Available online at www.sciencedirect.com
SCIENCE DIRECT
Cognitive Brain Research 25 (2005) 499 – 507
COGNITIVE BRAIN RESEARCH
www.elsevier.com/locate/cogbrainres

Research Report

Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration

Jordi Navarra[a,b,*], Argiro Vatakis[b], Massimiliano Zampini[b], Salvador Soto-Faraco[a], William Humphreys[b], Charles Spence[b]

[a]Grup de Recerca Neurociència Cognitiva, Parc Científic de Barcelona, Spain
[b]Department of Experimental Psychology, University of Oxford, UK

Accepted 31 July 2005
Available online 31 August 2005

Abstract

We examined whether monitoring asynchronous audiovisual speech induces a general temporal recalibration of auditory and visual sensory processing. Participants monitored a videotape featuring a speaker pronouncing a list of words (Experiments 1 and 3) or a hand playing a musical pattern on a piano (Experiment 2). The auditory and visual channels were either presented in synchrony, or else asynchronously (with the visual signal leading the auditory signal by 300 ms; Experiments 1 and 2). While performing the monitoring task, participants were asked to judge the temporal order of pairs of auditory (white noise bursts) and visual stimuli (flashes) that were presented at varying stimulus onset asynchronies (SOAs) during the session. The results showed that, while monitoring desynchronized speech or music, participants required a longer interval between the auditory and visual stimuli in order to perceive their temporal order correctly, suggesting a widening of the temporal window for audiovisual integration. The fact that no such recalibration occurred when we used a longer asynchrony (1000 ms) that exceeded the temporal window for audiovisual integration (Experiment 3) supports this conclusion.

Theme: Neural basis of behavior
Topic: Cognition

Keywords: Multisensory integration; Asynchrony; Temporal recalibration; Speech; Music; Temporal order judgment

1. Introduction

One of the best examples of human multisensory integration is provided by audiovisual speech perception (see [6]). The integration of audiovisual speech information occurs automatically whenever both acoustic and visual (lip movement) information are available simultaneously, even if this results in an illusory percept. This is perhaps best illustrated by the so-called McGurk effect, whereby the observer experiences hearing /da/ when presented with the sound /ba/ while viewing the lip movements associated with /ga/ (e.g., [22]; see also [33]).

Temporal coincidence has been identified as one of the most important factors determining whether or not multisensory integration takes place (see [6,8,31], for reviews). For instance, the audiovisual integration of speech breaks down if the asynchrony between the visual lip movements and the auditory speech sounds becomes too great (e.g., [9,13,19,20,24]). However, strict temporal overlap is not necessary, as the perceptual system can accommodate some degree of asynchrony, especially between correlated multisensory inputs (i.e., the lips match the sounds). This supports the idea that there is a temporal window within which multisensory integration can take place (e.g., [34,40]). The McGurk illusion, for

* Corresponding author. Departament de Psicologia Bàsica, Universitat de Barcelona, Pg. Vall d'Hebron, 171, 08035 Barcelona, Spain. Fax: +34 93 402 13 63.
E-mail address: jordi.navarra@ub.edu (J. Navarra).

0926-6410/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogbrainres.2005.07.009

example, persists even when the visual information leads (by up to 240 ms), or lags (by up to 60 ms) the auditory input [9,24]. Similarly, modern technology can also lead to asynchronous audiovisual stimulus presentation as, for example, with satellite TV broadcasts, in which there is often a lag between the auditory and visual signals (cf. [18,29,30]). The ability of the human perceptual system to reconcile small temporal asynchronies suggests a certain flexibility in the underlying mechanisms of multisensory integration.

Here, we investigated the nature of this temporal flexibility by addressing whether exposure to a continuous stream of complex audiovisual stimuli (such as speech or a recording of a hand playing a piano) presented asynchronously can induce a general temporal recalibration between audition and vision. Across three experiments, we analyzed the effects of monitoring asynchronous speech or music on performance in a temporal order judgment (TOJ) task where participants had to judge which of two events, a light flash and a burst of white noise, had been presented first (see [16]).

Several recent studies provide evidence for a temporal equivalent of the well-known spatial ventriloquism effect. In its spatial version, the location of a sound source is illusorily misplaced toward the position of a concurrent visual stimulus (e.g., [17]; see [3] for a recent review). In a recent demonstration of the existence of the temporal analogue of the ventriloquism effect, Morein-Zamir, Soto-Faraco, and Kingstone [23] reported that the perceived onset time of a light can be attracted temporally toward the onset time of a sound that is presented slightly later (see also [1,2,10,28,41]). Moreover, recent studies have also shown that it is possible to induce temporal recalibration after-effects by exposing the observer to a continuous stream of desynchronized audiovisual stimuli (e.g., tones and lights) [12,42]. In the present study, we investigated whether it is possible to demonstrate temporal recalibration using an online adaptation method during exposure to more complex and ecologically valid stimuli, such as a face talking or a hand playing notes on a piano. We measured the transfer of any temporal recalibration effect caused by exposure to desynchronized complex stimuli (speech or music) to the perception of a different kind of stimuli, consisting of a simple flash of light and a burst of white noise (cf. [12]).

