Article:
Stone, J.V., Hunkin, N.M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M. and Porter, N.R. (2001) When is now? Perception of simultaneity. Proceedings of the Royal Society B: Biological Sciences, 268 (1462). pp. 31-38. ISSN 1471-2954

http://dx.doi.org/10.1098/rspb.2000.1326

THE ROYAL SOCIETY

doi 10.1098/rspb.2000.1326

When is now? Perception of simultaneity

J. V. Stone[1*], N. M. Hunkin[2], J. Porrill[1], R. Wood[1], V. Keeler[1], M. Beanland[1], M. Port[1] and N. R. Porter[1]

[1]*Department of Psychology, and* [2]*Department of Clinical Neurology, University of Sheffield, Sheffield S10 2UR, UK*

We address the following question: Is there a difference (D) between the amount of time for auditory and visual stimuli to be perceived? On each of 1000 trials, observers were presented with a light–sound pair, separated by a stimulus onset asynchrony (SOA) between −250 ms (sound first) and +250 ms. Observers indicated if the light–sound pair came on simultaneously by pressing one of two (yes or no) keys. The SOA most likely to yield affirmative responses was defined as the point of subjective simultaneity (PSS). PSS values were between −21 ms (i.e. sound 21 ms before light) and +150 ms. Evidence is presented that each PSS is observer specific. In a second experiment, each observer was tested using two observer–stimulus distances. The resultant PSS values are highly correlated ($r = 0.954$, $p = 0.003$), suggesting that each observer's PSS is stable. PSS values were significantly affected by observer–stimulus distance, suggesting that observers do not take account of changes in distance on the resultant difference in arrival times of light and sound. The difference RT_d in simple reaction time to single visual and auditory stimuli was also estimated; no evidence that RT_d is observer specific or stable was found. The implications of these findings for the perception of multisensory stimuli are discussed.

Keywords: vision; audition; simultaneity; awareness; temporal

1. INTRODUCTION

When executing time-critical tasks, such as playing table tennis, knowing precisely when the ball made contact with the table is important for fast and accurate motor coordination. However, even if the the perception of audio-visual simultaneity is not veridical, it should at least be stable for a given observer. Such stability may permit the motor system to be temporally calibrated with respect to the perceived timing of auditory and visual events. These considerations suggest that the perceived timing of visual and auditory events should be highly accurate, or, at least, highly stable for a given observer.

Between 1861 and 1865 Hirsch used the clockwork Hipp chronoscope to demonstrate that reaction time (RT) to visual stimuli is greater than the RT to auditory stimuli (Woodworth & Schlosberg 1954, p. 10). Typical reaction times to auditory stimuli (RT_a) and visual stimuli (RT_v) are $RT_a = 140$ ms and $RT_v = 180$ ms (Woodworth & Schlosberg 1954). In a seminal paper, Hershenson (1962) varied the stimulus onset asynchrony (SOA) between a briefly presented flash and a sound (noise burst): observers responded as quickly as possible as soon as either stimulus was detected. The SOA between the flash and the subsequent sound of each flash–noise pair varied randomly between 0 and 85 ms. It was found that the mean RT to asynchronous flash–noise pairs RT_{av} was minimal at an SOA approximately equal to the difference $RT_d = (RT_v - RT_a)$ in mean RT to single audio and visual stimuli. This suggests that the mean RT to asynchronous flash–noise pairs can be accounted for in terms of the difference between mean RTs to single audio and visual stimuli.

It might be thought that RT_d (the difference in RT to single audio or visual stimuli), or RT_{av} (the minimal RT associated with asynchronous audio–visual pairs) can be used as a measure of the difference D in time required for auditory and visual stimuli to be perceived. This is a category error: a RT is the time required to execute a reaction, whereas D is the difference in time required for an auditory and visual stimulus to be perceived. This type of argument applies to both RT_d and RT_{av}. Moreover, the fact that a stimulus evokes a response does not imply that an observer was aware of the stimulus before initiating a response. To take an extreme example, the response to a painful stimulus is mediated by spinal reflexes, and the rapid startle reflex to a loud bang may occur before conscious awareness of the sound. A RT is, at best, an indirect measure of D, and relies on unspoken assumptions regarding the relationship between RT and perception.

Recently, Moutoussis & Zeki (1997*a*) used a novel method to demonstrate that colour information is perceived 60–80 ms before motion information. Observers looked at 30 randomly positioned squares which moved up and down with a periodic motion. The colour of all of the squares changed synchronously between red and green at the same frequency as the motion, but the phase of the colour and motion changes varied between trials. On each trial, observers indicated if the squares were both green when they moved upward and red when they moved downward. Given the frequency of oscillation, responses were translated into a time lag associated with perceived changes in motion relative to the lag associated with perceived changes in colour. Results indicated that colour is perceived *ca.* 70–80 ms before motion, although individual variations in the range 40–90 ms can be estimated from their results (Moutoussis & Zeki 1997*a*, p. 395, fig. 3*b*,*c*). On the basis of results obtained in a series of papers (e.g. Moutoussis & Zeki 1997*b*), Zeki & Bartels (1998) argue that conscious awareness of a particular physical attribute (such as colour) depends critically on the activity induced within the corresponding neocortical region. If this type of argument applies across different sensory modalities then it

*Author for correspondence (j.v.stone@sheffield.ac.uk).

Received 26 June 2000 *Accepted* 25 September 2000

implies that the timing of conscious awareness of visual and auditory stimuli should depend on the timing of processing within visual and auditory areas, respectively.

