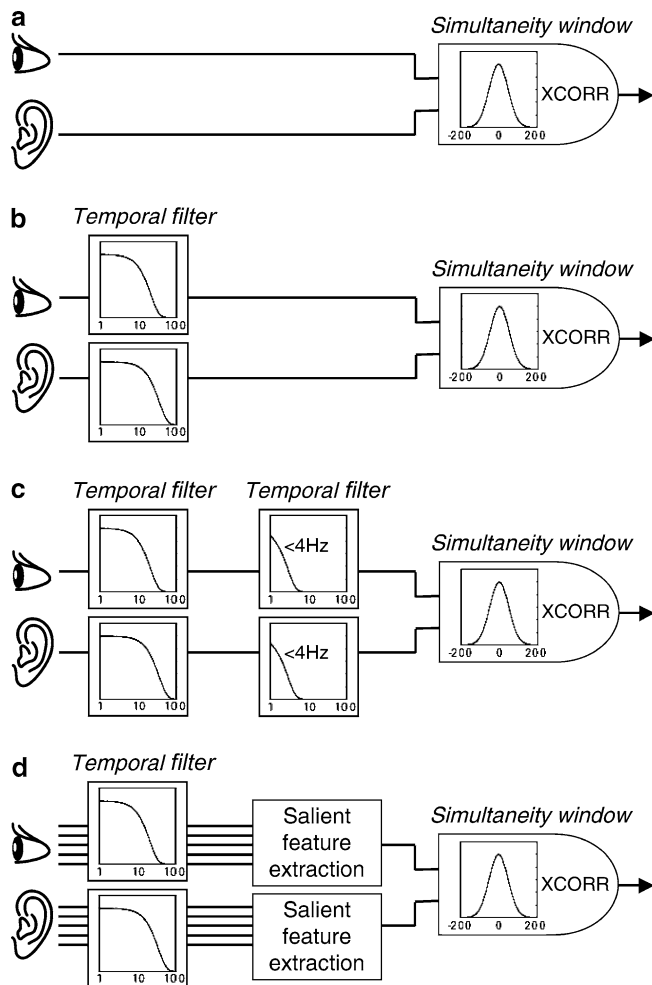


Fig. 1 Schematic representation of possible models of the algorithm for audio-visual synchrony detection. **a** Simple cross-correlation model. The operator at the right end takes cross correlation of audio-visual signals at various lags and checks whether there is a sufficient amount of correlated signals within a window of simultaneity (illustrated by a Gaussian function). **b** Peripheral (within-modal) low-pass filter model. **c** Central low-pass filter model. **d** Salient temporal feature-matching model

Experiment 1

Methods

Participants

Participants were the authors and three paid volunteers who were unaware of the purpose of the experiments. All had normal or corrected-to-normal vision and hearing. Informed consent was obtained after the nature and possible consequences of the studies were explained.

Apparatus and stimuli

Visual stimuli were presented with a VSG2/5 (Cambridge Research Systems). Auditory stimuli were presented with a TDT Basic Psychoacoustic Worksta-

¹An alternative method, i.e. measuring the probability of reporting apparent synchrony as a function of the audiovisual time lag (e.g. Dixon and Spitz 1980), could severely suffer from variation of the participants' criteria of "simultaneity". When a generous criterion is applied, the participants would judge different time lags as belonging to the same "synchrony" category even though they could discriminate one lag from another. This is also a serious problem under conditions in which auditory driving induces the perception of an illusory audiovisual synchrony irrespective of the audio-visual relationship

tion (Tucker-Davis Technologies). Both systems were controlled by Matlab (The MathWorks) running on a PC (Dell Precision 360). The time lags between audio-visual signals were defined at the onset and offset of the signals. Precise time control of audio-visual stimuli was accomplished by driving the TDT by a sync signal from the VSG. We used an oscilloscope to compare the time course of the visual signal (measured by a photocell attached to the face of the monitor) with that of the auditory signal (fed by the TDT) and determined system conditions that provided a stable alignment of the two signals to within a millisecond or better.

In a quiet dark room the participant sat 57 cm from a monitor (Sony GDM-F500, frame rate 160 Hz). The visual stimulus was a luminance-modulated Gaussian blob (standard deviation 2.0°) presented at the centre of the monitor screen (21.5 cd m^{-2} uniform field, 38.7° in width, 29.5° in height). The Gaussian blob had gradual luminance modulation to which the visual response is rapid (Kelly 1979). The luminance increment of the blob peak was temporally modulated between 0 and 43 cd m^{-2} . Nothing was visible during the off period. The fixation marker was a bullseye presented before stimulus presentation at the centre of the monitor screen. Participants were instructed to view the visual stimulus while maintaining their fixation at this location.

The auditory stimulus was a 100% amplitude-modulated white noise (54 dB sound pressure level (SPL) at the peak of modulation) presented diotically via headphones (Sennheiser HDA 200) with a sampling frequency of 24,420 Hz.

The audio-visual stimuli were modulated by the same periodic pulse train, resulting in a pair comprising a visual flicker and an auditory flutter, either in phase or out of phase between audio-visual signals. The durations of single pulses were one display frame for the visual stimulus (nominally $1,000/160 = 6.25 \text{ ms}$) and 6.25 ms for the auditory stimulus. To measure the width of the simultaneity window of a non-repetitive stimulus the audio-visual pair was also presented as a single pulse.

