Some methodological aspects for measuring asynchrony detection in audio-visual stimuli

Pacs Reference: 43.66.Mk, 43.66.Lj

Van de Par, Steven[1]; Kohlrausch, Armin[1,2]; and Juola, James F.[3]

1) Philips Research laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, the Netherlands,
Tel: +31 40 2745418, Fax: +31 40 2744675, email: Steven.van.de.Par@philips.com
2) Technische Universiteit Eindhoven, Den Dolech 2, 5600 MB Eindhoven, the Netherlands
3) Department of Psychology, University of Kansas, Lawrence KS 66045, USA

ABSTRACT

For audio-visual stimuli with a clear temporal structure, like impact events, a difference exists between physical and perceived synchrony. We compare the results of different methods to establish the point of subjective equality (PSE). These methods differ in the type of response categories subjects can use: 1) 3 categories: audio first, synchronous, video first, 2) 2 categories: synchronous, asynchronous, 3) 2 categories: audio first, video first. It appears that the estimates of the PSE obtained with methods 1 and 2 are rather robust and in agreement with each other. In contrast, method 3 allows for different decision strategies and results depend on which strategy the subject chooses.

INTRODUCTION

In daily life we seem to have little difficulty in integrating sensory information from various sensory modalities. For example, when a person is speaking to us, we can easily link the acoustic speech signal with the person that we see speaking. This ability to integrate multi-sensory information, however, should not be regarded as trivial, considering the difference in the structure and nature of the different sensory information streams. It is fair to assume that this integration is mediated by the underlying physical laws which lead to stimulus properties that are common across sensory modalities. In the example of the person that is speaking, the direction along which we see the person corresponds to the place where we localise the sound. In addition, the lip movements that we see correspond in some way with the temporal acoustical changes that occur in the speech. A third factor could be that we know that the characteristics of the particular voice match the person that we see. In this presentation we want to address the topic of auditory-visual time perception.

Various experiments have been reported that dealt with the sensitivity to timing in auditory-visual stimuli. Dixon and Spitz (1980) measured the point at which subjects where able to notice asynchronies between audio and video as the asynchrony gradually increased. The stimuli were recorded scenes of a human speaker and a hammer hitting a peg. They found thresholds of 188 ms for audio delays and 75 ms for video delays in the case of the hammer stimulus, but 258 ms and 131 ms for the speech stimulus. The centre-point between the audio delay and video delay thresholds can be defined as the point of subjective equivalence (PSE). Clearly the experiments of Dixon and Spitz reveal a PSE that is not equal to the point of objective equivalence (POE) where the stimuli are in physical synchrony. The PSE is 57 ms audio delay for the hammer stimulus and 64 ms audio delay for the speech stimulus.

More recently we measured thresholds for the detection of asynchrony using an adaptive staircase procedure (van de Par and Kohlrausch, 2000). On each trial, two stimuli were presented, one AV stimulus in physical synchrony and one stimulus with an AV asynchrony. Subjects had to indicate which interval contained the asynchronous stimulus. In this experiment, contrary to the experiments of Dixon and Spitz, feedback was provided to the subjects after each trial about the correctness of their reponse. The stimulus consisted of the same impact stimulus which will be described for the experiments in the current paper. We found considerably smaller thresholds as compared to those of Dixon and Spitz. They were about 30 ms on average for video delays, and about 85 ms for audio delays. The difference in our results is probably due to the specific measurement procedure that we used which was a discrimination task (including feedback). This procedure is particularly suitable to find the limits of detectability of asynchrony between auditory and visual stimuli, while the method of Dixon and Spitz is more suitable for measuring the AV delays that are needed to create a sensation of asynchrony for subjects.