In this paper, we define a measure (the PSS) of the difference D in time for auditory and visual stimuli to be perceived. Importantly, this measure is not contaminated by intermediate temporal processes, such as the RT associated with executing a motor response. We predicted that, even if the PSS is not veridical, it should be stable for a given observer in order to facilitate calibration of time-critical motor tasks.

A note on nomenclature: the time required for stimuli to be perceived is the same as the time for stimuli to reach conscious awareness. The term 'perceived' does not involve the many connotations associated with the term 'conscious awareness'. We therefore prefer to use the term 'perceived' wherever possible in this paper.

2. MATERIAL AND METHODS

(a) *Experiment 1: estimating the point of subjective simultaneity*

The experiment consisted of two tasks, a simultaneity judgement task, and a reaction time task. Before these tasks, the purpose of the experiment was explained to each subject, and a written instruction sheet was provided. The order in which the two tasks were executed was counter-balanced across observers, and both tasks were run automatically by computer. The experiment took about 50 min.

(i) *Observers*

The observers were nine male and 14 female undergraduate psychology students (mean age = 20.9 years, s.d. = 3.42 years, range = 18–36 years).

(ii) *Apparatus*

The light stimulus was a red light-emitting diode (LED), positioned on a computer keyboard in front of the observer at a distance of 50 cm (luminance = 11 cd m^{-2}), in a dimly lit room. The sound stimulus was a 250 Hz square-wave tone delivered through headphones at 71 dB. The intensities of both stimuli were well above threshold in order to minimize the differential effect of intensity on sensory integration time (Woodworth & Schlosberg 1954). The timing accuracy of the stimulus onset times was accurate to less than 1 ms. The stimuli were controlled from a Macintosh 8100 computer, via a National Instruments board.

(iii) *Simultaneity judgement task*

To estimate D (recall that D is the difference in time for audio and visual stimuli to be perceived) as directly as possible, the task requires a 'yes' or 'no' decision regarding the perceived simultaneity of a light and sound stimulus, presented with a SOA that varied randomly across trials. The SOA at which a given observer was most likely to respond in the affirmative is the point of subjective simultaneity (PSS), and was taken to be an estimate of D.

On each of 1050 trials, each observer was presented with a light and a sound, separated by a SOA. An observer indicated whether or not the sound and light came on simultaneously by pressing one of two (yes or no) response keys; both the light and sound stimulus were switched off automatically once a response was made. (The stimuli were kept on until a response was obtained to ensure that observers could not base their responses on the SOA between fixed-length stimuli being switched off). Observers were requested to respond as quickly and as accurately as possible. The SOAs varied between −250 ms (sound first) and +250 ms (light first). The first 50 trials were treated as practice trials, and were discarded. For the remaining 1000 trials, stimulus pairs with every SOA in the set $S = \{-250, -249, \ldots, -1, 1, \ldots, 249, 250\}$ were presented twice, with SOAs being chosen from S in the same random order for all subjects. The observer was given an opportunity to take a short break every 100 trials. The inter-trial interval varied randomly (with uniform probability) between 1300 and 1700 ms.

(iv) *Reaction time task*

Sixty light stimuli were presented, followed by 60 sound stimuli (the order of these was counterbalanced between observers). These were the same stimuli as used in the simultaneity judgement task. The inter-trial interval was varied randomly (with a uniform probability) between 1300 and 1700 ms. Each observer was required to respond as quickly as possible by pressing one key. The stimulus was switched off automatically as soon as a response was made.

(b) *Experiment 2: effect of observer–stimulus distance*

During two separate sessions, each of five observers judged the simultaneity of a sound–light stimulus at two observer–stimulus distances, with the sound stimulus delivered via a speaker. These 'near' and 'far' sessions were at least 24 h apart. Increasing the observer–stimulus distance effectively delays the sound stimulus, relative to the light stimulus. Therefore, the PSS and RT_d values should be altered by a change in observer–stimulus distance, unless observers discount the effects of distance.

(i) *Observers*

The observers were four males and one female, all aged 21 years.

(ii) *Apparatus*

The apparatus was the same as in experiment 1 except that the sound stimulus was delivered via a 5 cm speaker, and the LED was attached to the top of this speaker.

(iii) *Procedure*

This was identical to experiment 1, except for the following changes. Each observer was tested twice, once in each of two separate sessions. In the 'near' session, the stimulus (i.e. speaker and light) was placed 0.5 m away from the observer; in the 'far' session the stimulus was placed 3.5 m away from the observer. To provide cues to stimulus distance, the ambient lighting was increased slightly, and standard sized drink cans were placed on the table between the observer and the far stimulus. The order in which observers were tested in the near and far conditions was counterbalanced across observers. The SOA varied between −250 and 300 ms over a total of 1150 trials (including 50 practice trials, as in experiment 1). The interval between trials varied with uniform probability between 1100 and 1900 ms. The interval between the near and far sessions for each observer was −75, 45, 26, −31 and −164 h for observers 1, 2, 3, 4 and 5, respectively: a positive value indicates that the near condition preceded the far condition.