Procedure

Table 1 summarises the time lags between the audio-visual signals used for each condition. For periodic pulse

trains, the temporal frequency was changed between 1 and 16 Hz, and the audio-visual phase shift was 180° , 90° , 45° , 22.5° , or 11.25° . Note that the time lag for the condition of $X \text{ Hz}$ and Y° phase shift was $(1,000/X \times Y/360)$ (ms). The time lags for the single-pulse condition were 31–500 ms, which corresponded to the time lags of the 180° pulse train condition. Figure 2 shows an example of the modulation waveforms. Three pulse trains of 4, 2, and 1 Hz have a constant time lag of 0.125 s, which corresponds to 180° , 90° , and 45° phase shifts of each frequency, respectively.

The percentage correctness when discriminating synchrony–asynchrony for each stimulus condition (corresponding to a cell in Table 1) was measured in separate experimental blocks. Each block consisted of 20 trials, plus four initial practice trials, during which synchronous and asynchronous stimuli were presented in turn. In the next 20 trials, ten trials for each lag were presented in a random order. For a given stimulus condition, four blocks (80 trials) were conducted.

In a trial, 2 s after the last participant's response the fixation marker was removed and an audio-visual pair was presented, either synchronously or asynchronously. The participant had to make a two-alternative forced response by pressing a VSG response box key. A pulse train lasted 6 s with 2-s cosine ramps both at the onset and offset of the stimulus (Fig. 2). Also, to prevent the participants making a synchrony judgement based on the onset of the pulse train, the audio-visual phase delay started with a random value, gradually shifted to the intended phase over the initial 2 s, then kept that phase for the remaining 4 s. The participant was instructed to ignore the initial 2 s. Feedback was given after each response by the colour of the fixation marker, where blue indicated "synchronous" and red "asynchronous". We expected the feedback to exclude a type of error where the participants could discriminate the two lag conditions but not correctly label the physically "synchronous" pair as "synchronous".

Except for the 180° lag condition, audio-visual delays were positive (vision first) in half of the block and negative (audition first) in the other half. Each session consisted of four to nine blocks for the conditions selected mostly from the same row in Table 1. The time lags across blocks were in descending order (from large

Table 1 Audio-visual time lag conditions used in Experiment 1 (ms)

	Modulation frequency (Hz)								
	1	1.4	2	2.9	4	5.7	8	11.4	16
Phase shifts									
(Repetitive pulse trains)									
180°	500	356	250	175	125	88	63	44	31
90°	250	175	125	88	63	44	31		
45°	125	88	63	44	31				
22.5°	63	44	31						
11.25°	31								
Single pulses	500	356	250	175	125	88	63	44	31

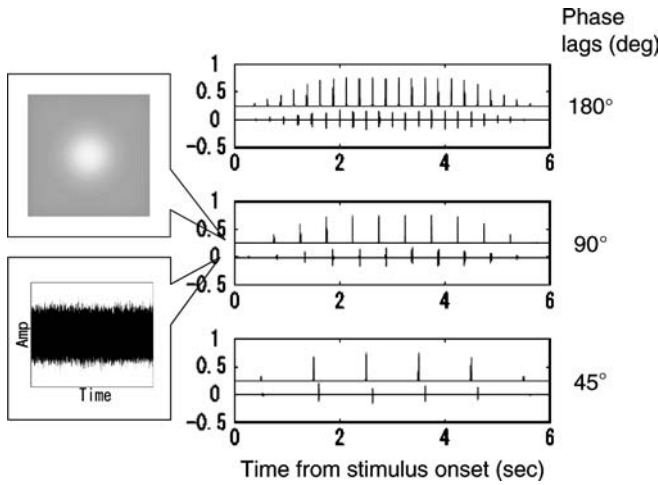


Fig. 2 Stimuli used in Experiment 1. The visual stimulus was a Gaussian blob (spatial profile is shown in the *upper inset*) and the auditory stimulus was white noise (carrier temporal waveform is shown in the *lower inset*). They were temporally modulated by a repetitive pulse train with or without an audio-visual lag. Each panel on the *right* shows the temporal waveform of the visual stimulus (*upper*) and the waveform of auditory stimulus for three temporal frequency conditions (4, 2, and 1 Hz), respectively. In these panels the three frequency conditions have a constant audio-visual time lag of 0.125 s, which corresponds to 180°, 90° and 45° phase shifts of each frequency, respectively

to small lags) in half of the sessions and in ascending order in the other half.

Results and discussion

For each lag condition, the discrimination performance (proportion correct) was plotted as a function of the time lag or temporal frequency, and then fitted by a logistic function of the form

$$P = 0.5 \times \frac{\gamma}{1 + \exp \{-\alpha(\log_{10} x - \log_{10} \beta)\}} + 0.5,$$

where α denotes the slope, β the horizontal position, and γ the maximum level of proportion correct ($P_{max} = 1$ when $\gamma = 1$). We used β as the estimate of discrimination threshold, which corresponds to the lag of the 75% correct point when $\gamma = 1$. Table 2 shows the threshold values estimated for each participant and for group data, with 95% confidence intervals estimated by the bootstrapping method (Efron and Tibshirani 1994).