We used a videotaped recording of a speaker pronouncing a list of words or a hand playing a piano. In half of the experimental blocks, the auditory signal was delayed relative to the visual signal, whereas in the remainder of the blocks, the auditory and visual stimuli were presented in synchrony (see Fig. 1). While monitoring the speech (or music) stream for targets (male first names or a break in the musical pattern, respectively), participants were asked to judge the order in which a pair of stimuli (a burst of white noise and a briefly flashing LED) was presented (i.e., they performed an audiovisual TOJ task).[1] We predicted that if any adaptation to the asynchronous complex audiovisual stream were to take place, then it might be possible to find a general temporal recalibration of audiovisual processing (cf. [12]), thus leading to an influence in TOJ performance for light flashes and noise bursts.

In particular, the occurrence of adaptation might affect the just noticeable difference (JND), and/or the point of subjective simultaneity (PSS) in the TOJ task. The JND refers the smallest temporal interval between two stimuli needed for participants to be able to judge correctly which one was presented first on 75% of trials. Our prediction was that monitoring the complex asynchronous stimuli (either audiovisual speech or a hand playing a piano) might modify the perceiver's 'online' temporal resolution (i.e., a widening of the audiovisual temporal window for integration), thus participants would require a longer interval in the TOJ task to decide whether the light or the sound came first (i.e., the JND should be larger). The PSS indicates how much time one stimulus has to lead the other in order for the two to be judged as occurring simultaneously (i.e., the average SOA at which participants make each response equally often), and is sensitive to differences in neural processing latencies between auditory and visual stimuli [34]. In the present study, any shift in the PSS would presumably reflect a realignment in the temporal processing of one sensory modality relative to the other, consequent on the brain's ability to adapt to audiovisual asynchrony.

2. Experiment 1

2.1. Materials and methods

2.1.1. Participants

Twelve participants took part in this experiment. All were naive as to the purpose of the experiment and all reported normal hearing and normal or corrected-to-normal vision. All of the participants gave their informed consent prior taking part in the study, and the majority received a £5 (UK Sterling) gift voucher in return for their participation. All of the experiments reported in this study were non-invasive, were conducted in accordance with the Declaration of Helsinki, and had ethical approval from the Department of Experimental Psychology, University of Oxford, UK.

2.1.2. Apparatus and materials

We used an 18-min videotaped recording of a male speaker (consisting of a close-up of the mouth area, from the

[1] Given that temporal recalibration effects are smaller when the adaptor stimuli and test stimuli (used to measure temporal after-effects) are different (see [12]), we decided to present the speech monitoring and TOJ tasks simultaneously (i.e., rather than sequentially as in many studies of post-exposure after-effects; cf. [12,42]).