Table 1. *Simultaneity judgement task and reaction time task*

(Simultaneity judgement task. The point of subjective simultaneity (PSS) is the SOA at which an observer is most likely to perceive the onset of a light and a sound as simultaneous. All times are in units of milliseconds, and all quantities (except n and RT_d) are maximum-likelihood (ML) estimates (see Appendix A). PSS is the ML estimate of the PSS, and $\sigma(PSS)$ is the ML estimate of its standard deviation (s.d.). $\hat{s}$ is the estimated s.d. of the distribution of 'yes' responses (see figure 1), and $\sigma(\hat{s})$ is its estimated s.d. $\hat{a}$ is the estimated probability of observing a 'yes' response at a SOA equal to the PSS, and $\sigma(\hat{a})$ is the estimated s.d. in $\hat{a}$. n is the total number of 'yes' responses out of 1000 trials. Observer data has been ordered according to PSS. RT task. RT_v is the ML estimate of the mode of the distribution of 60 RTs to visual stimuli presented alone, RT_a is the corresponding mode for auditory stimuli, and $RT_d = RT_v - RT_a$.)

	simultaneity judgement task							RT task		
observer	PSS	$\sigma(PSS)$	$\hat{s}$	$\sigma(\hat{s})$	$\hat{a}$	$\sigma(\hat{a})$	n	RT_v	RT_a	RT_d
1	−21	4.5	103	3.6	0.77	0.04	391	229	202	27.5
2	−6	4.0	98	3.1	0.89	0.04	413	181	165	16.1
3	3	6.2	158	6.6	0.80	0.03	563	226	195	31.1
4	8	5.7	109	4.9	0.57	0.03	304	233	175	58.6
5	16	5.6	153	5.6	0.91	0.04	596	193	171	22.3
6	30	3.9	105	3.1	0.98	0.04	480	182	160	22.0
7	32	7.6	202	9.1	0.90	0.03	704	214	178	36.5
8	37	3.8	109	3.2	1.00	0.04	546	202	172	29.7
9	43	4.8	123	4.1	0.88	0.04	504	249	208	40.4
10	52	5.1	139	4.6	0.97	0.04	593	201	198	2.9
11	75	7.4	153	6.9	0.78	0.03	514	220	188	32.3
12	81	16.9	241	20.7	0.65	0.03	507	199	207	−8.4
13	82	5.7	145	5.1	0.95	0.04	608	189	177	12.8
14	87	14.0	217	15.5	0.68	0.03	529	218	197	20.5
15	90	10.4	264	13.3	0.99	0.03	788	225	203	21.8
16	102	13.7	333	20.6	1.00	0.03	837	206	180	25.9
17	150	17.4	198	14.3	0.70	0.03	465	218	205	12.6
mean	51	8.0	168	8.5	0.85	0.03	550	211	187	23.8

Table 2. *Stability of PSS and RT_d*

(Five observers were tested twice on the simultaneity judgment task and the reaction time task. The correlation between corresponding PSS values across both test sessions is $r = 0.95$ ($p < 0.01$), and the corresponding correlation between RT_d values is $r = 0.06$ ($p > 0.05$). The interval between the first and second sessions for each observer was 75, 45, 26, 31 and 164 h, respectively. See table 1 for a description of each column heading.)

	simultaneity judgement task							RT task		
observer	PSS	$\sigma(PSS)$	$\hat{s}$	$\sigma(\hat{s})$	$\hat{a}$	$\sigma(\hat{a})$	n	RT_v	RT_a	RT_d
near condition										
1	−25	6.4	148	6.5	0.75	0.03	504	204	178	26.5
2	−24	3.1	73	2.2	0.96	0.05	356	217	186	31.2
3	−1	3.5	96	2.7	0.98	0.04	469	212	183	29.2
4	48	3.8	106	3.1	1.00	0.04	531	212	196	16.2
5	8	7.8	192	9.3	0.81	0.03	633	237	196	40.9
far condition										
1	−24	6.0	150	6.0	0.81	0.03	547	205	190	15.0
2	−38	3.2	71	2.3	0.93	0.05	337	206	184	21.8
3	−17	3.4	93	2.6	1.00	0.04	479	217	178	39.5
4	35	4.8	130	4.2	0.93	0.04	574	216	201	14.8
5	11	8.8	201	10.9	0.77	0.03	608	230	229	1.2

The intensities of the sound and light stimuli, as measured at the observer's position, were adjusted to be equal to those in experiment 1, in both the near and far conditions. Data were analysed as in experiment 1: no observer's data failed the goodness-of-fit tests described in Appendix A.

3. RESULTS

(a) *Experiment 1: estimating the point of subjective simultaneity*

Results for the simultaneity judgement task and the reaction time task are summarized in table 1.

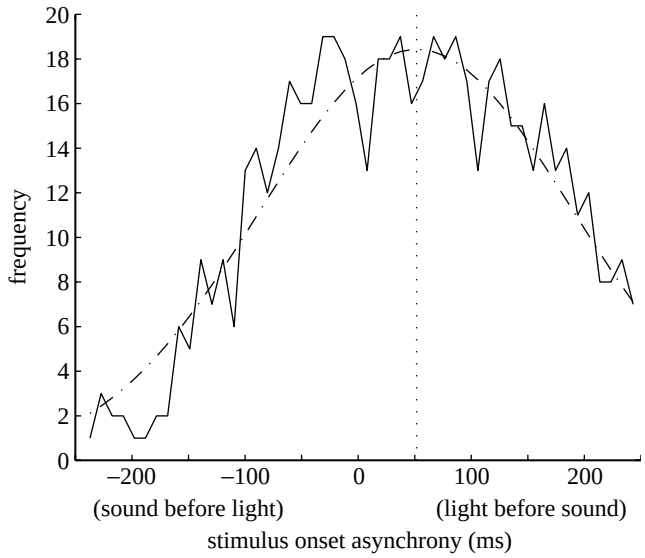

Figure 1. Histogram of 'yes' responses to the question, 'Were the onset times of the light and sound simultaneous?', as a function of sound–light SOA, for one observer. ML estimation was used to estimate the mode of the distribution of responses, which is defined as PSS. Solid line, frequency of 'yes' responses; dotted line, Gaussian function fitted using ML estimation (see Appendix A). In this example, the PSS is 52 ms (i.e. if the light came on 52 ms before the sound then the sound and light were perceived as having simultaneous onset times). The ML estimate of the standard deviation in the above distribution of 'yes' responses is $\hat{s} = 139$ ms, and the ML estimate of the standard deviation in the value of PSS is $\sigma(\mathrm{PSS}) = 5.1$ ms.