Figure 3a shows the proportion correct for pulse trains and single pulses averaged over the five participants, plotted as a function of the time lags between audio-visual signals. When the audio-visual phase shift angle was smaller than or equal to 90°, the discrimination performance was nearly the same as that obtained for single pulses. The discrimination thresholds, estimated from the group-averaged psychometric function, were 74.6 ms for the single pulse condition and 80.9 ms for the 90° phase shift condition. For the 180° phase

Table 2 The discrimination threshold (the lag of the 75% correct point when the maximum level of proportion correct is 1) estimated for each participant for all conditions in Experiments 1, 2, and 3 (ms). The 95% confidence intervals estimated by the bootstrapping method are shown in parentheses. Mean: the average of individual thresholds. Pooled: threshold estimated from the group data

Participant	Experiment 1		Experiment 1 sub		Experiment 2		Experiment 3			
	Single pulses	180° Rep. pulses	90° Rep. pulses	Single pulses same location	Single Pulses Diff. location	180° Rep. pulses same location	180° Rep. pulses diff. location	Within-vision	Within-audition	180° Sine Waves
1(KT)	69.7 (65.2-75.0)	129.2 (123.9-134.3)	74.7 (69.0-81.0)					15.4 (14.7-16.2)	4.1 (4.0-4.2)	174.1 (173.3-175.1)
2(SO)	122.4 (117.2-128.2)	125.2 (121.7-127.6)	88.4 (84.8-92.1)					32.6 (30.9-34.3)	7.4 (7.3-7.5)	180.2 (175.7-184.9)
3(AM)	50.2 (47.9-52.6)	124.8 (123.2-126.6)	85.3 (80.9-90.3)	64.0 (61.9-66.0)	109.2 (105.3-113.7)	116.6 (114.5-121.4)	109.2 (105.3-113.7)	18.5 (18.4-18.6)	4.0 (3.9-4.0)	240.7 (231.3-249.4)
4(SN)	73.2 (69.7-76.9)	126.6 (123.9-129.5)	79.0 (75.7-82.5)	86.4 (81.2-91.6)	120.2 (117.6-122.8)	120.3 (120.0-120.7)	120.2 (117.6-122.8)	13.1 (12.8-13.5)	6.86 (6.86-6.87)	159.2 (147.1-165.3)
5(WF)	73.3 (69.9-77.0)	93.2 (90.3-96.5)	71.8 (68.7-75.2)	53.4 (51.0-55.9)	104.4 (101.2-107.8)	104.4 (101.2-107.8)	104.4 (100.8-108.1)	20.9 (19.9-22.1)	6.3 (5.8-7.3)	184.0 (180.1-188.8)
Mean	77.7	119.8	79.8	68.0	113.8	113.8	111.3	20.1	5.7	187.7
SD	26.7	14.9	7.0	16.8	8.3	8.3	8.1	7.6	1.6	31.1
Pooled	74.6 (72.9-76.4)	119.3 (117.5-121.2)	80.9 (79.4-82.6)	64.8 (63.2-66.6)	65.9 (64.2-67.6)	119.3 (119.0-119.6)	112.6 (110.2-114.8)	19.2 (18.6-19.7)	5.6 (5.3-6.4)	189.6 (186.7-191.8)

shift condition, however, the discrimination threshold was much larger (119.3 ms). A one-way analysis of variance (ANOVA) based on the log-transformed threshold for each participant indicated a significant main effect of the stimulus condition [$F_{(2,8)}=8.54$, $P<0.05$]. Tukey's HSD test indicated significant differences between the single pulse and 180° shift conditions ($P<0.05$) and between the 180° and 90° shift conditions ($P<0.05$), but not between the single pulse and 90° shift conditions ($P>0.10$).

The results indicate that the window of simultaneity is a common limiting factor for audio-visual synchrony detection of single pulses and repetitive pulse trains when the phase lag is equal to (or less than) 90° (Fig. 1a), but this is not the case for the 180° phase shift. In Fig. 3b, the same data are plotted as a function of the temporal frequency. This figure suggests that audio-visual correspondence cannot be established when the pulse repetition rate exceeds ~4 Hz.

Our data indicate that the best threshold lag for discriminating synchronous from asynchronous audio-visual pairs was approximately 80 ms. Although this is comparable with the size of the simultaneity window estimated from synchrony judgements (Dixon and Spitz 1980; Fujisaki et al. 2004), it is slightly larger than the minimum detectable audio-visual lag (just noticeable difference) estimated from the slope of the temporal order judgement (e.g. Spence et al. 2003). This is presumably because our task compared the magnitudes of two audio-visual lags, not the timings of audio and visual signals per se. Our simple simulation based on signal detection theory (Green and Swets 1966), indicated that the threshold should be $\sqrt{2}$ times larger for our task than for the temporal order judgement given audio-visual signal timings are independently perturbed in time.

