Compensation for a large gesture-speech asynchrony in instructional videos

Andrey Anikin, Jens Nirme, Sarah Alomari, Joakim Bonnevier, Magnus Haake

Lund University Cognitive Science, Sweden

rty_anik@yahoo.com, jens.nirme@lucs.lu.se, sjawdat@hotmail.com,
joakim.bonnevier.213@student.lu.se, magnus.haake@lucs.lu.se

Abstract

We investigated the pragmatic effects of gesture-speech lag by asking participants to reconstruct formations of geometric shapes based on instructional films in four conditions: sync, video or audio lag (±1,500 ms), audio only. All three video groups rated the task as less difficult compared to the audio-only group and performed better. The scores were slightly lower when sound preceded gestures (video lag), but not when gestures preceded sound (audio lag). Participants thus compensated for delays of 1.5 seconds in either direction, apparently without making a conscious effort. This greatly exceeds the previously reported time window for automatic multimodal integration.

Index Terms: gesture-speech synchronization, multimodal integration, temporal synchronization, comprehension

1. Introduction

Manual gestures facilitate speech production, evidenced by the fact that they persist when blind people speak among themselves [1] or when the listener is not visible [2]. Furthermore, gestures may improve listening comprehension, especially when speech is ambiguous [3] or when there is a lot of background noise [4]. But how exactly are gestures temporally related to speech? How important is this temporal relation to successful communication?

An influential view is that speech and gesture share a common origin and are best seen as two forms of the same communicative process [5],[6]. Their temporal relationship is determined by the semantic and pragmatic synchrony rules: if speech and gestures co-occur, they must either present the same semantic information or perform the same pragmatic function. It is well established that gestures are generally initiated simultaneously with – or slightly before – the onset of their lexical affiliates [7],[8],[9],[10]. But a new question immediately arises: Are they synchronized because this is necessary for successful comprehension or simply because speech and gesture stem from the same "idea unit"? [5],[6]

One way to answer this question is to see how a disruption of the natural synchronization affects comprehension. Since speech and gesture exploit different modalities, this is a case of multisensory integration, which is affected by the synchronicity of the two channels [11]. Of course, the time-window of tolerance for asynchrony varies depending on the nature of stimuli.

Several studies have found effects of gesture asynchrony on event-related potentials elicited around 400 ms after the onset of a word (N400) indicative of integration difficulty. Habets et al. [12] found a greater N400 to mismatched versus matched gesture-speech sequences only when speech lagged by 0 and 160, but not by 360 ms. The authors conclude that gesture and speech are integrated automatically when they fall within 160 ms of each other, so that a gesture which does not semantically match speech leads to effortful processing. Obermeier and Gunter [13] found an N400 effect for gestures related to either dominant or subordinate meanings of an ambiguous word from approximately -200 ms (speech lag) to +120 ms (gesture lag). Other studies have found a greater perceived emphasis on words when they are synchronized with gestures [10],[14].

A view emerges that gesture and speech may be integrated either automatically or with some conscious effort, depending on how precisely they are synchronized. The window for automatic integration is, however, well within the time frame reported for naturalistic conversations. For example, Morrel-Samuels & Krauss [15] discovered that gestures were never preceded by their lexical affiliates in their data bank, but the onset of gesture usually preceded its lexical affiliate. In fact, the mean reported delay was 1 second, and for less familiar words it could be as long as 3.8 seconds! This result emphasizes the simple fact that we should not underestimate the variability of gesture-speech synchronization in natural conversation. In fact, delays too large for automatic integration may be a normal feature of conversation, for which humans must possess a compensatory mechanism.