An alternative methodology to study the synchrony perception of AV stimuli is to examine the temporal order in which these stimuli are perceived. The AV delay for which an equal number of stimuli are perceived as video leading and audio leading is then termed the PSE. This method has been studied extensively in relation to the difference in reaction times to auditory and visual stimuli. The assumption in these studies is that the transduction time along the auditory and visual neural pathways is different. This difference can be seen in reaction times that are 50 to 60 ms shorter for auditory stimuli than visual stimuli (e.g. Jaśkowski et al. 1990). This suggests that auditory information travels faster along sensory pathways than visual information. This difference might explain why the PSE is often observed to be in the range of 40 ms audio delay. However, temporal order judgment (TOJ) experiments do not give an unequivocal result in this respect. Some studies indeed report PSE at a positive audio delay (Jaśkowski et al. 1990), while, other TOJ experiments show a negative audio delay for the PSE (e.g. video leading) (Rutschmann and Link, 1964; Aschersleben and Müsseler, 1999). Clearly such results would not be in line with a faster neural transduction time of auditory stimuli. Nor are they in line with the results of Dixon and Spitz (1980).

In a combined study, Smeele (1994) performed a TOJ experiment and an experiment in which subjects had to judge whether an audio-visual speech stimulus was synchronous or asychronous. Even in this study with the same group of subjects and stimuli, there was a difference in the PSE that was found in the two types of experiments. This result was interpreted as an indication that different mechanisms may be underlying the detection of asynchrony and the perception of temporal order.

To get more insight into whether different mechanisms do indeed exist for TOJ and AV asynchrony perception, this paper presents experiments which measured both temporal order judgements and the perception of synchrony versus asynchrony with the same stimuli. In addition an experiment was conducted which was a combination of these two types of experiments.

EXPERIMENTS

In each experiment, a series of short synthetic AV stimuli were presented. The visual component of each stimulus, displayed on a computer monitor, consisted of a white disk on a black background. On each stimulus trial, the disk began at rest and accelerated linearly downwards until it reached a white bar at the bottom of the screen, and then deflected upwards, decelerating linearly, until it came to rest at its initial position. The moment of incidence of the white bar was random, but at least 500 ms away from the beginning and ending of the interval. The auditory stimulus was a 500-Hz tone with a sharp onset with cosine phase and was damped exponentially with a time constant of 30 ms. This onset had an AV delay relative to the moment of incidence of the visual stimulus which varied from –350 ms audio delay (e.g. audio was leading) to 350 ms audio delay, in steps of 50 ms.

Subjects gave their response after each stimulus presentation. The response categories were different for the three experiments: In the first experiment response categories were “audio-first”, “synchronous”, or “video-first” (ASV); in the second experiment response categories were, “synchronous” or “asynchronous” (SA); in the third experiment response categories were, “audio-first” or “video-first” (because these response categories resemble a temporal order judgement this experiment it is labelled TOJ).

The three experiments were conducted in sequential order. Within each experiment, the various AV delays that could occur were presented in random order and each AV delay presentation was repeated 60 times. Four subjects participated in all three experiments.

RESULTS

In Fig.1, the results of all three experiments are shown for the four subjects that participated in the experiment. The results of the first experiment (ASV) are depicted by the filled black symbols. In this experiment subjects could respond “audio first”, “synchronous”, and “video first”. As would be expected, the “synchronous” responses (black circles) are highest in the range where the audio delay is close to zero. The synchronous curve, however, is not centred exactly around zero, but is shifted somewhat toward positive audio delays. This is especially clear for subjects S1 and S2. Thus, the point of subjective equivalence (PSE) does not correspond to the point of objective equivalence (POE) in this experiment (see also Table I). We find that the “audio first” responses (black diamonds) and the “video first” responses (black squares) have high response frequencies for negative and positive audio delays, respectively, in line with expectations.

	ASV	SA	TOJ (1st)	TOJ (2nd)
S1	55	72	-28	-
S2	55	47	50	-
S3	16	29	-15	-43
S4	14	19	-29	-138

Table I: *PSE’s in miliseconds for the three experiments (ASV, SA, TOJ 1st) and for the second TOJ experiment (TOJ 2nd) for all three subjects.*

The results of the second experiment (SA) are shown by the white symbols. The white circles denote “synchronous” responses, and the white triangles denote “asynchronous” responses. As can be seen the synchronous curve of this experiment matches the synchronous curve of the first experiment quite well for each of the four subjects individually. Consequently, very similar differences between the PSE and the POE are found as for the ASV data (see Table I). Only for subject S3 does there seem to be a tendency to accept stimuli with an audio delay to be synchronous over a slightly larger range in the second experiment. Another observation from the data of both the ASV and SA experiment is that for subjects S1 and S2 we see that the transition between “audio first” and “synchronous” responses is sharper than between “video first” and “synchronous”.