Subsidiary experiment: effects of spatial location

In the main experiment, the visual stimulus was presented centrally, and the auditory stimulus was presented diotically through headphones. Because there have been several reports that temporal order judgement is more accurate when the light and sound are presented at different spatial locations (in different hemispheres) (Spence et al. 2003; Zampini et al. 2003a, b; but also see, Lewald and Guski 2003; Guski 2004), this pattern of results might change when the stimuli are presented at the same location or different hemispheres. In a subsidiary experiment, the effect of spatial location was tested for a limited number of participants (two authors and one naïve). The monitor was shifted rightwards in such a way that the fixation point at the left edge of the monitor screen was located right in front of the observer. A visual stimulus was presented 24.7° (25 cm) right of the fixation point. Two small identical speakers were positioned 24.7° right and left of the fixation point, respectively. They were set at approximately 4.02° below the height of fixation, just in front of the monitor screen,

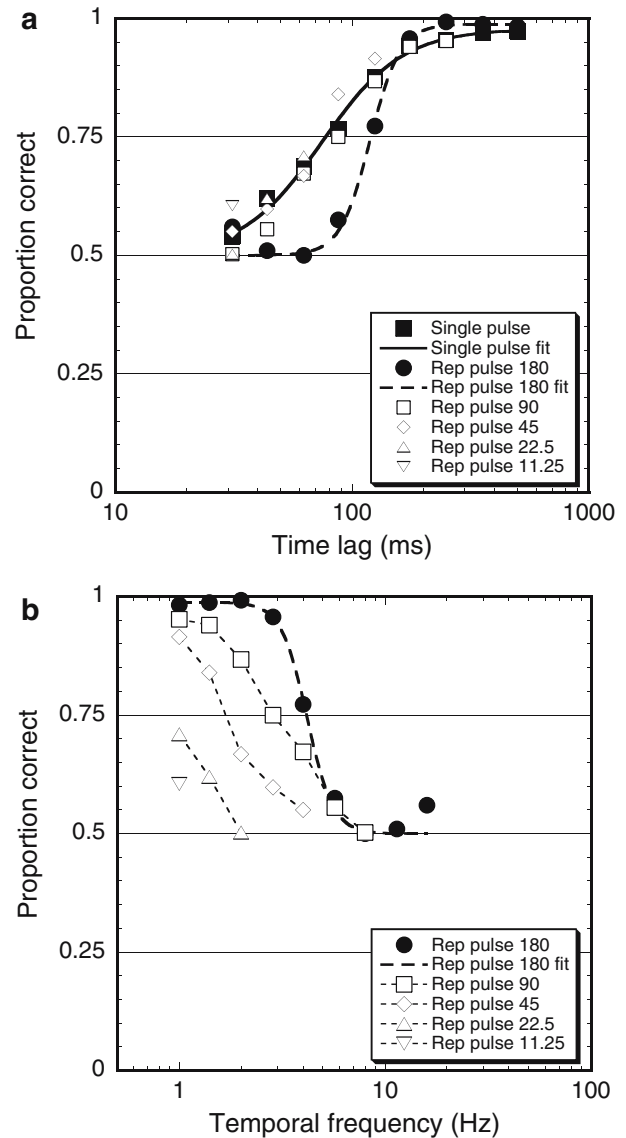


Fig. 3 **a** The proportion correct of audio-visual synchrony judgements for five audio-visual phase shift conditions (180°, 90°, 45°, 22.5°, and 11.25°) of repetitive pulse trains (*Rep pulse*) and for the single pulse condition, plotted as a function of the time lags between audio-visual signals. Average of results from the five participants. The *dashed line* indicates the fit by a logistic function for the 180° shift condition and the *solid line* indicates the fit for the single-pulse condition. **b** The same data (excluding the single-pulse data) re-plotted as a function of the temporal frequency

and at the same distance from the participant. Stimulus position was either “same” (vision right, auditory right) or “different” (vision right, auditory left), and stimulus waveform was either single pulse or periodic pulse train (180° phase shift).

The results (Table 2, “Experiment 1 sub”) show that the threshold obtained with the same location condition was very similar to that obtained with the different location condition. In addition, these thresholds were close to the values obtained in the main experiment,

where the sounds were localised inside the participants' heads.² The results indicate that audio-visual synchrony–asynchrony discrimination deteriorates at high temporal frequencies irrespective of the locations of audio-visual stimuli.

Experiment 2

The results of Experiment 1 can be interpreted as indicating that audio-visual synchrony detection is limited not only by the window of simultaneity but also by the stimulus temporal frequency. If this interpretation is correct, then what causes the deterioration of synchrony detection at high temporal frequencies? One possibility is the low-pass filtering by the peripheral within-modal processing (Fig. 1b). Vision is known to have lower temporal resolution than audition, which forms the basis of the modality appropriateness hypothesis (see “Introduction”). If the sluggish within-modal response is indeed a limiting factor of the audio-visual synchrony detection, the phase discrimination thresholds for purely visual or auditory tasks should be comparable to that for the cross-modal task. The second experiment tested this hypothesis.

Method

There were two conditions: vision–vision (VV) and audition–audition (AA). The stimulus for the VV condition was a luminance-modulated Gaussian blob. It was divided into left and right halves separated by a 0.15° gap. The two half blobs were modulated by the same periodic pulse train either in phase or 180° out of phase. The task was to judge whether they were synchronous or asynchronous. The modulation frequency was varied from 8 to 40 Hz. The stimulus for the AA condition was amplitude-modulated white noise (flutters) presented dichotically via headphones. Flutters presented in the left and right ears were modulated by the same periodic pulse train either in phase or 180° out of phase. The carriers were uncorrelated between the ears. The task of the participants was to judge whether the flutters presented in the left and right ears were synchronous or asynchronous. The modulation frequency was varied from 8 to 128 Hz for the AA condition. The participants and the other methods were the same as in Experiment 1.