Practical implications of gesture-speech lag remain relatively unexplored, partly due to methodological problems with generating naturalistic sequences with mismatched speech and gestures. In particular, lip movements quickly give the manipulation away, unless the face is hidden or computer animation is used to separate facial from bodily movements. Practical implications of gesture-speech synchronization are, however, more relevant today, when digital agents are becoming increasingly common as chatterbots or virtual service desk personnel. Woodall & Burgoon [16] found that actors who purposefully delayed their gestures by up to 1 sec were perceived as less persuasive, and this delay impaired recall. However, in this paradigm speech is not identical in different conditions. Further study of the effects of gesture-speech (de)synchronization on overall comprehension as well as the perceived competence of digital agents is an essential part of the effort to make computer-human communication smooth and effortless. The results could drastically change the way digital agents speak and move.

There is some preliminary evidence that people tolerate much larger speech-gesture delays than the window in which multimodal integration occurs automatically. In a study by Kirchhof [17] 60% of participants accepted gesture-speech pairs as natural with delays from -600 to +600 ms. Furthermore, when asked to synchronize the audio and video tracks, participants chose delays from approximately -1.8 s (gestures first) to +1.2 s (speech first). The author concludes that gestures and speech are more closely synchronized in production than is necessary for successful comprehension. A limitation of Kirchhof's approach is that the perceived naturalness of a clip or the chosen audio-video offset time are both explicit measures that tap into subjective evaluation rather than implicit comprehension. To examine the latter, we would need to assess the pragmatic effects of the multimodal message on observable behavior.

Accordingly, we designed a practical task that required the participant to integrate the visual and the auditory channels, so that performance could be a measure of how successfully speech was integrated with gestures at different time lags. The perceived difficulty of the task and quality of instructions were assessed in a short questionnaire and provide explicit measures of the effects of gesture-speech lag. The main question is how overall comprehension is affected by a large audio or video lag and whether it is associated with subjectively experienced cognitive effort and/or dissatisfaction with the speaker.

2. Methodology

2.1. Participants

83 participants were students recruited and tested at Lund University. Data collection followed the recommendations of *Good Research Practice* by the Swedish Research Council [18] with respect to information to participants, consent, debriefing, confidentiality, and data use.

2.2. Experimental task

Participants were asked to recreate arrangements of five geometric shapes (Figure 1) after watching short videos, in which an instructor spoke and used gestures but did not show the physical objects. The shapes had to be selected from an array of 8 objects: two boxes, two sticks, a ball, a can, a tube, and a small cylinder. The task could be performed incrementally; missing a single step of the instructions did not preclude successful completion of later steps. The videos were presented in one of four conditions (sync, video-lag, audio-lag or audio-only). "Sync" in this case stands simply for original, unmodified clip; the files were not manipulated to ensure perfect synchronization of gesture strokes with their semantic referents. The lag conditions operated with delays of 1,500 ms. This value was chosen based on the results of a pilot study with 11 participants and delays up to +/- 2 s. In the audio-only condition the soundtracks were presented without any accompanying video.

2.3. Materials

Six short instructional videos from 42 to 82 seconds in duration were filmed with a hand-held camera, which was placed high over the shoulder of the instructor so as to provide an unobstructed view of his manual gestures but not his face. The instructor was a male student, who was told that the focus of this study was the effectiveness of communication but was naive to the exact purpose of the study. He was not specifically asked to gesture but encouraged to describe what needed to be done "as well as he could". For each trial, a picture of the target formation was shown on the screen of a smartphone, which the instructor kept in his lap (off camera) while explaining how to build the formation. The recordings were split into separate video- and soundtracks. Each soundtrack was converted to a sample rate of 44100 Hz, filtered to remove background noise, and normalized. In case of mistakes or unwanted noise, such as the sound of a hand slapping the desk, a new take was filmed. All 11 participants in a pilot study confirmed that they could hear the instructions clearly and were not bothered by the camera angle.