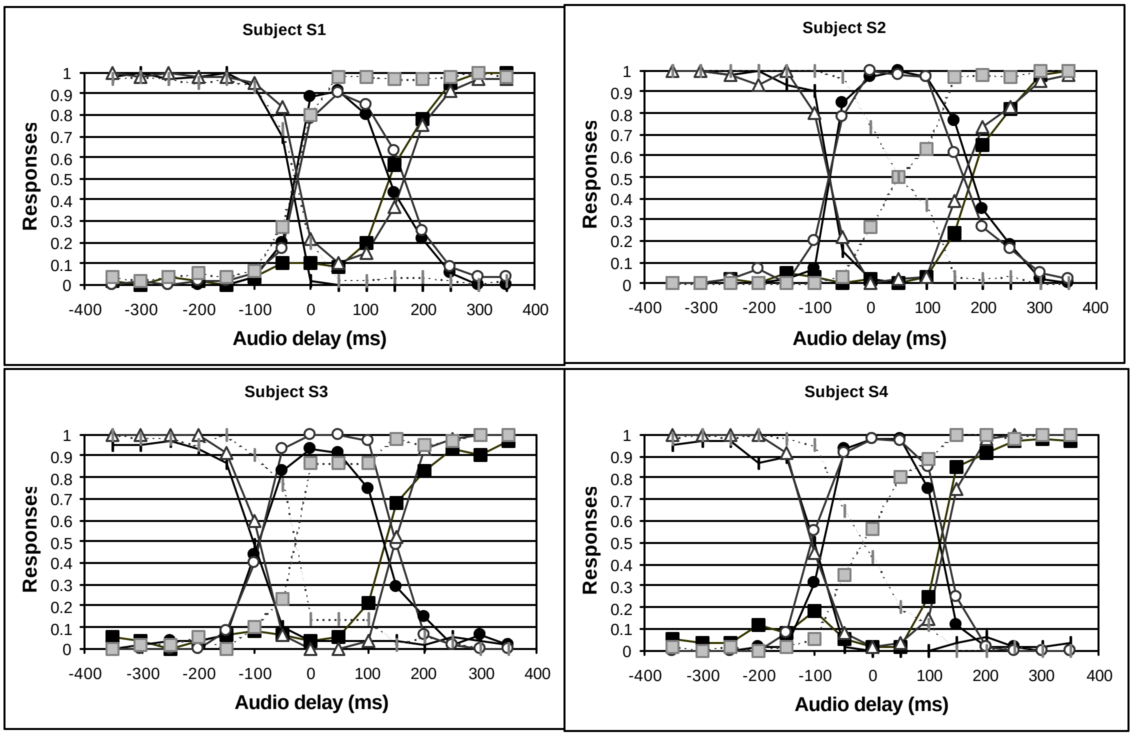

Figure 1: *The results of the three experiments are shown for each subject in a separate panel as a function of audio delay in terms of the proportion of responses. The results of the ASV, SA, and TOJ experiments are shown by the black, the white, and the grey stimuli, respectively. Responses could be (depending on experiment) "video first" (squares), "audio first" (diamonds), "synchronous" (circles), and "asynchronous" (triangles).*

The results of the third experiment (TOJ) are shown by the grey symbols. When the PSE is defined as the crossing point of the "audio first" and "video first" curves, we see that one subject (S2) shows a positive PSE, two subjects (S1 and S3) show a negative PSE and subject S4 has a PSE that is negative but closer to zero. For three out of the four subjects the crossing of the curves for "audio first" responses (grey diamonds), and "video first" responses (grey squares) falls within the range where subjects predominantly responded with "synchronous" in the first two experiments. For one subject S1, the transition coincides with the crossing of "synchronous" curve with the "audio first" curve of the first experiment.