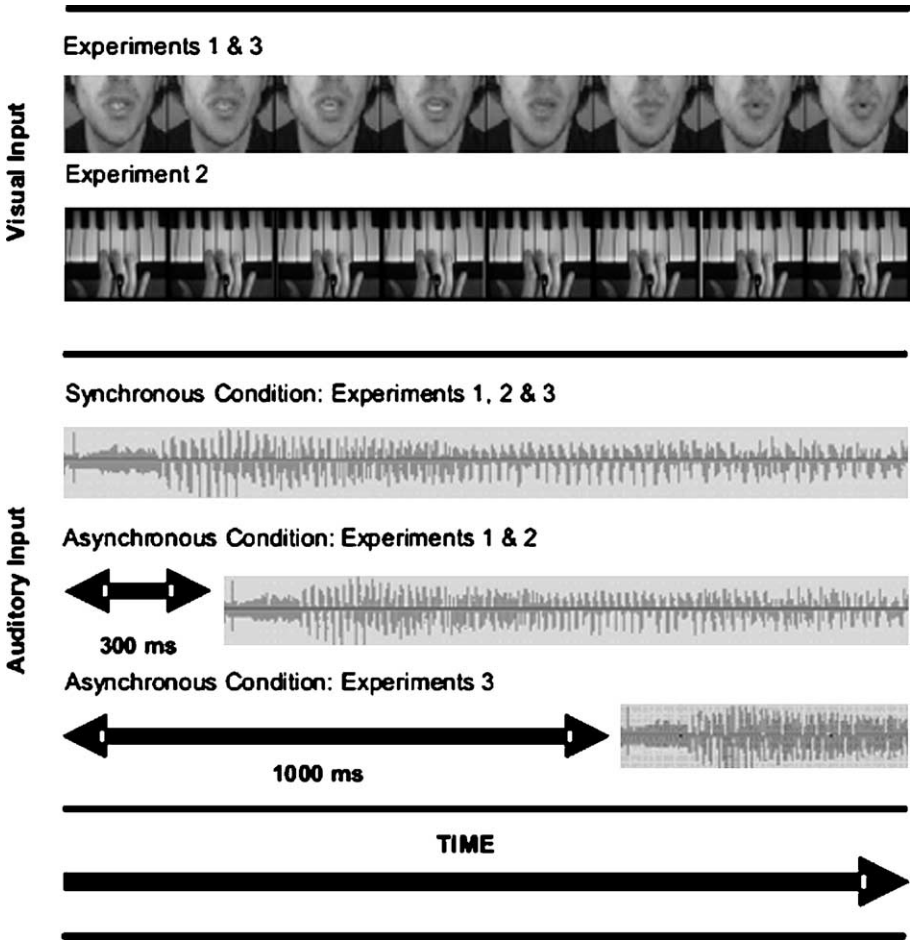

Fig. 1. In the synchronous condition (Experiments 1–3), the auditory and visual signals were presented in synchrony. In the asynchronous condition, the visual signal was presented 300 ms before the sound in Experiments 1 and 2, and 1000 ms before the sound in Experiment 3. In Experiments 1 and 3, the stimulus consisted of a continuous audiovisual speech stream, while in Experiment 2, it consisted of a recording of a hand playing a simple melody on three adjacent piano keys. As in Experiment 1, the image of the hand was presented 300 ms before the sound in the asynchronous condition and simultaneously in the synchronous condition.

upper part of the nose to approximately 3 cm below the chin) reading a list of 1000 words at a rate of approximately 60 words/min. One hundred of these 1000 words were target words (consisting of male first names: e.g., John, Peter, etc.) and were inserted pseudo-randomly into the list. The videotape was replayed on a videorecorder and presented on a 35-cm-wide television monitor located in a dark sound-attenuated booth. The speech signal from the videotape was presented through television speakers (centrally located above and below the monitor, 13 cm from the center of the screen) at approximately 78 dB(A) (as measured from participant's head position). The videotape was rewound during 3 pauses inserted between blocks. A video delay box (Pixel Instruments Corporation, AD 2100; Los Gatos, California) was interfaced with the video player in order to manipulate the asynchrony between the auditory and visual channels. White noise was presented continuously at 75 dB(A) from two additional loudspeakers (100% correlated), located 43 cm to either side of the television, throughout the experimental blocks. The white noise was introduced in order to reduce the intelligibility of the auditory input and, consequently, to enhance the participant's reliance on visual lip movements [14,20]. Note, however, that the speech signal, presented at a speech-to-noise ratio of +3 dB, was intelligible (see [37]).

A red light-emitting diode (LED) (luminance of 64.3 cd/m^2) was attached to the front of the television screen to present the visual stimuli in the TOJ task. In Experiment 1, the LED was located 2 cm below the speaker's lower lip, and did not cover any relevant part of the image needed for visual speech perception. Two identical loudspeakers (Audax, VE100AO) were placed 5 cm to either side of the television monitor and were used to present the auditory stimuli for the TOJ task, which consisted of 9-ms bursts of white noise [82 dB(A), as measured from the participant's head position]. These auditory stimuli were perfectly audible during the experiment. The loudspeakers (which delivered a coherent output) produced a sound that was perceived centrally by the participants, at approximately the same location as the LED, thus reducing the possibility of participants using redundant spatial cues to facilitate their TOJ performance (see [35,43,44]). A hand-held response pad with two response

keys placed horizontally (i.e., one to the left of the other) was used to record the participant's responses in the TOJ task. The left key was used to respond when the white noise burst was presented first, while the right key was used to respond when the onset of the LED occurred first. The auxiliary LEDs located at the bottom right hand corner of the television screen were used to inform participants that their manual response had been registered. The experimental protocol was controlled by a computer program written in Turbo Pascal 6.0. The participant's eye position was monitored via a CCTV infrared camera (Panasonic, BP310) attached to the top of the television screen, and connected to a monitor outside the booth. The experimenter checked the monitor periodically throughout the session to ensure that the participants were directing their gaze to the screen as instructed.

2.1.3. Procedure

The experiment was conducted in a completely dark sound-attenuated booth. The participants sat in front of the monitor (at a distance of approximately 125 cm) and were instructed to fixate the screen throughout each block of experimental trials. The experimental session consisted of eight blocks (each lasting approximately 5 min) in which participants monitored the spoken list for male first names which they were instructed to count. At the same time (i.e., online), the participants performed 80 TOJ trials. The participants first completed three practice blocks. First, speech monitoring only, then a second practice block of 15 TOJ trials in which the SOAs were twice as large as those used in the actual experiment (in order to familiarize participants with the task), and finally a third practice block of combined video monitoring with 15 TOJ trials, again with the SOAs doubled. The LED placed on the screen was illuminated at the beginning of each block of trials and acted as a warning signal. In the TOJ task, participants had to press a key with their (left or right) thumb in the TOJ task according to which stimulus (sound or light) came first. One stimulus was presented 750 ms after the offset of the warning light. The other stimulus was presented at one of ten different SOAs (−500 ms, −350 ms, −250 ms, −150 ms, −50 ms, 50 ms, 150 ms, 250 ms, 350 ms, 500 ms;[2] negative SOAs indicate that the auditory stimulus was presented first, whereas positive values indicate that the visual stimulus was presented first) using the method of constant stimuli (e.g., [35]). The task was unspeeded, and participants were informed that they should respond when confident of their response (although within the 3500 ms allowed before the termination of the trial). If participants responded prior to the onset of the first TOJ stimulus, or failed to respond before the trial was terminated (less than 1% of trials), error feedback (consisting of the flickering of the fixation light for 1000 ms) was presented. Otherwise, the participant's response was indicated by the illumination of one of the feedback lights located at the bottom right hand corner of the television screen for 500 ms after their response was detected. The warning light was illuminated to indicate the start of the next trial 750 ms after the end of the preceding trial.