(i) *Simultaneity judgement task*

As the SOA was varied from −250 to 250 ms, the probability of an observer responding 'yes' (to the question, 'Were the onset times of the light and sound simultaneous?') increased and then decreased (see figure 1). The resultant distribution of responses was fitted to a Gaussian function for each observer, using maximum-likelihood estimation (see Appendix A). The mode of this fitted distribution is an estimate of the PSS for one observer. The goodness of this Gaussian fit was tested, which resulted in six out of 23 data sets being discarded (see Appendix A). Most (five out of six) of these data sets were discarded because the distribution of responses was essentially flat, as if observers were responding at random. The remaining 17 out of 23 data sets form the basis of the results reported here.

The PSS values vary across observers between −21 and 150 ms. Most PSS values are positive, implying that sound stimuli are perceived before light stimuli. Typical values for the estimated standard deviation in PSS are *ca.* 9 ms. Fourteen observers' PSS values are more than 1.96 s.d. away from zero, and are therefore statistically different from zero ($p < 0.05$).

The variation in PSS values (and their small standard deviations $\sigma(\mathrm{PSS})$) across observers suggests that each observer has a PSS that is statistically different from most other observers. Moreover, the distribution of PSS values across observers appears to be non-Gaussian. Evidence for this can be obtained by evaluating the difference between each PSS value and the estimated population mean PSS. Based on the 17 PSS values and their standard deviations $\sigma(\mathrm{PSS})$, the estimated population mean and standard deviation are $\overline{\mathrm{PSS}} = 29.321$, $\overline{\sigma(\mathrm{PSS})} = 1.402$ (see Appendix A), respectively. If the observed PSS values were derived from a Gaussian distribution then 95% of observed PSS values would fall within 1.96 s.d. of the estimated population mean. In fact, only two PSS values are within 1.96 observer s.d. of the estimated population mean. This implies that the underlying distribution of PSS values is not described by a Gaussian function. Additionally, a Kolmogorov–Smirnov goodness-of-fit test shows that the 17 PSS values are not significantly different from a uniform distribution ($z = 0.961$, $p > 0.05$). Together, these results indicate that the distribution of PSS values is essentially uniform, and that each observer has a PSS value that is specific to that observer.

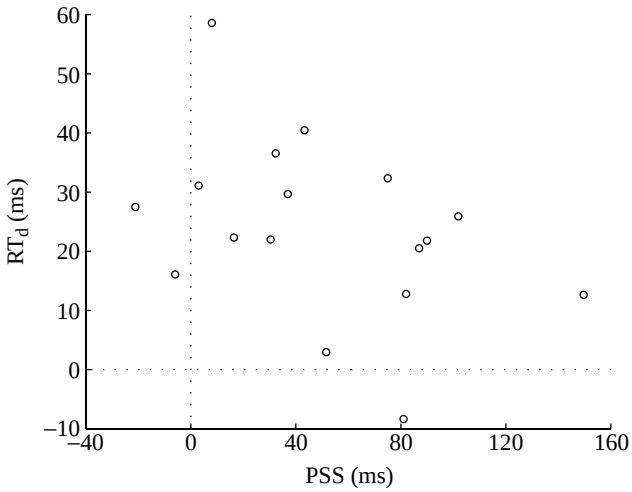

Figure 2. Plot of PSS versus $\mathrm{RT_d}$ for the 17 observers in experiment 1.

(ii) *Reaction time task*

Each observer's RTs to auditory and visual stimuli ($\mathrm{RT_a}$ and $\mathrm{RT_v}$, respectively) were estimated using maximum-likelihood estimation (see Appendix B). These were then used to estimate a value $\mathrm{RT_d} = \mathrm{RT_v} - \mathrm{RT_a}$ for each observer.

If both $\mathrm{RT_d}$ and PSS measure the difference in time D for visual and auditory stimuli to be perceived then they should be positively correlated across our sample of 17 observers, and they should have the same sample mean. However, a paired t-test shows that the difference between the mean PSS (51 ms) and the mean $\mathrm{RT_d}$ (21.4 ms) approaches significance ($t = 2.095$, $p = 0.0524$, d.f. = 16). Additionally, observer-specific PSS and $\mathrm{RT_d}$ values are negatively correlated $r = -0.400$ ($p = 0.055$). However, examination of figure 2 reveals that this correlation depends largely on a small number of outliers, rather than a general trend. Omitting one of several of these outliers substantially reduces the significance of this correlation. For example, omitting observer number 4 yields $r = -0.327$ ($p = 0.108$).

(b) *Experiment 2: effect of observer–stimulus distance*

(i) *Simultaneity judgement task*

The PSS for each observer in each condition was calculated using ML estimation, as in experiment 1, and results are shown in table 2. The significance of the

Table 3. *Effect of distance on PSS and PSS values adjusted for travel time of sound over 3* ms

(The PSS of five observers was measured with the audio-visual stimulus at two different distances, near (0.50 m) and far (3.5 m). The predicted reduction in PSS from PSS_N (near condition) to PSS_F (far condition) is significant (one-tailed z-test) for three observers with small PSS standard errors. The first five columns of this table are copied from table 2. Adding 11 ms (the time for sound to travel 3 m) to PSS_F values changes the three significant z-values into non-significant z'-values (using a two-tailed z-test).)

observer	PSS_N	PSS_F	σ_N	σ_F	z	$p(z)$	z'	$p(z')$
1	−25	−24	6.4	6.0	0.10	0.54	1.36	0.18
2	−24	−38	3.1	3.2	−3.27	0.00	−0.79	0.43
3	−1	−17	3.5	3.4	−3.18	0.00	−0.95	0.34
4	48	35	3.8	4.8	−2.14	0.02	−0.34	0.73
5	8	11	7.8	8.8	0.26	0.60	1.19	0.23

difference ($PSS_N - PSS_F$) for each observer was evaluated using z-tests as described in Appendix C, and results are summarized in table 3.