One of the authors identified gesture phases in the videos [5]. The gestures produced by the instructor were short in duration (duration: M = 580 ms, SD = 270 ms). The videos contained 124 gestures in total with a gesture stroke on average every 4.80 words. A concern with our video manipulations was the lack of control over what the temporally offset gestures ended up being synchronized with. Even with the large temporal offset used, chances are that gesture strokes still overlap with congruent speech (referring to the same position, orientation or shape of an object as the gesture). We thus categorized the overlapping speech in the manipulated videos as being either congruent or incongruent (overlapping with irrelevant or contradicting speech or silence). The proportion of congruence was very similar in the video lag (40.1%) and audio lag (38.8%) conditions. In the synchronized videos some natural "asynchrony" was present, but generally it was well within the magnitude of the delay introduced in the manipulated videos: 28.2% of the gesture strokes preceded the stressed syllable of their lexical affiliates (median offset 130 ms) and, conversely, 5.6% of the affiliated stressed syllables preceded the onset of the gesture strokes (median offset 80 ms).

2.4. Procedure

At the beginning of the experiment participants were asked to evaluate their ability to read maps (a skill judged to be functionally similar to the demands of the main test). As a practical pre-test, they also had to arrange small pieces of paper "furniture" in the drawing of a room based on verbal instructions. The instructions were read by the experimenter slowly, but without repetitions, and the resulting arrangement was informally assessed on a scale of 1 (poor) to 3 (good).

After this the participants were randomly assigned to an experimental condition (except the audio-only group, which was tested after the others) and started the main experiment, which consisted of six trials. In each trial the participant was asked to reconstruct a formation of five geometric shapes after watching an instructional video presented in *PsychoPy* [19]. Participants were instructed to watch the instruction videos first and then choose and arrange the correct five objects, so that they would not have to divide their attention between the videos and the objects. The reconstructed array was photographed for future coding, and the participant proceeded to the next trial. If a participant could not recall all five objects, they were not pressed to guess but their incomplete arrays were accepted as they were. All trials, except in the audio-only condition, were double-blind: neither participants nor experimenters knew which condition was being tested.

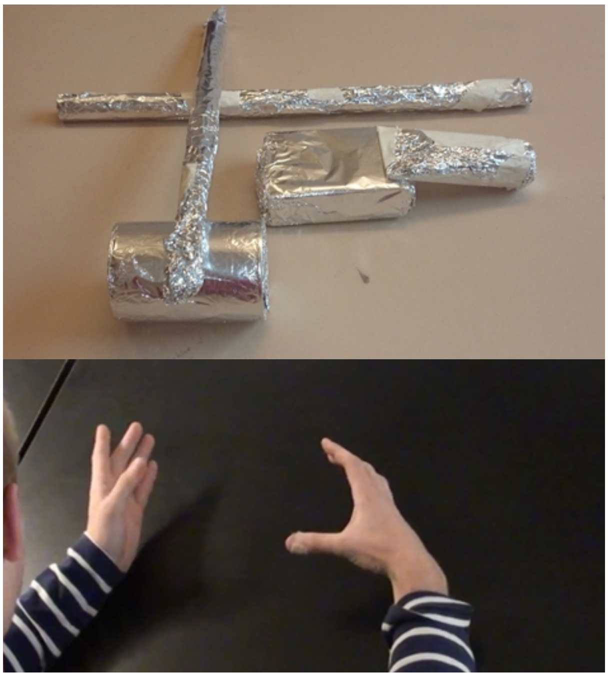

Figure 1. *(Upper) An example of the original formation in trial 6; (Lower) Frame from the (synchronized) instructional video in trial 6, extracted from segment when the instructor describes the position of the rectangular shape.*

After completing six trials, participants filled out a short questionnaire (Table 1) rating the difficulty of the experimental task and the efficiency of the instructor on a visual analogue scale (VAS). They were also encouraged to leave free-text feedback, once after rating the task and once after rating the instructor. Finally, each participant was debriefed and asked whether they had noticed anything strange about the video and sound. If they did not report noticing anything unusual, they were then asked directly whether the video- and soundtracks were synchronized. The entire procedure took 15-20 minutes.