At the end of each block of trials, the participants were instructed to report the number of male first names that they had detected in the preceding block (the actual number varied between 15 and 25). If the number of first names reported by participants was more than 3 names over or below the number of names that were actually presented, then the experimenter verbally reminded the participant of the importance of performing the monitoring task accurately. In general, performance in this task was maintained within this limit of ±3 names.

The auditory speech signal was delayed by 300 ms (relative to the visual speech signal) in half (i.e., 4) of the experimental blocks, while the audiovisual speech signal was presented synchronously in the remainder of the blocks. The synchronous and asynchronous speech conditions were alternated on a block-by-block basis, with the starting block (synchronous vs. asynchronous) counterbalanced across participants.

2.2. Results

The proportion of 'light first' responses in the TOJ task was converted to their equivalent Z scores assuming a cumulative normal distribution (see [11]). The intermediate eight SOAs were used to calculate a best-fitting straight line for each participant and condition.[3] The slopes and intercepts from these best-fitting lines were used to calculate the JND (JND = 0.675/slope since ±0.675 point corresponds to the 75% and 25% points on the cumulative normal distribution) and the PSS (PSS = −slope/intercept) for each participant and synchrony condition (see [7]).

The JND and PSS data (see Table 1) were submitted to paired-samples t tests to compare TOJ performance during exposure to the synchronous and asynchronous conditions. The analysis of the JND data revealed a statistically significant difference [$t(11) = -2.61$, $P = .034$], with participants requiring a shorter time interval (i.e., a smaller JND) to correctly perceive the order of sound and light while monitoring synchronous speech (mean JND = 120 ms) than while monitoring asynchronous speech (mean JND = 145 ms; see Table 1). Thus, our data suggest that the temporal window for audiovisual integration became larger as a result of monitoring the asynchronous speech stream. This finding supports the central prediction of our study,

[2] We used a broader range of SOAs than that used in many previous TOJ studies (e.g., [32,35]) given that pilot testing using SOAs of ±50 ms, ±100 ms, ±150 ms, ±250 ms, ±350 ms yielded error rates in excess of 30%.

[3] The ±500-ms points were excluded from this computation because most participants performed nearly perfectly at this interval; hence, no additional variance was accounted for by these points, and their inclusion in the data analysis would actually have resulted in an artifactual reduction in the slope (see [35] on this point).

Table 1
Participant's performance on the TOJ task in Experiments 1–3

Experiment	Condition	JND (mean)	JND (SE)	PSS (mean)	PSS (SE)
Experiment 1 (speech)	Synchronous	120	8.03	−22	17.47
	Asynchronous (300 ms)	145	11.31	−15	22.51
Experiment 2 (piano)	Synchronous	116	14.57	−26	19.16
	Asynchronous (300 ms)	125	15.77	−23	28.74
Experiment 3 (speech)	Synchronous	137	11.06	−7	24.27
	Asynchronous (1000 ms)	129	12.98	−10	25.85

Note. The mean JND and PSS values (in milliseconds), and their standard errors in the TOJ were obtained while participants monitored either synchronous or asynchronous speech (Experiments 1 and 3), or else a hand playing a piano (Experiment 2). Significant differences in the JND between the synchronous and the asynchronous conditions were demonstrated when the visual stream led the auditory stream by 300 ms (Experiments 1 and 2), but not when it led by 1000 ms (i.e., when the stimuli fell outside the window of audiovisual integration). There was no effect of the synchrony factor on the PSS in any of the experiments.

namely, that adaptation to desynchronized audiovisual complex stimuli (i.e., speech) can influence the temporal resolution in the perception of simple stimuli (i.e., light flashes and noise bursts).

A paired-samples *t* test revealed no significant differences between the mean PSS reported when participants monitored synchronous versus asynchronous speech [$t(11) = -.741$, $P = .474$], suggesting that monitoring asynchronous speech did not result in a temporal realignment (such as a systematic delay) of the information from one sensory modality relative to the other. In line with several previous studies (e.g., [12,38]), our data indicate that, in general, the auditory stimuli had to lead the visual stimuli by approximately 20 ms for the PSS to be achieved in the TOJ task (though see [43,44] for a different pattern of results).

Thus far, our results suggest that the monitoring of an asynchronous audiovisual speech stream can lead to a temporal recalibration that can affect the perception of simple non-speech stimuli. Given the debate over the question of whether speech may represent a special case of multisensory integration (see [21,39]), we attempted in our second experiment to generalize this adaptation effect to another type of complex stimulus (in particular, music) that can also be, in some circumstances, experienced audiovisually. The similarities and differences, in terms of brain processing, between music and speech have been studied intensively (see [4,25,27,45]). Some cues (such as pitch and timing patterns) present both in music and prosodic aspects of speech are thought to be processed in the same brain areas (i.e., the right parietal cortex and the right frontal lobe; e.g., [25,27]). Interestingly, the planum temporale, a brain area associated with lipreading, has also been shown to be activated in trained piano players when watching melodies being played on a piano in the absence of any piano-related sound [15].