The observer–stimulus distance in the near and far conditions differed by 3 m, a distance travelled by sound in only 11 ms. The predicted difference in PSS is therefore 11 ms, with the far condition having smaller predicted PSS values. A one-tailed z-test revealed a significant effect of distance for three out of five observers, for whom the difference between PSS_N and PSS_F values also has the predicted sign. These three observers have small values for the estimated standard deviation in PSS, which suggest that the noise levels of the remaining two observers is simply too high to enable a difference as small as 11 ms to be evaluated.

As a simple test of the hypothesis that PSS_F was reduced by 11 ms relative to PSS_N, the differences ($PSS_N - PSS_F$) were re-evaluated with a two-tailed z-test after 11 ms had been added to each PSS_F value, as shown in table 3. If an observer's difference between PSS_N and PSS_F is accounted for by the extra 11 ms travel time of sound in the far condition then no significant difference should remain after adding 11 ms to PSS_F. As predicted, none of the modified differences ($PSS_N - (PSS_F + 11)$) were significant. This suggests that observers do not discount the effects of distance when making judgements of simultaneity.

As a further test of this hypothesis, a simple regression of the near session PSS values (PSS_N) against the far session PSS values (PSS_F) yielded the regression equation: $PSS_N = 0.971\ PSS_F + 7.607$ ms ($R^2 = 0.909$, $t = 5.478$, $p = 0.012$). Given that the two data sets were acquired in the near and far conditions, the predicted intercept value is 11 ms. However, the standard deviation associated with the estimated intercept of 7.607 ms is $\sigma_m = 5.643$ ms. Thus, the estimated intercept (7.607 ms) is not significantly different from the predicted intercept (11 ms) ($z = 0.601$, $p > 0.05$).

These near and far sessions were at least 24 h apart. Consequently, the results of this experiment were also used to test the stability of PSS and RT_d over time. The correlation between PSS values obtained in the near and far conditions is $r = 0.953$ ($p = 0.008$). Additionally, the slope of the regression line (see above) is approximately equal to unity. Thus, the PSS_N and PSS_F values are not only highly correlated, they co-vary with a ratio of approximately 1:1. Further evidence that the PSS is stable is given by the stability of its associated standard deviation σ(PSS): the correlation of σ(PSS) values between near and far sessions is $r = 0.967$ ($p = 0.0014$).

(ii) *Reaction time task*

Given that the sound stimulus takes 11 ms longer to reach the observer in the far condition than in the near condition, we would predict that RT_a is 11 ms longer in the far condition than in the near condition. A one-tailed paired t-test indicated no significant difference between the values of RT_a in the near and far conditions ($t = 1.250$, $p = 0.140$, d.f. $= 4$). The correlation between RT_v in the near and far conditions is $r = 0.814$ ($p = 0.036$, d.f. $= 4$), and that for RT_a is $r = 0.747$ ($p = 0.062$, d.f. $= 4$). Despite these reasonably stable values for RT_v and RT_a between conditions, the correlation between RT_d values obtained in the near and far conditions is $r = -0.282$ ($p = 0.319$, d.f. $= 4$). However, it should be noted that these results are based on a relatively small number (60) of RTs for each condition.

4. DISCUSSION

Given an operational definition PSS of D, we set out to answer the following question: Is there a difference D between the amount of time required for auditory and visual stimuli to be perceived?

Our answer can be summarized as follows. First, most observers have a PSS value that is significantly different from zero. Second, PSS values are observer specific; each observer's PSS value is significantly different from most other observers' PSS values and from the estimated population mean PSS value. Third, the difference between the mean observer PSS and the mean observer difference RT_d (between RTs to audio and visual stimuli) approaches significance.

Additionally, experiment 2 provides evidence that the value of PSS, but not the value of RT_d, is stable over time for each observer; and that observers do not take account of changes in observer–stimulus distance on the difference in arrival times of light and sound when making judgements of simultaneity.

One possible confound might exist if visual and auditory stimuli were able to mask each other. However, results reported by Massaro & Kahn (1973) exclude the possibility that sound is masked by light. On each trial, observers were presented with an 800 Hz target sound for 20 ms. This was followed 0–500 ms later by a masking stimulus, which was either a light or an 800 Hz square-wave sound. Observers were required to report whether the target sound was sharp (saw-tooth waveform) or dull (sinusoidal waveform). Performance increased from 60 to 90% as the target–mask interval increased from 0 to 500 ms for the sound mask. In contrast, performance remained essentially unaltered at 90% at all target–mask intervals for the light mask. Whilst this result suggests that there was no masking of sound by light in our

experiments, the masking of light by sound remains a logical possibility.

(a) *Cortical mediation of the point of subjective simultaneity*

Conscious awareness of the simultaneity of audio-visual aspects of stimuli self-evidently requires activity within the visual and auditory systems to be monitored. Critical questions are: Which parts of these systems are monitored, and which 'higher-order' cortical circuits are responsible for monitoring them?