Table 1. *Questionnaire items.*

How difficult was it to understand: (difficult / not difficult) – the instructor's speech? – which shapes to use? – the relations between the shapes? – what to build?
How did you find the instructor: – clear / not clear? – certain / not certain? – professional / not professional?

2.5. Coding

The "furniture" pre-test was coded informally by one of the authors for all participants. All trials were coded independently by two other authors based on an algorithm which awarded points for the correct choice of each object as well as its position in two dimensions, three dimensions, and in relation to the reference object used to describe the location of the object in question. A maximum of 19 points could be awarded for each of the six trials, i.e. maximum 114 points per participant. The coders were (except in the audio-only condition) blind to the experimental condition. Any disagreements in coding were discussed by the two coders, and then either a compromise solution was reached or two different scores were entered in the database and averaged.

2.6. Analysis

All statistical analyses were performed in *R* [20]. Implicit comprehension was operationalized as the total score out of the maximum of 114 on all six experimental trials. The scores from two coders were averaged and rounded to the nearest integer (where different) and modeled with binomial generalized linear mixed models (GLMM) with a random intercept per participant using the *lme4* package [21] and Bayesian modeling with Markov chain Monte Carlo (MCMC) method using *rjags* [22],[23]. Explicit comprehension and satisfaction with the instructor were measured on a VAS and analyzed using ANOVA and *rjags*. Free-text comments were categorized by attitude (neutral/critical) as well as direction (towards the task/the instructor/oneself). Note that commenting was optional and no participant made positive comments.

3. Results

A total of 83 participants (45 females and 38 males) completed the experiment in one of four conditions (Table 2). There were no significant differences between experimental groups in baseline characteristics, such as gender composition ($\chi2(3, N = 83) = 3.68$, $p = .30$) and the score on the “furniture” pre-test ($\chi2(6, N = 83) = 2.07$, $p = .91$). ANOVA of self-assessed ability to read maps also failed to discover any effect of condition ($F(3,79) = 0.62$, $p = .60$). Total scores awarded by both coders were very strongly correlated, demonstrating high inter-rater reliability (Spearman’s rho: $\rho = .98$). Two participants reported noticing that the audio and video were out of sync; both were in the audio-lag group and both performed extremely well on all trials. Another seven participants (4 in audio-lag and 3 in video-lag condition), when told during debriefing that there might have been a delay, were not sure whether they had noticed it or not; their performance was a bit below average.

Table 2. *Baseline characteristics of study groups*

Group	Number of participants (male / female)	Self-rated ability to read maps ($M \pm SD$)	Pre-test score low/med/high (%)
Sync	20 (12 / 8)	66.7 ± 22.1	5 / 35 / 60
Audio lag 1.5 s	23 (11 / 12)	68.3 ± 24.5	13 / 22 / 65
Video lag 1.5 s	20 (9 / 11)	61.6 ± 23.6	10 / 25 / 65
Audio only	20 (6 / 14)	60.3 ± 20.2	15 / 30 / 55

3.1. Implicit comprehension

Implicit understanding of the instructional videos was assessed by comparing each reconstructed array with the original and adding up the scores on all trials.

Individual variation of the total score per subject proved to be very considerable, but the overall level of success was high ($M = 78.7\%$, $SD = 10.1\%$). The mean total score per participant in each condition was as follows ($M \pm SD$ as proportion of maximum): sync = 82.4% ± 8.0%, audio lag = 81.9% ± 11.5%, video lag = 78.4% ± 9.7%, audio only = 71.7% ± 7.2% (Figure 2).