²Why we did not find any effect of spatial location is not obvious, but a few points are worth mentioning. First, past studies showing positional facilitation employed tasks that are not free from response bias. If the effect of spatial location is to stabilise the response bias, it will not affect synchrony–asynchrony discrimination performance. Second, Keetels and Vroomen (2004) reported that the effect of location was small for participants giving a good performance. Because the participants of the subsidiary experiment were all well-trained, their performance might have already saturated

Results and discussion

Figure 4 shows the proportion correct for the VV and AA conditions averaged over five participants as a function of the modulation frequencies, together with the result obtained with the 180° periodic pulse trains in Experiment 1. The threshold temporal frequency for the VV condition, estimated from the group data, was 26.0 Hz (time lags: 19.2 ms), which is close to the known temporal resolution of the visual system. The threshold frequency for the AA condition was 89.3 Hz (time lags: 5.6 ms). These visual and auditory temporal frequency limits were far above the temporal frequency limit of the cross-modal condition (4.19 Hz, 119.3 ms). A one-way ANOVA of the log-transformed thresholds indicated a significant effect of the stimulus ($F_{(2,8)} = 173.28$, $P < 0.01$). Tukey's HSD test showed that the differences among cross-modal, VV, and AA conditions were all significant ($P < 0.01$). The results therefore indicate that the temporal limitation of audio-visual synchrony detection cannot be ascribed to peripheral low-pass filtering of each modality (Fig. 1b).

Previous studies have shown that the temporal frequency limit of within-vision synchrony detection is reduced when the stimulus separation is large (e.g. Forte et al. 1999; Victor and Conte 2002) or when different attributes are compared (Holcombe and Cavanagh 2001). These lower performances however should not be ascribed to the temporal resolution of the visual system, but to the properties of the synchrony detection process (see also “General discussion”). Under the conditions we used, within-modal synchrony–asynchrony detection was likely to be subserved by sensitive low-level sensors,

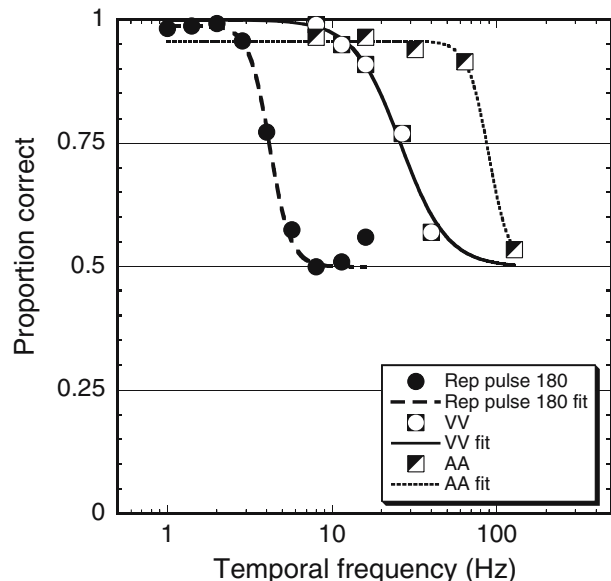
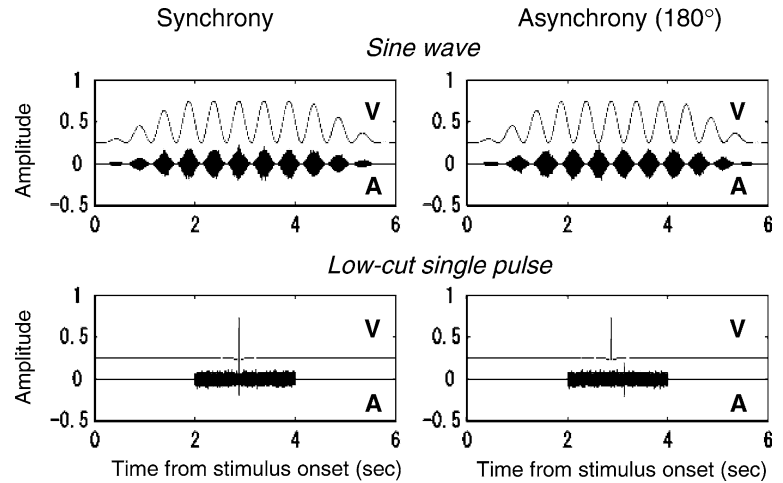


Fig. 4 The proportion correct for within-modal synchrony–asynchrony judgements as a function of the modulation frequency. Average for the five participants. VV within-vision, AA within-audition, Repulse 180 is the 180° periodic pulse train condition in Experiment 1 (cross-modal)

Fig. 5 Temporal waveforms for a sine wave (*upper*) and a low-frequency-cut single pulse (*lower*) used in Experiment 3. In each panel, the *upper and lower traces* indicate the amplitudes of visual stimulus (a Gaussian blob) and auditory stimulus (amplitude-modulated white noise), respectively. In the low-frequency-cut single pulse, as a result of the removal of slow temporal changes, the amplitude briefly falls below the baseline before and after a narrow pulse



such as first-order motion detectors for vision and binocular localisation detectors for audition. This point, however, does not undermine our conclusion, because the temporal resolution of the visual/auditory temporal responses feeding to the sensors should be equal to or better than the temporal-frequency thresholds obtained.