The earliest neuronal activation induced by auditory and visual stimuli occur within the superior colliculus (SC). In cats, an auditory stimulus evokes SC activity within 13 ms, whereas a visual stimulus evokes SC activity within 65–100 ms (Stein & Meredith 1993). In humans, evoked response potential (ERP) studies suggest that the mean P_1 ERP component occurs 104 ms after onset of a visual stimulus, and 76 ms after onset of an auditory stimulus (Andreassi & Greco 1975).

The mean difference in the earliest ERP component (P_1) of 28 ms is consistent with the mean 23.8 ms ($\sigma_m = 3.65$ ms) of 17 RT_d values (the difference in RT to auditory and visual stimuli) observed here in experiment 1, and with the mean 23.6 ms ($\sigma_m = 3.88$ ms) of ten RT_d values in experiment 2, where σ_m is the estimated standard error in the mean. It is also consistent with the values of RT_d reported by Hershenson (1962) and Andreassi & Greco (1975), and with the SOA (between visual and auditory stimuli) associated with a minimal RT (RT_{av}) (Hershenson 1962). In experiment 1, the estimated population mean is $\overline{PSS} = 29.321$, $\overline{\sigma(PSS)} = 1.402$, which is consistent with the ERP value of 28 ms.

As stated in § 1, Zeki & Bartels (1998) argue that conscious awareness of visual and auditory stimuli depends critically on activation with associated cortices. Following this line of reasoning, we can hypothesize that conscious awareness of the simultaneity of audio-visual aspects of stimuli depends critically on the timing of activity in associated cortices. The simplest prediction based on this hypothesis is that audio-visual stimuli are perceived as being simultaneous if they activate auditory and visual cortices at exactly the same time. According to the ERP study reported by Andreassi & Greco (1975), simultaneous activation of auditory and visual cortices suggests that the SOA between light and sound should be 28 ms. As described in the preceding paragraph, the estimated population mean PSS value ($\overline{PSS} = 29.321$, $\overline{\sigma(PSS)} = 1.402$) is consistent with a value of 28 ms. On the basis of these mean figures, we cannot therefore reject the hypothesis that audio-visual stimuli are perceived as simultaneous if they activate auditory and visual cortices at exactly the same time.

(b) *The point of subjective simultaneity is observer specific*

The individual variation in PSS and σ(PSS) values suggests that the estimated population mean PSS is derived from observer-specific PSS values. Indeed, in experiment 1, only two PSS values are not significantly different from the estimated population mean. Additionally, a Kolmogorov–Smirnov goodness-of-fit test shows that the 17 PSS values in experiment 1 are not significantly different from a uniform distribution. Together, these results suggest that each observer has a PSS value that is specific to that observer.

The inter-observer variability in PSS is consistent with results reported by Moutoussis & Zeki (1997*a*), which used colour and motion. We analysed data derived from Moutoussis & Zeki (1997*a*), p. 395, fig. 3*b*,*c*), and estimated that the difference in time required to perceive colour and motion varies between 40 and 90 ms in different individuals. Results presented here and by Moutoussis & Zeki (1997*a*) are therefore consistent with the hypothesis that the time required for different components of the perceptual system to process information is observer specific.

(c) *The point of subjective simultaneity is stable*

One critical requirement of D (the difference in time between conscious awareness of simultaneous visual and auditory aspects of a single stimulus) is that it is stable for a given observer. If D were variable then the apparent simultaneity between visual and auditory stimuli would vary from day to day. Such variability could disrupt time-critical motor tasks involving multisensory stimuli (e.g. playing squash, hunting). The observer-specific PSS clearly meets this requirement, as demonstrated in experiment 2. In contrast, RT_d does not appear to be stable. However, our results with regard to RT_d should be interpreted with caution because of the relatively small number (60) of trials used.

(d) *Observer–stimulus distance and the point of subjective simultaneity*

Experiment 2 was designed to test the hypothesis that observers take account of the effect of observer–stimulus distance when making judgements of simultaneity. Three of the five observers had significantly different PSS values in the near and far conditions. After taking account of the predicted effect of distance on the PSS in the far condition (PSS_F) (by adding 11 ms to PSS_F), all five differences ($PSS_N - PSS_F$) became non-significant. This suggests that observers do not discount the effects of distance when making judgements of simultaneity.

It might be supposed that the PSS confers some advantage in terms of discounting the different speeds of sound and light, and thereby discounts their different arrival times at sensory organs. For a given (positive) PSS, there exists a stimulus–observer distance at which physically simultaneous visual and audio aspects of a stimulus would be perceived as simultaneous. This is because, as a stimulus is moved further away the arrival time of sound is progressively delayed, whereas the arrival time of light is essentially unaffected. This PSS-equivalent distance might be at arm's length (for manual work), or typical of the distance between two people in conversation. For example, given an observer with a PSS of 50 ms, an audio-visual stimulus with physically simultaneous audio and visual components would be perceived as simultaneous only if the stimulus–observer distance was 16.6 m (assuming sound travels at 331.3 m s^{-1} and that the travel time of light is negligible). Whilst most positive values of PSS obtained here are difficult to reconcile with this type of interpretation, the negative values of PSS would be associated with a (physically impossible) negative PSS-equivalent distance. Together, these results are

inconsistent with the hypothesis that the PSS acts to discount the different arrival times of audio and visual aspects of stimuli at sensory organs.

5. CONCLUSIONS

We have defined a measure, PSS, of the difference D in time for auditory and visual stimuli to be perceived. Importantly, the PSS does not depend on RT, and is not therefore contaminated by intermediate temporal processes associated with executing a fast motor response. Based on maximum-likelihood estimation, our results indicate that the PSS is observer specific, and that it is stable over time. We have argued that such stability is critical for accurately calibrating the timing of motor commands involved in time-critical tasks. Thus, whilst the inter-observer variability of PSS values remains unexplained, the stability of observer-specific PSS values has a compelling ecological explanation.