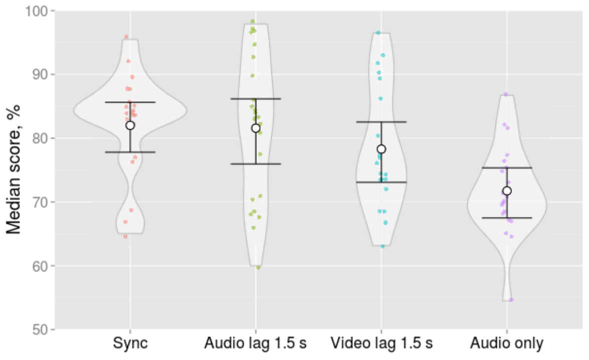

Figure 2. *The distribution of scores for each group (median and 95% credibility intervals).*

According to the MCMC model, there is evidence for higher scores in all three video groups compared to the audio-only group. The most credible difference (median (%) and 95% highest density interval) for sync / audio-lag / video-lag *vs* audio-only conditions is 10.3 [5.0, 16.1], 9.8 [3.3, 16.2] and 6.6 [0.0, 12.3], respectively. In contrast, sync and audio-lag group have essentially the same average scores, while the most credible difference in scores between sync and video-lag groups is only 3.7% [-2.6%, 9.6%]. The video-lag group thus appears to score in between sync and audio, but closer to the former. The difference between all conditions, including audio-only, is small relative to variance within each condition, which translates into low statistical power. A retrospective power analysis shows that we were 84% likely to prove that all 3 video conditions exceed the audio-only condition, 55% likely to prove the difference between the video-lag and audio-only conditions, and only 23% likely to prove the difference between the sync and video-lag conditions.

Naturally, performance on the experimental task may be strongly affected by the individual spatial abilities of each participant, and the effect of condition may depend on these abilities. GLMM models were therefore fitted to investigate possible interactions between condition and each of two measures of underlying spatial ability: (1) the direct question (“How do you evaluate your ability to read maps?”) and (2) the score on the “furniture” pre-test, in which the participant had to arrange furniture based on verbal instructions. The interaction between self-rated spatial ability and experimental condition is strong (likelihood ratio test: $L = 14.1$, $df = 3$, $p = .003$). The same holds for the score on the “furniture” pre-test ($L = 15.8$, $df = 3$, $p = .001$). Better results on the pre-test thus predict higher scores on the main task, but primarily in the audio-lag condition.

3.2. Explicit comprehension

Individual scores on the four questions related to the difficulty of the task are strongly correlated (Cronbach’s alpha = .86), therefore they were combined and analyzed as a single item, with a significant main effect of condition in ANOVA: $F(3,79) = 12.1$, $p < .001$. The overall rating of task difficulty was higher in the audio-only condition compared to any other condition (the most credible difference is 27% [18%, 36%]). The evidence for any difference between the video conditions is very weak (the highest-density intervals include zero for each comparison). The task was thus judged to be considerably more difficult by participants in the audio-only group, but with no difference between the three video groups (Figure 3).

As for the three questions in which participants rated their satisfaction with the instructor, scores on the individual items were also strongly correlated (Cronbach’s alpha = .93). These three questions were therefore combined. There is a noticeable main effect of condition on the combined score on these three items ($F(3,78) = 4.6$, $p = .006$). Compared to sync condition, the instructor received lower ratings in the audio-only and audio-lag conditions (the most credible difference is 24.9% [10.8%, 41.2%] and 20.3% [6.1%, 36%], respectively). The uncertainty is high, but it appears that satisfaction with the instructor was highest in the sync condition, lowest in the audio-only condition and intermediate in the audio/video-lag conditions (see Figure 3).

Participants provided in all 77 free-text comments (out of 168 opportunities). As can be seen in table 3, the distribution of comments of different types across conditions was not