3. Experiment 2

3.1. Materials and methods

3.1.1. Participants

Six new participants took part in this experiment. As in Experiment 1, all of the participants were naive as to the purpose of the study, and all reported normal hearing and normal or corrected-to-normal vision.

3.1.2. Apparatus and materials

We used a 35-min videotape of a hand repeating a simple musical pattern in which 3 fingers of one hand pressed 3 contiguous keys on the piano keyboard at a constant rhythm (approximately 65 times/min). The videotape was rewound during a pause inserted between blocks. Approximately 150 targets (changes to this pattern) were inserted pseudo-randomly into the sequence. Targets consisted of all three fingers forcefully pressing down on the three keys at the same time. Apart from these changes, the apparatus and materials were exactly the same as those used in the previous experiment.

3.1.3. Procedure

The procedure was identical to that used in Experiment 1 with the following exceptions: The video monitoring task now consisted of counting changes in a musical pattern executed by a hand on the keys of a piano, and participants did not receive any feedback concerning their performance. As in Experiment 1, participants also performed an online TOJ task regarding the LED and the burst of white noise. The LED did not cover any of the relevant parts of the image, being placed on the lower central part of the screen. The instructions explicitly asked participants to focus their gaze on the fingers playing the piano, rather than on the LED.

3.2. Results

The JND values obtained in the synchronous and asynchronous conditions of Experiment 2 (see Table 1) were submitted to a paired-samples *t* test, resulting in a statistically significant difference [$t(5) = -2.984$, $P = .031$]. When participants monitored the synchronous stimulus their temporal resolution in the audiovisual TOJ task was better (mean JND = 116 ms) than when they monitored the asynchronous stream (mean JND = 125 ms), just as in Experiment 1. This finding is in line with the temporal recalibration hypothesis whereby the temporal window of integration for audiovisual inputs becomes wider (cf. [36]). The PSS data in Experiment 2 indicated that simultaneous perception of the audiovisual stimulus pairs required the auditory stimulus to lead by 25 ms. A paired-samples *t* test comparing the PSS while monitoring the asynchronous piano playing (mean PSS = −26 ms) to the synchronous condition (mean PSS = −23

ms) again revealed no significant differences [$t(5) = -.209$, $P = .842$].

The results of Experiment 2 generalize the finding of Experiment 1, in support of the idea that exposure to either class of complex asynchronous audiovisual stimuli can affect the window of perceived simultaneity for any pair of (unrelated) audiovisual events. These results also suggest that even when the monitoring task could be performed unimodally (e.g., by concentrating visually on when the hand broke the pattern), participants perceived both visual and auditory aspects of the stimulus and integrated them into a single percept (as indicated by the temporal recalibration effect). However, another explanation of the results should also be considered. In particular, one could argue that participants in Experiments 1 and 2 may have been less accurate in the TOJ task while monitoring the asynchronous stream (and, consequently, needed more time between the light and the sound to judge which one came first) simply because they were being exposed to an unnatural stimulus that was perhaps distracting and/or attention-grabbing, thereby producing a general decrease in their ability to concentrate on the TOJ task. We conducted a final experiment to address this possibility.

The goal of Experiment 3 was to study whether the temporal recalibration reported in Experiments 1 and 2 would still occur if the temporal interval between the auditory and visual inputs clearly exceeded the temporal interval for successful audiovisual integration. In contrast to Experiments 1 and 2, in which the asynchrony was at the limit of the temporal window for audiovisual integration (see [10,24,40]), the visual input was presented 1000 ms before the auditory input in the asynchronous condition in Experiment 3 (see Fig. 1) to ensure that participants were not able to match the visual and the auditory word streams (namely, to recalibrate in time). Yet, as in Experiment 1, participants were still clearly exposed to audiovisually non-coincident stimuli (thus, equally unnatural and putatively attention-grabbing). According to the temporal recalibration hypothesis, when the asynchrony exceeds the window for successful audiovisual integration [24], no recalibration, and therefore no change in TOJ performance should occur. If, on the other hand, the effects reported in Experiments 1 and 2 simply reflect the increased difficulty of monitoring the somewhat unnatural (i.e., mismatching) signal, then the effect found on JND should, once again, be observed.

4. Experiment 3

4.1. Materials and methods

4.1.1. Participants

Eleven new participants with normal hearing and normal or corrected-to-normal vision participated in this experiment.

4.1.2. Materials, apparatus, and procedure

These were the same as in Experiment 1, with the following exception: In the asynchronous condition, the visual information was presented 1000 ms before the auditory information (instead of 300 ms before as in Experiments 1 and 2). In the asynchronous condition, the heard and visually mouthed words were completely mismatched in terms of content (sound vs. lip movement did not correspond). Moreover, they did not perfectly match in terms of onset/offset time, as the length of the words did not coincide. This manipulation should make the asynchronous condition of this experiment look at least as unnatural as the asynchronous condition of Experiment 2, but prevent any possible integration from occurring. As in Experiment 2, participants did not receive feedback regarding their performance (i.e., which response they had made).