DISCUSSION

The first experiment (ASV) shows that the point of subjective equivalence (PSE) does not coincide with the point of objective equivalence (POE). This general observation is in line with the findings of many previous experiments (e.g. Dixon and Spitz, 1980). In our data the PSE is found at an audio delay of about 35 ms. In trying to understand the difference between PSE and POE, it is interesting to consider that the speed of sound is limited and that at a distance between the source of sound and the observer of about 10 metres, sound would take about 35 ms longer to reach the observer than light. Thus, the difference between PSE and POE may be regarded as some kind of accommodation to compensate for AV delays that occur in daily life when observing remote objects. There are some indications that such an accommodation is not acquired by learning, because a difference between PSE and POE is already found in children a few months old (Lewkowicz, 1996). Also from physiological sources it is known that neural transduction times from the peripheral sensors towards the Superior Colliculus are about 50 ms shorter for auditory stimuli than for visual stimuli (Meredith et al., 1987).

The results of the second experiment (SA) show that the response category “synchronous” in this and the first experiment are practically identical and consequently the “asynchronous” responses category in this experiment is the sum of the “audio first” and “video first” response categories. The response curves also show that when a stimulus is perceived as asynchronous, subjects are generally able to decide whether audio or video was first. This observation does not lend support to the hypothesis that different perceptual mechanisms mediate the perception of synchrony and temporal order in AV time perception such as suggested by Smeele (1994) in the context of speech stimuli. According to this hypothesis there could be a perceptual state where a stimulus is perceived as asynchronous while it is not possible to determine the temporal order.

It is noteworthy to mention that between the first and the last two experiments a time span of two years elapsed. Considering this, the correspondence between the results of the first and second experiments indicates that the perception of AV synchrony, albeit being different *between* subjects, is highly consistent over time *within* subjects.

In the last experiment (TOJ), we found PSE’s to be inconsistent across subjects; finding both negative and positive values. Also the relative position of response curves found in the TOJ experiment compared to those from the first two experiments varied across subjects. For subjects S2, S4, and to a lesser extent for subject S3, the crossing point of the TOJ response curves is centred within the “synchronous” response curve of the first two experiments. For subject S1, this crossing point coincides with the crossing of the “synchronous” and “audio first” curves of the first experiment.

An explanation for the variable results across subjects that we obtained in the TOJ experiment may be that subjects adopted different decision criteria for determining whether audio or video was leading. If we assume that there are three perceptual states that may occur, audio leading, synchronous, or video leading, the strategy that may have been adopted by subjects S2 and S4 would be to place the criterion somewhere within the synchronous perceptual state. In other words, the strategy would be to make the best possible decision about whether audio or video was leading even if the stimulus is perceived as synchronous. A different response strategy would be to respond with “audio first” if the stimulus is perceived as audio leading and to respond with “video first” when the stimulus is perceived as synchronous or video leading. This would mean that the decision criterion would be put at the transition between the audio-leading and synchronous perceptual states. As was seen for subjects S1 and S2, this transition is rather sharp and may therefore be more preferable in order to make a clear distinction between the two cases. Indeed subject S1 has a very sharp transition in the TOJ experiment which coincides with the transition that was found in the first experiment, while subjects S2 and S4 have considerably shallower transitions.

To test whether indeed it is possible to adopt different response strategies, subjects S3 and S4 repeated the experiment, while being instructed to use the boundary between the audio leading and synchronous perceptual states as the decision criterion. Fig. 2 shows the results of the repeated TOJ experiment together with the results of the first two experiments. For subject S3 the repeated results are barely different from those of the first TOJ experiment (see also Table I). For subject S4, however, there is a clear shift in the position of the transition between the “audio first” and “video first” curves and now the transition between the curves nearly coincides with the transition of the “audio first” and “synchronous” response curves of the first experiment. It is also interesting to see that the transition is now sharper than in the first TOJ experiment, in line with our expections. Apparently, the decision strategy that can be employed in a TOJ experiment is not uniquely defined by the task given to subjects and it is possible for a subject to change the decision strategy.