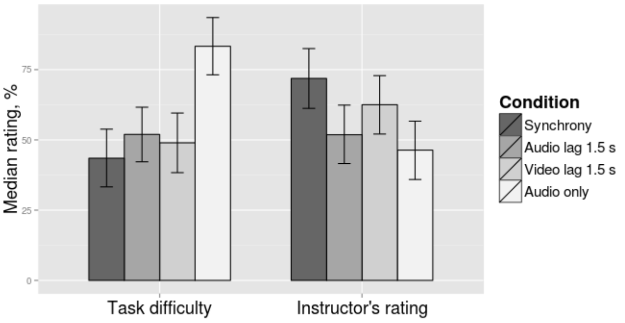

Figure 3. *Subjective ratings of the difficulty of the tasks and satisfaction with the instructor (median and 95% credibility intervals).*

uniform. Comments directed towards the difficulty of the task (e.g. “that was a lot of information”) were rare, and critical comments directed towards the instructor (e.g. “he did not seem to know what he was doing”) more frequent in the audio-lag and video-lag conditions. The participants in the audio-only group were more likely to direct criticism towards themselves (e.g. “I had trouble keeping all that information in my head”).

Table 3. *Number of free-text responses classified as neutral or critical per group and comment direction.*

Directed towards	Number of comments, neutral / critical			
	Sync N=20	Audio lag N=23	Video lag N=20	Audio only N=20
Task	0 / 2	0 / 1	1 / 0	1 / 3
Instructor	0 / 1	4 / 9	1 / 9	4 / 5
Self	3 / 1	5 / 0	5 / 2	4 / 9

4. Discussion

As Woodall and coauthors long ago pointed out, it is important to establish how closely verbal and nonverbal behaviors are synchronized during communication and describe the nature of this synchronization process, but *“an equally important issue is how it affects communication outcomes such as information exchange and persuasion”* [16]. The latter point has been largely neglected since, but today the ubiquity of digital agents makes this straightforward question of great practical significance: what degree of gesture-speech desynchronization is tolerated before communication breaks down and/or the receiver gets annoyed?

The task used in this study was designed to be solvable only if the audio and video channels are integrated. The fact that scores in the audio-only condition were significantly lower than in the three video conditions (full sync, audio lag 1.5 s, and video lag 1.5 s) indicates that both modalities were needed to solve the task. Our result does not reveal that a delay of 1.5 seconds in either direction prevents the receiver from integrating gestures with speech, despite a weak tendency for lower performance in the video-lag group. Furthermore, compared to the sync condition, the task was rated as considerably more difficult in the audio-only condition but not in the audio/video lag conditions. Not only could the participants integrate gestures and speech despite the large delay, but they did so without experiencing the task as more difficult. However, in contrast to the ratings of the *task*, ratings of the *instructor* were affected by delay, as the instructor in audio-lag condition was rated as worse than in the sync condition and almost as low as in the audio-only condition. More free-text criticism was also directed towards the instructor in both the audio- and video-lag conditions. Unexpectedly, criticism was less likely to be directed towards the instructor when he was not visible, despite the low VAS ratings that he received in this condition.

Clearly, individual variation in spatial abilities may influence the results. Indeed, we discovered a highly significant interaction between both measures of spatial skills (self-evaluated ability to read maps and performance in the “furniture” pre-task) and experimental condition. Participants with good spatial skills in the audio lag group were able to fully compensate for the temporal mismatch, while those with poor spatial skills were unable to compensate and performed worse compared to participants in the sync group. Intriguingly, spatial skills had very little effect on performance in the video-lag condition and none at all in the audio-only and full-sync conditions. Given the small sample sizes, this difference could be spurious, or it could indicate that certain cognitive skills are involved in compensating for the lack of synchrony which are not manifest in other experimental conditions.

On the one hand, it is somewhat surprising that the participants could compensate relatively successfully for such a large delay as 1,500 ms, when previous studies have found that the time window for automatic integration spans no more than a few hundred milliseconds [12],[13]. It is especially impressive when the audio track is advanced relative to the video track - the “atypical” direction, since speech hardly ever precedes gestures in natural conversation [10],[14].

On the other hand, integration of visual and auditory stimuli with very large delays has been reported before. In a study of the McGurk effect, Campbell and Dodd [24] presented participants with short words using audio lags of 400, 800 and 1600 ms. Phoneme identification was optimal in the full sync condition, but even at the longest delay identification was better compared to the audio-only control condition. In a recent series of studies Kirchhof [17] discovered that surprisingly large temporal mismatches of gestures were accepted as natural.