4.2. Results

As in Experiments 1 and 2, the JND and PSS were obtained for each participant in each condition. One participant performed well below average (i.e., near to chance levels even at the long SOAs; e.g., 60% correct at the 250-ms SOA), and his data were therefore removed from the analyses. A paired-samples t test did not reveal any significant difference in JND between the synchronous (mean JND = 137 ms) and asynchronous (mean JND = 129 ms) conditions [$t(9) = 1.012$, $P = .338$] (note that the trend here, toward worse performance in the synchronous condition was actually the opposite of that seen in the two previous experiments). A second paired t test on the PSS data also failed to reveal any significant differences between the synchronous and asynchronous conditions [$t(9) = .163$, $P = .874$].

In order to compare the effects of monitoring small (300 ms) versus large audiovisual asynchronies (1000 ms) on TOJ performance, a mixed design analysis of variance (ANOVA) was performed on the JND data including one between-participants factor, Experiment (1 vs. 3), and one within-participants factor, Synchrony (Synchronous vs. Asynchronous). No significant effects of synchrony [$F(1,21) = 1.793$, $P = .196$] or experiment ($F < 1$) were found but, critically, the analysis revealed a significant interaction between the two factors [$F(1,21) = 6.068$, $P = .023$]. This interaction indicates that in Experiment 1, monitoring speech with an asynchrony of 300 ms induced larger JNDs than the synchronous speech condition, whereas participants in Experiment 3, who monitored the speech with a large asynchrony (1000 ms), had equivalent JNDs in both synchronous and asynchronous conditions (see Table 1). This pattern of data supports the view that the effects reported in Experiments 1 and 2 resulted from a temporal recalibration effect, and not from some kind of distraction effect caused by mere exposure to

uncorrelated (i.e., as opposed to correlated) audiovisual input.[4]

5. Discussion

The results of the present study suggest that monitoring complex asynchronous audiovisual signals (such as speech or the playing of a musical instrument) results in a significant modulation in the temporal processing of audiovisual stimuli. Our findings are consistent with the idea that the temporal window for audiovisual integration widens as a consequence of adaptation to asynchronous signals. However, the results of Experiment 3 reveal that this temporal widening does not always occur. It seems that the audiovisual asynchrony must remain within certain temporal limits which, in our case, coincided with the so-called temporal window of integration of speech. Several studies have shown that effective audiovisual speech integration breaks down when the asynchrony is larger than about 300 ms; e.g., [9,13]. The present recalibration effect supports recent claims about the existence of a synchrony detection mechanism [12], which would be critical to the binding across sensory modalities (e.g., [5]). It will be interesting to investigate the particular neural mechanisms underlying the *phasic* changes involved this temporal recalibration effect across sensory modalities (cf. [5,26]).

The fact that exposure to one kind of complex stimulus (i.e., speech) produces recalibration effects that can affect one's perception of other kinds of stimuli (i.e., light flashes and noise bursts) is consistent with a general (rather than stimulus-specific) perspective on audiovisual integration (see [39]). An additional question for future research, in order to clarify this matter further, will be to investigate systematically whether the exposure to asynchrony in any kind of information (e.g., complex or simple) can affect the concomitant perception of any other kind of complex or simple multisensory stimuli. The occurrence of such an effect would provide more robust evidence concerning the existence of general, rather than specific, processes for the temporal aspects of audiovisual integration.

Fujisaki et al. [12] (see also Vroomen et al. [42]) recently reported that repeated exposure to temporally misaligned audiovisual stimuli (such as brief tones and light flashes) can induce a transient shift in the PSS between the auditory and visual modalities. In Fujisaki et al.'s study [12], the temporal recalibration was investigated by measuring after-effects (i.e., participant's performance was measured after exposure to asynchronous stimuli) using both direct and indirect measures of temporal perception (e.g., TOJ tasks and judging whether two balls streamed through one another, or else bounced; cf. [31]). Interestingly, Fujisaki et al.'s findings were also demonstrated when the type of stimuli used during adaptation was different to the test stimuli (showing that the effects were not stimulus-specific, as also seen in the present study). However, in this latter case, the recalibration was smaller than when the nature of the adaptation and the test stimuli was similar, perhaps explaining the small size of the effects found in the present study. In order to increase the opportunities of capturing 'cross-stimulus' temporal recalibration effects (that Fujisaki et al. found tend to be small; [12]), we used a different experimental paradigm where possible temporal recalibration effects were measured online (during exposure to the asynchronous stimulus). It remains an interesting question for future research to determine why we found that the monitoring of asynchronous audiovisual stimuli always led to an increase in the JND but no change in the PSS, whereas Fujisaki et al. [12] found the opposite pattern of results when using adaptation after-effects. An obvious difference between the two studies is the time at which the TOJ measures were taken (online effects vs. after-effects, respectively). This raises one intriguing but at the present time necessarily speculative possibility. Namely, that the mechanism by which temporal recalibration occurs may involve an initial widening of the temporal window for audiovisual integration (producing the increase in the JND) followed by a temporal realignment of the two modalities (indicated by a shift in the PSS), and