Experiment 3

The results of Experiment 2 showed that the temporal frequency limit of audio-visual correspondence is much lower than the limit predicted by peripheral within-modal low-pass filtering. Therefore, the temporal frequency limit of audio-visual correspondence should be ascribed to the property of a more central mechanism presumably located at or before cross-modal signal comparison. One possibility is a secondary linear low-pass filter whose cut-off frequency was ~ 4 Hz (Fig. 1c). To test this hypothesis, Experiment 3 measured the performance of synchrony detection for two modulation waveforms—sine waves and a low-frequency-cut single pulse. If there is a low-pass filter on luminance contrast and/or sound amplitude modulation, the cross-modal process should not be able to distinguish sine waves and periodic pulse trains at around the upper temporal limit, because the two waveforms differ from each other only in the second and higher harmonics. It was also predicted that the synchrony detection would be impossible for a low-cut single pulse consisting only of the frequency components well above 4 Hz, because no temporal modulation should be left after the pulse passes through the low-pass filter.

Method

As shown in Fig. 5, there were two stimulus conditions—sine wave and low-cut single pulse. In the former, the stimulus was modulated by a sine wave, either in phase or 180° out of phase between audio-visual signals.

As in a pulse train, the luminance increments against the background were 0 cd m^{-2} at the bottom of modulation and 43 cd m^{-2} at the peak of modulation. The auditory stimulus was 100% amplitude-modulated white noise (54 dB SPL at the peak of modulation) presented diotically via headphones. The data collection followed the procedure used for the 180° phase lag condition in Experiment 1.

The low-cut single pulse used in the latter condition was made by removing temporal frequency components lower than 16 Hz from a single pulse. The pulse modulation was transferred into the frequency domain by the fast Fourier transform (FFT), filtered by a hatbox low-cut filter, then transferred back to the time domain by the inverse FFT. Although this procedure allows a non-causal influence of the filtering, it introduces no relative phase shift. To present the filtered modulation that contained negative modulations, a noise floor (48 dB SPL) was added to the auditory stimulus. As a result, the amplitude of auditory signals changed between 30 (44 dB SPL) and 130% (57 dB SPL) relative to the amplitude of unfiltered single pulses. The time lag was either +250 ms (vision first) or -250 ms (audition first).

Results and discussion

Figure 6 shows the results obtained with the sine waves averaged over participants as a function of the modulation frequencies. The results obtained with the 180° periodic pulse trains in Experiment 1 are also shown in the same graph. The upper frequency limit for sine waves (2.64 Hz, time lags 189.6 ms) was significantly worse than that of the periodic pulse trains (4.19 Hz, time lags 119.3 ms) ($F_{(1,4)} = 22.70$, $P < 0.01$). This is incompatible with the prediction by the central low-pass filter model that the threshold should be nearly the same for sine waves and pulse trains.

For the low-cut single pulses, the discrimination performance was nearly perfect (85, 100, 92.5, 100, and 100; average: 95.5%) whereas the central low-pass model predicts chance-level (50%) performance. This

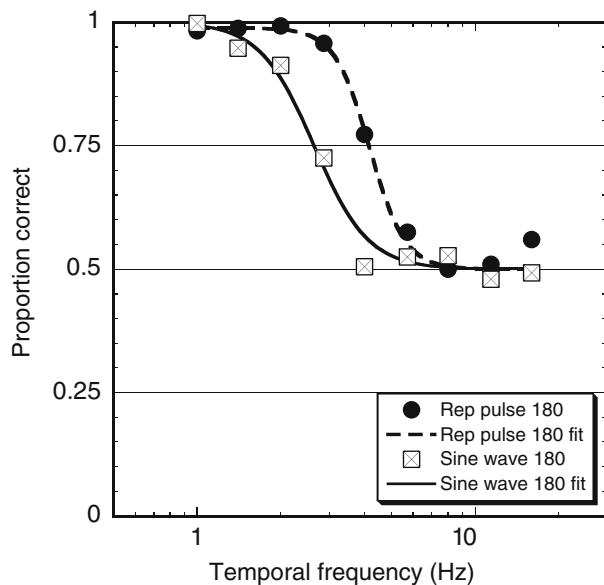


Fig. 6 The proportion correct for the sine waves plotted as the function of the modulation frequencies. Average for the five participants. *Repulse 180* is the 180° periodic pulse train condition in Experiment 1

provides another line of evidence against the central low-pass filter model.