An important question to ask pertains to the mechanisms of cross-modal integration at these longer delays. What exactly happens if gestures and speech are poorly matched temporally and fail to be integrated “automatically” back into a single “idea unit”? An influential position in psychology invokes the notion of “mental models” [25] or “situation models” [26],[27] – holistic representations of the described situation, which are integrated across sentences, modalities, sometimes even languages and multiple documents or conversations. The temporal structure of messages is not always linear. Grammatical rules being what they are, the order of events in a narrative does not always correspond to the order in which they are mentioned in a sentence: for instance, we may say: “Before I opened the door, I had to search for my keys for a few minutes”. Seen in this light, a gesture-speech lag of a second or so is a special case of integrating information arriving from different modalities and at different times into a unified situation model. In line with Massaro's “fuzzy logical model of perception” [28], the two modalities will probably be integrated as long as they are perceived as belonging to the same perceptual event. Then again, gesture and speech can hardly be attended to as two completely independent channels. Instead, it seems likely that speech sets up a context for interpreting gestures, and vice versa [29]. This integration may not be automatic, but judging by the rating of task difficulty in the four groups, it requires very little conscious effort.

A limitation of the task used in this study is that both average performance and individual variability were high, making it harder to detect differences between groups. In other words, the auditory channel alone contained enough information for some participants to perform near the ceiling, while others struggled even when presented with unmanipulated videos. As a result, it is hard to be certain whether the tendency for lower scores in video-lag compared to sync condition is an artifact. It would be desirable to try other experimental tasks, in which the informational load of gestures is higher.

Similarly, the tendency for somewhat lower satisfaction with the instructor in the audio/video lag conditions is suggestive, but the evidence is inconclusive. An independent measure of effort could help reveal if this tendency stems from an increased effort manifested as frustration with the instructor without attribution of difficulty to the task itself. Given the high natural variability in gesture-speech temporal coordination, the strokes of the instructor’s gestures did not necessarily have a tight temporal coupling with their lexical referents in the original unmanipulated videos. In fact, despite the large temporal offset, around 40% of the gestures in manipulated videos still overlapped with semantically congruent speech (although this effect was balanced between the video-lag and audio-lag conditions). The stimuli also included instances of spoken deictic expressions referencing the gestures (“this”, “here”). In these cases instructions were clearly incomplete when the associated gestures were missing in the desynchronized videos. Eliminating congruent overlap and such obvious mismatches by a strict selection of instruction videos from a larger set might reveal effects that our results did not.

In summary, this study investigated whether desynchronized speech and gestures can still communicate task-relevant information. The answer, at least for the task investigated here, is a clear yes. Not only is compensation nearly perfect, but the participants fail to notice a delay of 1.5 seconds in either direction and do not make a conscious effort to integrate desynchronized gestures and speech. Asynchrony may, however, cause the speaker to appear less competent. Many issues, such as the generalizability of this outcome, the nature of integration processes and the cognitive skills involved, await further research.

5. References

[1] Iverson, J. M., and Goldin-Meadow, S., “Why people gesture when they speak”, Nature, 396(6708):228-228, 1998.

[2] Wagner, P., Malisz, Z., and Kopp, S., “Gesture and speech in interaction: An overview”, Speech Communication, 57:209-232, 2014.

[3] Thompson, L. A., and Massaro, D. W., “Evaluation and integration of speech and pointing gestures during referential understanding”, Journal of Experimental Child Psychology, 42(1):144-168, 1986.

[4] Rogers, W. T., “The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances”, Human Communication Research, 5(1):54-62, 1978.

[5] Kendon, A., “Gesticulation and speech: Two aspects of the process of utterance”, in M. Key [Ed], The Relationship of Verbal and Nonverbal Communication, 207-227, Mouton, 1980.