[4] The overall poor performance on the TOJ task in the present study, as reflected by the large JNDs, when compared with that reported in many previous studies (e.g., [32,35]), probably reflects the result of some overall dual-task cost and/or the presentation of complex (and perhaps attentionally demanding) stimuli such as speech or music, as well as the broad range of SOAs used (relative to previous studies). In a follow-up experiment, 15 participants performed the same TOJ task as that reported here under the same display conditions but without the instruction to monitor the speech stream for targets (i.e., under single task conditions). As in Experiment 1, the (now irrelevant) speech stream could either be presented in synchrony or with the auditory stream delayed by 300 ms. We found that JNDs (mean JND = 95 ms, in the synchronous speech condition, and 88 ms in the asynchronous speech condition) were significantly smaller than those reported in Experiment 1 [$F(1,26) = 5.971$, $P = .022$, for the synchronous condition, and $F(1,26) = 22.449$, $P = .0001$, for the asynchronous condition]. These results suggest that the overall large JNDs reported in Experiment 1 were at least in part a consequence of the attentional cost associated with participants having to perform two tasks at the same time (i.e., a dual task deficit). We also compared the JNDs obtained in our Experiment 1 (under conditions of synchronous monitoring) with the JNDs obtained in the first 80 trials (mean JND = 79 ms) of a similar previous study by Zampini et al. [43] (see Experiment 1; same position) where no extra audiovisual streams were present (i.e., we matched the number of trials analyzed for the two studies to compensate for any practice effects that may have been present). We found a significant difference between the JNDs reported in Experiment 1 and those reported in Zampini et al.'s study [$F(1,20) = 7.489$, $P = .013$] (thus again reflecting the dual-task cost). However, we did not find a statistical difference between the JNDs in our follow-up experiment (no speech monitoring) and Zampini et al.'s results [$F(1,23) = 1.309$, $P = .265$]. Taken together, this pattern of results clearly indicates that the larger JNDs reported in Experiment 1 were primarily caused by an attentional cost due to dual-task or speech monitoring. However, the critical point is that, despite the fact that the JNDs reported in the present study were quite large, we nevertheless still found (and replicated) a robust temporal recalibration effect.

perhaps ending up with a narrowing of the window for integration centered around this new PSS.

Acknowledgments

This study was supported by a Network Grant from the McDonnell-Pew Centre for Cognitive Neuroscience in Oxford to S.S.-F. and C.S., and a grant from the James S. McDonnell Foundation JSMF-20002079.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cogbrainres.2005.07.009.