Thanks to P. North, M. Westby, D. Buckley, D. Johnston, S. Booth, P. Coffey and P. Furness for useful discussions, and to two anonymous referees for their comments. This research was supported by a Mathematical Biology Wellcome Fellowship (grant no. 044823).

APPENDIX A. MAXIMUM-LIKELIHOOD ESTIMATION OF THE POINT OF SUBJECTIVE SIMULTANEITY

Given $n = 1000$ binary responses for each observer, the probability of a 'yes' response appeared to vary as a Gaussian function of SOA. Accordingly, the responses of each observer were fitted to a Gaussian function. The mode of this fitted distribution is an estimate of the PSS. To avoid any misunderstanding, note that the ML estimation procedure described here does not involve fitting a Gaussian function to a histogram of responses.

A Gaussian function is defined by three parameters $\theta = (\mu, \sigma, a)$, where μ is the mean, σ is the standard deviation, and a is the maximum amplitude of the Gaussian function. The mean and mode are equal for a Gaussian distribution, so that the mode can be estimated as μ. If an observer's responses can be modelled with a Gaussian distribution then the probability p_1 of observing a 'yes' response $r_i = 1$ at an SOA equal to x_i ms is

$$p_1(r_i = 1|x_i, \mu, \sigma, a) = a \exp[-(\mu - x_i)^2/2\sigma^2\ ], \qquad \text{(A1)}$$

where μ is the SOA at which a 'yes' response is most likely to be observed, a is the probability associated with a 'yes' response at the SOA $x_i = \mu$, and σ is the standard deviation associated with responses (see figure 1). It follows that the probability p_0 of a 'no' response $r_i = 0$ at an SOA equal to x_i ms is $(1 - p_1)$:

$$p_0(r_i = 0|x_i, \mu, \sigma, a) = 1 - a \exp[-(\mu - x_i)^2/2\sigma^2]. \qquad \text{(A2)}$$

The probability of observing a particular set of responses can be computed if we assume that responses to different SOAs are made independently of each other. For a given set of n_1 'yes' responses $\boldsymbol{r}_1 = \{r_1, \ldots, r_{n_1}\}$, with corresponding SOAs $\boldsymbol{x}_1 = \{x_1, \ldots, x_{n_1}\}$, the probability P_1 of observing these responses at $\boldsymbol{x}_1$ is

$$P_1(\boldsymbol{r}_1|\boldsymbol{x}_1, \mu, \sigma, a) = \prod_{i=1}^{n_1} a \exp[-(\mu - x_i)^2/2\sigma^2]. \qquad \text{(A3)}$$

Similarly, the probability of observing n_0 'no' responses $\boldsymbol{r}_0 = \{r_1, \ldots, r_{n_0}\}$, at SOAs $\boldsymbol{x}_0 = \{x_1, \ldots, x_{n_0}\}$ is

$$P_0(\boldsymbol{r}_0|\boldsymbol{x}_0, \mu, \sigma, a) = \prod_{i=1}^{n_0} (1 - a \exp[-(\mu - x_i)^2/2\sigma^2]). \qquad \text{(A4)}$$

Given the combined set of $n = (n_1 + n_0)$ (n_1 'yes' and n_0 'no' responses), the probability of observing responses $\boldsymbol{r} = \{\boldsymbol{r}_0, \boldsymbol{r}_1\} = \{r_1, \ldots, r_n\}$ at corresponding SOAs $\boldsymbol{x} = \{\boldsymbol{x}_0, \boldsymbol{x}_1\} = \{x_1, \ldots, x_n\}$ is defined by the likelihood function $L(\mu, \sigma, a)$:

$$\begin{aligned} L(\mu, \sigma, a) &= P_1 \times P_0 \\ &= \prod_{i=1}^{n_1} a \exp[-(\mu - x_i)^2/2\sigma^2] \\ &\quad \times \prod_{i=1}^{n_0} (1 - a \exp[-(\mu - x_i)^2/2\sigma^2]) \\ &= \prod_{i=1}^{n} (a \exp[-(\mu - x_i)^2/2\sigma^2])^{r_i} \\ &\quad \times (1 - a \exp[-(\mu - x_i)^2/2\sigma^2])^{(1-r_i)}. \end{aligned} \qquad \text{(A5)}$$

If we consider (μ, σ, a) to be variables of the likelihood function L with fixed parameters $(\boldsymbol{x}, \boldsymbol{r})$ then we can seek values of (μ, σ, a) that maximize L. These are known as the maximum-likelihood estimates of (μ, σ, a) (Cowan 1998). For each observer, the maximum-likelihood estimate $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{a})$ of $\theta = (\mu, \sigma, a)$ was obtained by maximizing L with respect to θ. This was achieved by minimizing $-L$ with the simplex method, using the Matlab function *fmins*. The value of θ was initialised to $(0, 100, 0.9)$, and different initial values had negligible effects on results. The standard deviation associated with each parameter in θ was obtained for each observer as the square root of diagonal elements of the matrix $V = -H^{-1}$, where H is the Hessian of the function $\log L(\mu, \sigma, a)$ (Cowan 1998, p. 78). This Hessian was estimated numerically at $\theta = \hat{\theta}$.