[6] McNeill, D., “So you think gestures are nonverbal?”, Psychological Review, 92(3):350-371, 1985.

[7] McNeill, D., “ Hand and mind: What gestures reveal about thought”, University of Chicago Press, 1992.

[8] Nobe, S., “Where do most spontaneous representational gestures actually occur with respect to speech?”, in D. McNeill [Ed], Language and Gesture, 186-198, Cambridge University Press, 2000.

[9] Loehr, D., “Aspects of rhythm in gesture and speech”, Gesture, 7(2):179-214, 2007.

[10] Treffner, P., Peter, M., and Kleidon, M., “Gestures and phases: The dynamics of speech-hand communication”, Ecological Psychology, 20(1):32-64, 2008.

[11] Liu, B., Jin, Z., Wang, Z., and Gong, C., “The influence of temporal asynchrony on multisensory integration in the processing of asynchronous audio-visual stimuli of real-world events: an event-related potential study”, Neuroscience, 176: 254-264, 2011.

[12] Habets, B., Kita, S., Shao, Z., Özyurek, A., and Hagoort, P., “The role of synchrony and ambiguity in speech–gesture integration during comprehension”, Journal of Cognitive Neuroscience, 23(8):1845-1854, 2011.

[13] Obermeier, C., and Gunter, T. C., “Multisensory Integration: The Case of a Time Window of Gesture–Speech Integration”, Journal of Cognitive Neuroscience, 27(2):292-307, 2015.

[14] Krahmer, E., and Swerts, M., “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception”, Journal of Memory and Language, 57(3):396-414, 2007.

[15] Morrel-Samuels, P., and Krauss, R. M., “Word familiar ity predicts temporal asynchrony of hand gestures and speech”, Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3):615-622, 1992.

[16] Woodall, W. G., and Burgoon, J. K., “The effects of nonverbal synchrony on message comprehension and persuasiveness”, Journal of Nonverbal Behavior, 5(4):207-223, 1981.

[17] Kirchhof, C., “Desynchronized speech-gesture signals still get the message across”, in the 7th International Conference on Multimodality (7ICOM), Hongkong, 2014.

[18] The Swedish Research Council, “God forskningssed”, Vetenskapsrådets rapportserie, 1:2011, 2011.

[19] Peirce, J., PsychoPy – Psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2):8-13, 2007.

[20] R Core Team, “R: A language and environment for statistical computing (R version 3.1.3)”, [Statistical software], R Foundation for Statistical Computing, 2015. Available: www.R-project.org/

[21] Bates D., Maechler, M., Bolker, B., and Walker, S., “lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7)” [Statistical software], 2014. Available: CRAN.R-project.org/package=lme4.

[22] Plummer, M., “rjags: Bayesian graphical models using MCMC (R package version 3-14)”, [Statistical software], 2014. Available: CRAN.R-project.org/package=rjags

[23] Kruschke, J., “ Doing Bayesian data analysis: A tutorial introduction with R”, Academic Press, 2010.

[24] Campbell, R., and Dodd, B., “Hearing by eye”, Quarterly Journal of Experimental Psychology, 32(1):85-99, 1980.

[25] Johnson-Laird, P. N., “Mental models: Towards a cognitive science of language, inference, and consciousness”, Harvard University Press, 1983.

[26] Van Dijk, T. A., and Kintsch, W., “ Strategies of discourse comprehension”, Academic Press, 1983.

[27] Zwaan, R. A., and Radvansky, G. A., “Situation models in language comprehension and memory”, Psychological bulletin, 123(2):162-185, 1998.

[28] Massaro, D. W., Cohen, M. M., and Smeele, P. M., “Perception of asynchronous and conflicting visual and auditory speech”, The Journal of the Acoustical Society of America, 100(3):1777-1786, 1996.

[29] Kelly, S. D., Barr, D. J., Church, R. B., and Lynch, K., “Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory”, Journal of Memory and Language, 40(4):577-592, 1999.