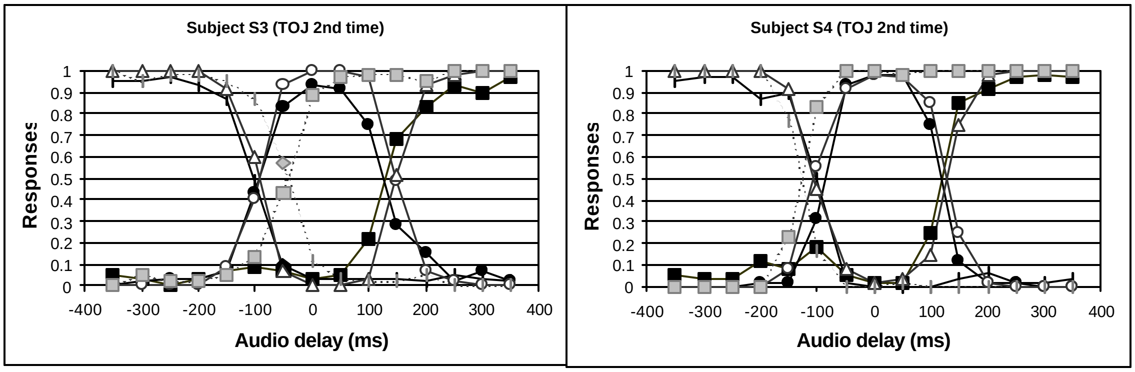

Figure 2: *The same as in Fig. 1 only now the TOJ data that are shown were obtained in a second experiment where subjects were instructed to adopt a different decision strategy.*

CONCLUSIONS

The point of subjective equivalence in auditory-visual synchrony perception is shifted towards audio delays by about 35 ms compared to the point of objective equivalence. This result is found both in experiments in which subjects could respond with “audio first”, “video first”, or “synchronous” (first experiment) as in which they could respond with “synchronous” or “asynchronous” (second experiment). The results of temporal order judgement experiments gives PSE’s with both video and audio delays. We suggest that this may be due to different interpretations of the task that subjects have to perform in a TOJ experiment which results in different decision strategies employed by the subjects.

LITERATURE

Aschersleben, G., and Müsseler, J. (1999), “Dissociations in the timing of stationary and moving stimuli,” J. of Experimental Psychology, **Vol.** 25, pp. 1709-1720

Dixon, N.F., and Spitz, L. (1980), “The detection of auditory visual desynchrony,” *Perception*, **Vol.** 9, pp. 719-721

Jaśkowski, P., Jaroszyk, F., and Hojan-Jezierska, D. (1990), “Temporal-order judgment and reaction time for stimuli of different modalities,” *Psychol. Res.*, **Vol.** 52, pp. 35-38

Meredith, M.A., Nemitz, J.W., and Stein, B.E. (1987), “Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors,” *J. of Neuroscience*, **Vol.** 7, pp. 3215-3229

Lewkowicz, D.J. (1996), “Perception of auditory-visual temporal synchrony in human infants,” *J. Experimental Psychology: Human Perception and Performance*, **Vol.** 22, pp. 1094-1106

Rutschmann, J, and Link, R. (1964), “Perception of temporal order of stimuli differing in sense mode and simple reaction time,” *Perceptual and Motor Skills*, **Vol.** 18, pp. 345-352

Smeele, P.M.T. (1994), “Perceiving Speech: Integrating Auditory and Visual Speech,” *Ph.D. thesis Delft University of Technology*, Delft

van de Par, S., and Kohlrausch, A. (2000), “Sensitivity to Auditory-Visual Asynchrony and to Jitter in Auditory-Visual Timing,” *Human Vision and Electronic Imaging V*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Editors, *Proceedings of SPIE*, **Vol.** 3959, pp. 234-242