General discussion

Experiment 1 showed that in terms of absolute lag the performance of synchrony–asynchrony discrimination was worse for the 180°-shift condition than for the other conditions. We interpreted this finding as indicating that temporal frequency is a limiting factor of audio-visual synchrony detection. However, another factor might make the 180°-condition harder than the other conditions. One might argue that the synchrony detector is activated not only at 0°, but also at 180° out-of-phase. This could be true if, for instance, one uses a square wave modulation in which the visual onset and auditory offset (or vice versa) are aligned at 180° shift.³ However, we used a train of brief pulse in which no alignment occurred at 180° shift. Neural offset responses may be delayed due to impulse responses of the sensory system

³In a preliminary observation, we measured the rate of perception of synchrony as a function of audio-visual lag. When the stimulus waveform was a square wave modulation whose mean was equal to the background, some participants made a significant number of false synchrony responses at ~180° shift. This error however disappeared when we reduced the background luminance to make the visual stimulus invisible during the off phase. To reduce the potential contribution of offset responses we used a similar background setting for all the stimuli in the main experiments (except for the low-cut single pulse). This could be an objection to the argument that offset response had some effects in the case of sinusoidal modulation.

(e.g. Roufs and Blommaert 1981), but the magnitude of delay is very unlikely to be as large as a half cycle of the upper frequency limit (~120 ms). In addition, the results obtained do not support the alternative interpretation either. First, if putative synchrony detectors are activated at 180° shift, at a given temporal frequency the discrimination performance should be worse for this phase shift than the 90° and other phase shift conditions. This relationship, however, was not observed at any temporal frequencies (Fig. 3b). Second, when plotted as a function of absolute lag (Fig. 3a), the psychometric function of the 180° condition overlaps with the other curves at larger lags, but falls with a steeper slope as the lag decreases. This pattern suggests that the performance of the 180° phase shift is as good as the others as long as the temporal frequency is low, but that it is impaired at higher temporal frequencies by a factor distinct from the window of simultaneity. We therefore interpret these results as supporting the notion that rapid stimulus changes make audio-visual synchrony detection very difficult.⁴

Experiments 2 and 3 suggest that the temporal limit reflects a central mechanism whose functional behaviour is not simply approximated by a linear low-pass filter.

The difficulty in judging the temporal phase of stimuli oscillating at or above 2–4 Hz is not a property specific to audio-visual synchrony detection. It is known that some within-modal inter-attribute bindings have temporal limits of the same order. For instance, the upper temporal limit is ~3 Hz for in-phase vs out-of-phase discrimination between an alternation of colour and an alternation of orientation (presented at separate locations) (Holcombe and Cavanagh 2001). The upper temporal limit is equally low for relative phase discrimination between alternation of colour and alternation of motion direction (presented either in the same object or in separate objects). Furthermore, for the colour and motion oscillating at 1–2 Hz, although relative phase judgement is possible, the point of perceptual synchrony largely deviates from the point of physical synchrony (Moutoussis and Zeki 1997; Nishida and Johnston 2002). These temporal limitations could be ascribed to the involvement of a high, slow, and attentive process for binding different visual attributes, which computes the temporal relationships among the signals carried by separate channels based on the salient features individuated from each channel (Lu and Sperling 1995; Holcombe and Cavanagh 2001; Nishida and Johnston 2002).

The results obtained in this study are consistent with a hypothesis that a similar slow, attentive process underlies the temporal synchrony detection of audio-visual signals (Fig. 1d). The poor performance at high temporal modulation frequency can be explained by the difficulty in individuating salient temporal features from

⁴Another line of support comes from the finding that audio-visual synchrony detection is nearly impossible for randomly generated pulse trains when the pulse temporal density is high (Fujisaki and Nishida 2004)

rapid repetitive signals. The lower temporal limit of sine waves relative to pulse trains can be ascribed to the quality of temporal features. That is, peaks and troughs, which could be salient features, are not sharply defined in sine waves. On the other hand, because a low-cut single pulse contains a sharp peak, temporal correspondence of audio-visual signals should be easy.

Previous studies have shown the effect of the within-modal grouping on cross-modal binding. When the simultaneous sound was embedded into the sequence of other identical sounds, the effect of auditory-induced visual motion illusion was attenuated (Watanabe and Shimojo 2001), and the effect of facilitation for a detection of a visual target in a sequence of distracters disappeared (Vroomen and de Gelder 2000). The effect of visual grouping on the visual capture of auditory motion has also been reported (Sanabria et al. 2003). In our pulse train stimulus, when the flicker/flutter frequency was increased the visual and auditory pulses were more likely to be grouped together within each modality. The accompanying increase in the within-modal grouping may explain the failure of cross-modal binding at high temporal frequencies. Note however that this interpretation is not incompatible with our hypothesis, because the increase in the within-modal grouping must be the main limiting factor for the attentional process to extract individual features from the stimulus stream.

At temporal frequencies slightly higher than 4 Hz, auditory driving often occurs (Shipley 1964; Recanzone 2003). According to the current results, this is the condition where visual and auditory changes are perceptible, but detection of audio-visual asynchrony is hard. Auditory driving demonstrates how the brain binds the audio-visual signals under this paradoxical situation.

Finally, the present findings may contribute to diagnosis in electrophysiology and brain imaging to find neural correlates of perceptual synchrony among many multisensory areas responding to audio-visual signals (Meredith et al. 1987; Bushara et al. 2003). Specifically, those areas should respond differently to in-phase and out-of-phase audio-visual repetitive stimuli if, and only if, the stimulus temporal frequency is less than ~ 4 Hz.

Acknowledgements We thank Makio Kashino (NTT), Shinsuke Shimojo (CalTech), Alan Johnston (UCL), Derek Arnold (UCL), and the Human Frontier Science Program (RGP0070/2003-C).