References

[1] G. Aschersleben, P. Bertelson, Temporal ventriloquism: crossmodal interaction on the time dimension. 2. Evidence from sensorimotor synchronization, Int. J. Psychophysiol. 50 (2003) 157–163.
[2] P. Bertelson, G. Aschersleben, Temporal ventriloquism: crossmodal interaction on the time dimension: 1. Evidence from auditory–visual temporal order judgment, Int. J. Psychophysiol. 50 (2003) 147–155.
[3] P. Bertelson, B. de Gelder, The psychology of multimodal perception, in: C. Spence, J. Driver (Eds.), Crossmodal Space and Crossmodal Attention, Oxford Univ. Press, Oxford, 2004, pp. 141–177.
[4] D. Bolinger, Intonation and its Uses: Melody in Grammar and Discourse, Stanford Univ. Press, Stanford; 1989.
[5] K.O. Bushara, J. Grafman, M. Hallett, Neural correlates of auditory–visual stimulus onset asynchrony detection, J. Neurosci. 21 (2001) 300–304.
[6] G. Calvert, C. Spence, B.E. Stein (Eds.), The Handbook of Multisensory Processing, MIT Press, Cambridge; 2004.
[7] S. Coren, L.M. Ward, J.T. Enns, Sensation and Perception, 6th ed., Harcourt Brace, Fort Worth; 2004.
[8] B. de Gelder, P. Bertelson, Multisensory integration, perception and ecological validity, Trends Cogn. Sci. 7 (2003) 460–467.
[9] N.F. Dixon, L. Spitz, The detection of auditory visual desynchrony, Perception 9 (1980) 719–721.
[10] R. Fendrich, P.M. Corballis, The temporal cross-capture of audition and vision, Percept. Psychophys. 63 (2001) 719–725.
[11] D.J. Finney, Probit Analysis: Statistical Treatment of the Sigmoid Curve, Cambridge Univ. Press, London; 1964.
[12] W. Fujisaki, S. Shimojo, M. Kashino, S. Nishida, Recalibration of audiovisual simultaneity, Nat. Neurosci. 7 (2004) 773–778.
[13] K.W. Grant, S. Greenberg, Speech intelligibility derived from asynchronous processing of auditory–visual information, Proceedings of the AVSP 2001 International Conference of Auditory–Visual Speech Processing, Scheelsminde, Denmark, 2001, pp. 132–137.
[14] K.W. Grant, P.F. Seitz, The use of visible speech cues for improving auditory detection of spoken sentences, J. Acoust. Soc. Am. 108 (2000) 1197–1208.
[15] T. Hasegawa, K.I. Matsuki, T. Ueno, Y. Maeda, Y. Matsue, Y. Konishi, N. Sadato, Learned audio–visual cross-modal associations in observed piano playing activate the left planum temporale: an fMRI study, Cognit. Brain Res. 20 (2004) 510–518.
[16] I.J. Hirsh, C.E. Sherrick Jr., Perceived order in different sense modalities, J. Exp. Psychol. 62 (1961) 423–432.
[17] I.P. Howard, W.B. Templeton, Human Spatial Orientation, John Wiley, London; 1966.
[18] ITU-T, Television and sound transmission: tolerances for transmission time differences between the vision and sound components of a television signal. International Telecommunication Union, Telecommunication standardization sector of ITU, Recommendation J.100, CMTT 717 in CCIR Recommendations, 12, Düsseldorf; 1990.
[19] A. Kopinska, L.R. Harris, Simultaneity constancy, Perception 33 (2004) 1049–1060.
[20] E. Macaluso, N. George, R. Dolan, C. Spence, J. Driver, Spatiotemporal contributions to audiovisual speech perception: a PET study, Neuroimage 21 (2004) 725–732.
[21] D.W. Massaro, M.M. Cohen, P.M. Smeele, Perception of asynchronous and conflicting visual and auditory speech, J. Acoust. Soc. Am. 100 (1996) 1777–1786.
[22] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 265 (1976) 746–748.
[23] S. Morein-Zamir, S. Soto-Faraco, A.F. Kingstone, Auditory capture of vision: examining temporal ventriloquism, Cognit. Brain Res. 17 (2003) 154–163.
[24] K.G. Munhall, P. Gribble, L. Sacco, M. Ward, Temporal constraints on the McGurk effect, Percept. Psychophys. 58 (1996) 351–362.
[25] K.G. Nicholson, S. Baum, A. Kilgour, C.K. Koh, K.G. Munhall, L.L. Cuddy, Impaired processing of prosodic and musical patterns after right hemisphere damage, Brain Cogn. 52 (2003) 382–389.
[26] T. Noesselt, C. Tempelmann, H.-J. Heinze, J. Driver, Neural correlates of audiovisual synchrony and asynchrony in the human brain, Poster Presented at the Meeting of the 5th International Multisensory Research Forum, Barcelona, Spain; 2004.
[27] A.D. Patel, I. Peretz, M. Tramo, R. Labreque, Processing prosodia and musical patterns: a neuropsychological investigation, Brain Lang. 61 (1998) 123–144.
[28] G.H. Recanzone, Auditory influences on visual temporal rate perception, J. Neurophysiol. 89 (2003) 1078–1093.
[29] B. Reeves, D. Voelker, Effects of audio–video asynchrony on viewer's memory, evaluation of content and detection ability, Research Report Prepared for Pixel Instruments, Los Gatos, California, USA; 1993.
[30] S. Rihs, The influence of audio on perceived picture quality and subjective audiovisual delay tolerance, in: R. Hamberg, H. de Ridder (Eds.), Proceedings of the MOSAIC Workshop: Advanced Methods for the Evaluation of Television Picture Quality, 1995, pp. 133–137 (Eindhoven).
[31] R. Sekuler, A.B. Sekuler, R. Lau, Sound alters visual motion perception, Nature 385 (1997) 308.
[32] D. Shore, C. Spence, M. Klein, Visual prior entry, Psychol. Sci. 12 (2001) 205–212.
[33] S. Soto-Faraco, J. Navarra, A. Alsius, Assessing the automaticity in audiovisual speech integration: evidence from the speeded classification task, Cognition 92 (2004) B13–B23.
[34] C. Spence, S. Squire, Multisensory integration: maintaining the perception of synchrony, Curr. Biol. 13 (2003) 519–521.
[35] C. Spence, D.I. Shore, R.M. Klein, Multisensory prior entry, J. Exp. Psychol. Gen. 130 (2001) 799–832.
[36] Y. Sugita, Y. Suzuki, Implicit estimation of sound-arrival time, Nature 421 (2003) 911.
[37] W.H. Sumby, I. Pollack, Visual contribution to speech intelligibility in noise, J. Acoust. Soc. Am. 26 (1954) 212–215.
[38] G. Teatini, M. Farnè, F. Verzella, P. Berruecos, Perception of temporal order: visual and auditory stimuli, G. Ital. Psicol. 3 (1976) 157–164.
[39] J. Tuomainen, T.S. Andersen, K. Tiippana, M. Sams, Audio–visual speech is special, Cognition 96 (2005) B13–B22.
[40] V. Van Wassenhove, K.W. Grant, D. Poeppel, Temporal integration in the McGurk effect, Poster presented at the annual meeting of the Cognitive Neuroscience Society (2002), San Diego, CA.
[41] J. Vroomen, B. de Gelder, Temporal ventriloquism: sound modulates the flash-lag effect, J. Exp. Psychol. Hum. Percept. Perform. 30 (2004) 513–518.

[42] J. Vroomen, M. Keetels, B. de Gelder, P. Bertelson, Recalibration of temporal order perception by exposure to audio–visual asynchrony, Cognit. Brain Res. 22 (2004) 32–35.
[43] M. Zampini, D.I. Shore, C. Spence, Audiovisual temporal order judgments, Exp. Brain Res. 152 (2003) 198–210.
[44] M. Zampini, D.I. Shore, C. Spence, Multisensory temporal order judgments: the role of hemispheric redundancy, Int. J. Psychophysiol. 50 (2003) 165–180.
[45] R. Zatorre, P. Belin, V.B. Penhune, Structure and function of auditory cortex: music and speech, Trends Cogn. Sci. 6 (2002) 37–46.