Each observer's data set was evaluated with three goodness-of-fit tests, using a significance criterion of $p = 0.05$ or $p = 0.95$ (as appropriate) for each test. First, a χ^2-test was used to test if the frequency distribution of 'yes' responses was uniformly distributed; this resulted in five data sets being discarded. Next, a different χ^2-test and a Kolmogorov–Smirnov test were used to evaluate the goodness of fit of the remaining 18 data sets to a Gaussian distribution. One data set failed both of these tests, and was discarded. The three tests therefore collectively excluded six from a total of 23 data sets.

(a) *Maximum-likelihood estimation of the population mean point of subjective simultaneity*

The $\mathcal{N} = 17$ PSS values and their estimated standard deviations $\sigma(\text{PSS})$ can be combined (Sivia 1996) to form a ML estimate of the population mean $\overline{\text{PSS}}$ and standard deviation $\overline{\sigma(\text{PSS})}$:

$$\overline{\mathrm{PSS}} = \frac{\sum_{i=1}^{N} w_i \mathrm{PSS}_i}{\sum_{i=1}^{N} w_i}, \qquad \overline{\sigma(\mathrm{PSS})} = \left(\sum_{i=1}^{N} w_i \right)^{-1/2}, \qquad \text{(A6)}$$

where $w_i = \sigma(\mathrm{PSS}_i)^{-2}$.

APPENDIX B. MAXIMUM-LIKELIHOOD ESTIMATION OF $\mathrm{RT_a}$ AND $\mathrm{RT_v}$

We describe a method for obtaining the ML estimate of the mode for a single set $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ of $n = 60$ RTs; this method was applied to obtain both $\mathrm{RT_a}$ and $\mathrm{RT_v}$. Having executed this procedure for both the visual and auditory tasks, the quantity $\mathrm{RT_d}$ was computed as $\mathrm{RT_d} = (\mathrm{RT_v} - \mathrm{RT_a})$.

The set of 60 RTs associated with each (visual and auditory) task was used to form a histogram of RTs. This histogram is an approximation to the probability density function (PDF) of the RTs, and has a characteristic positively skewed distribution. Accordingly, a log–normal function was used to model this PDF, using ML estimation. The result of the ML fitting procedure is an estimate of two parameters: the mean μ and the standard deviation σ of the observer's log–normal PDF. The RT most likely to be elicited by the (sound or light) stimulus is given by the mode of the fitted log–normal PDF.

If RT values y are distributed according to a log–normal distribution $f(\mu, \sigma, y)$ with mean μ and standard deviation σ then the likelihood function $L(\mu, \sigma)$ is

$$L(\mu, \sigma) = \prod_{i=1}^{n} f(\mu, \sigma, y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y_i} \exp\left(-(\mu - \log y_i)^2 / 2\sigma^2 \right). \qquad \text{(B1)}$$

The ML estimates $(\hat{\mu}, \hat{\sigma})$ of (μ, σ) were obtained by minimizing the function $-L$ with the MatLab simplex method *fmins*. Having obtained the maximum-likelihood estimate of μ and σ, the mode of the fitted PDF was computed by finding the value of y (i.e. RT) such that $\mathrm{d}f(\hat{\mu}, \hat{\sigma}, y)/\mathrm{d}y = 0$. The value of $\mathrm{RT_d}$ is the difference between the estimated modes associated with the distributions for auditory and visual RTs.

APPENDIX C. EVALUATING $\mathrm{PSS_N} - \mathrm{PSS_F}$

In experiment 2, the difference between the two PSS values of each observer obtained in the near and far conditions ($\mathrm{PSS_N}$ and $\mathrm{PSS_F}$, respectively) was evaluated as follows. Each PSS value has an associated ML estimate of its standard deviation $\sigma(\mathrm{PSS})$, which can be used to compare $\mathrm{PSS_N}$ and $\mathrm{PSS_F}$ for each observer. This is because each PSS is a ML estimate, and is therefore approximately normally distributed for the large sample sizes (1100 trials) used here (Cowan 1998). The significance of the difference $(\mathrm{PSS_N} - \mathrm{PSS_F})$ can be evaluated as a z-score, $z = (\mathrm{PSS_N} - \mathrm{PSS_F})/\sqrt{\hat{\sigma}_N^2 + \hat{\sigma}_F^2}$. Each z-score can then be associated with a significance value p using a simple one- or two-tailed z-test.

REFERENCES

Andreassi, J. & Greco, J. 1975 Effects of bisensory stimulation on reaction time and the evoked cortical potential. *Physiol. Psychol.* **3**, 189–194.

Cowan, G. 1998 *Statistical data analysis*. Oxford: Clarendon Press.

Hershenson, M. 1962 Reaction time as a measure of intersensory facilitation. *J. Exp. Psychol.* **63**, 289–293.

Massaro, D. & Kahn, B. 1973 Effects of central processing on auditory recognition. *J. Exp. Psychol.* **97**, 51–58.

Moutoussis, K. & Zeki, S. 1997*a* A direct demonstration of perceptual asynchrony in vision. *Proc. R. Soc. Lond.* B **264**, 393–399.

Moutoussis, K. & Zeki, S. 1997*b* Functional segregation and temporal hierarchy of the visual perceptive systems. *Proc. R. Soc. Lond.* B **264**, 1407–1414.

Sivia, D. 1996 *Data analysis: a Bayesian tutorial*. Oxford: Clarendon Press.

Stein, B. & Meredith, M. 1993 *The merging of the senses*, p. 136. Cambridge, MA: MIT Press.

Woodworth, R. & Schlosberg, H. 1954 *Experimental psychology*, 3rd edn. London: Methuen.

Zeki, S. & Bartels, A. 1998 The asynchrony of consciousness. *Proc. R. Soc. Lond.* B **265** 1583–1585.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.