References

- Bushara KO, Hanakawa T, Immisch I, Toma K, Kansaku K, Hallett M (2003) Neural correlates of cross-modal binding. *Nat Neurosci* 6:190–195
- Dixon NF, Spitz L (1980) The detection of auditory visual desynchrony. *Perception* 9:719–721
- Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. *Monographs on statistics and applied probability*, no. 5. Chapman & Hall, New York
- Fendrich R, Corballis PM (2001) The temporal cross-capture of audition and vision. *Percept Psychophys* 63:719–725
- Forte J, Hogben JH, Ross J (1999) Spatial limitations of temporal segmentation. *Vision Res* 39:4052–4061
- Fujisaki W, Nishida S (2004) Temporal characteristics of synchrony detection of audiovisual signals. Paper presented at the Fifth Annual Meeting of the International Multisensory Research Forum, Universitat de Barcelona, Barcelona, Spain
- Fujisaki W, Shimojo S, Kashino M, Nishida S (2004) Recalibration of audiovisual simultaneity. *Nat Neurosci* 7:773–778
- Gebhard JW, Mowbray GH (1959) On discriminating the rate of visual flicker and auditory flutter. *Am J Psychol* 72:521–529
- Green DM, Swets JA (1966) Signal detection theory and psychophysics. Krieger, Huntington
- Guski R (2004) Spatial congruity in audiovisual synchrony judgements. *Proceedings of 18th International Congress on Acoustics*, pp 2059–2062
- Holcombe AO, Cavanagh P (2001) Early binding of feature pairs for visual perception. *Nat Neurosci* 4:127–128
- Johnston A, Nishida S (2001) Time perception: brain time or event time? *Curr Biol* 11:R427–R430
- Keetels M, Vroomen J (2004) Spatial separation affects audio-visual temporal synchrony. Paper presented at the Fifth Annual Meeting of the International Multisensory Research Forum, Universitat de Barcelona, Barcelona, Spain
- Kelly DH (1979) Motion and vision: II. Stabilized spatio-temporal threshold surface. *J Opt Soc Am* 69:1340–1349
- Lewald J, Guski R (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Brain Res Cogn Brain Res* 16:468–478
- Lewkowicz DJ (1996) Perception of auditory-visual temporal synchrony in human infants. *J Exp Psychol Hum Percept Perform* 22:1094–1106
- Lu ZL, Sperling G (1995) The functional architecture of human visual motion perception. *Vision Res* 35:2697–2722
- McGurk H, MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264:746–774
- Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J Neurosci* 7:3215–3229
- Morein-Zamir S, Soto-Faraco S, Kingstone A (2003). Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res* 17:154–163
- Moutoussis K, Zeki S (1997) A direct demonstration of perceptual asynchrony in vision. *Proc R Soc Lond B Biol Sci* 264:393–399
- Munhall KG, Gribble P, Sacco L, Ward M (1996) Temporal constraints on the McGurk effect. *Percept Psychophys* 58:351–362
- Nishida S, Johnston A (2002) Marker correspondence, not processing latency, determines temporal binding of visual attributes. *Curr Biol* 12:359–368
- Recanzone GH (2003) Auditory influences on visual temporal rate perception. *J Neurophysiol* 89:1078–1093
- Roufs JA, Blommaert FJ (1981) Temporal impulse and step responses of the human eye obtained psychophysically by means of a drift-correcting perturbation technique. *Vision Res* 21:1203–1221
- Sanabria D, Soto-Faraco S, Chan J, Spence C (2003) Does perceptual grouping precede multisensory integration? Evidence from the crossmodal dynamic capture task. Paper presented at 4th Annual Meeting of the International Multisensory Research Forum, McMaster University, Hamilton, Ontario, Canada
- Sekuler R, Sekuler AB, Lau R (1997) Sound alters visual motion perception. *Nature* 385:308
- Shimojo S, Shams L (2001) Sensory modalities are not separate modalities: plasticity and interactions. *Curr Opin Neurobiol* 11:505–509
- Shipley T (1964) Auditory flutter-driving of visual flicker. *Science* 145:1328–1330
- Spence C, Baddeley R, Zampini M, James R, Shore DI (2003) Multisensory temporal order judgements: when two locations are better than one. *Percept Psychophys* 65:318–328
- Victor JD, Conte MM (2002) Temporal phase discrimination depends critically on separation. *Vision Res* 42:2063–2071

- Vroomen J, de Gelder B (2000) Sound enhances visual perception: cross-modal effects of auditory organization on vision. *J Exp Psychol Hum Percept Perform* 26:1583–1590
- Wada Y, Kitagawa N, Noguchi K (2003) Audio-visual integration in temporal perception. *Int J Psychophysiol* 50:117–124
- Watanabe K, Shimojo S (2001) When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol Sci* 12:109–116
- Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. *Psychol Bull* 88:638–667
- Welch RB, DuttonHurt LD, Warren DH (1986) Contributions of audition and vision to temporal rate perception. *Percept Psychophys* 39:294–300
- Zampini M, Shore DI, Spence C (2003a) Audiovisual temporal order judgments. *Exp Brain Res* 152:198–210
- Zampini M, Shore DI, Spence C (2003b) Multisensory temporal order judgments: the role of hemispheric redundancy. *Int J Psychophysiol* 50:165–180