Temporal integration for live conversational speech

Ragnhild Eg[1], *Dawn M. Behne*[2]

[1]Simula Research Laboratory, Lysaker, Norway
[2]Department of Psychology, Norwegian University of Science and Technology, Norway
rage@simula.no, dawn.behne@svt.ntnu.no

Abstract

The difficulty in detecting short asynchronies between corresponding audio and video signals demonstrates the remarkable resilience of the perceptual system when integrating the senses. Thresholds for perceived synchrony vary depending on the complexity, congruency and predictability of the audiovisual event. For instance, asynchrony is typically detected sooner for simple flash and tone combinations than for speech stimuli. In applied scenarios, such as teleconference platforms, the thresholds themselves are of particular interest; since the transmission of audio and video streams can result in temporal misalignments, system providers need to establish how much delay they can allow. This study compares the perception of synchrony in speech for a live two-way teleconference scenario and a controlled experimental set-up. Although methodologies and measures differ, our explorative analysis indicates that the windows of temporal integration are similar for the two scenarios. Nevertheless, the direction of temporal tolerance differs; for the teleconference, audio lead asynchrony was more difficult to detect than for the experimental speech videos. While the windows of temporal integration are fairly independent of the context, the skew in the audio lead threshold may be a reflection of the natural diversion of attending to a conversation.

Index Terms: audiovisual speech, temporal integration, synchrony perception, teleconference

1. Introduction

Perception of synchrony is frequently applied as a tool to evaluate the temporal integration of audiovisual events. Historically, researchers have looked at very simple stimuli, such as flash and tone combinations, to study basic perceptual processes (e.g., [1,2]). In later years this methodology has been extended to more complex audiovisual events, in particular to speech (e.g., [3–5]). The growing body of research on the temporal integration of audiovisual events has established the ability of human perception to realign sensory signals so that short temporal offsets go unnoticed, with an inherent asymmetry that favours the precedence of visual over auditory signals [e.g., 6]. This buffer is likely in place to avoid perceptual conflicts and to ensure coherent sensory experiences, similar to the ventriloquist effect that compensates for spatial displacements [7], and the McGurk effect [8], which illustrates perceptual strategies for incongruent speech tokens. In this respect, the detection of asynchrony is an informative measure of the limitations of perceptual integration. The window of temporal integration will vary according to both the context [9,10] and the applied experimental methodology [11,12] such that the difference between conditions is often of greater interest than the specific temporal thresholds (e.g., [13,14]). Temporal offsets between simple audiovisual events such as light and sound combinations are typically detected at fairly short offsets [12]. More complex events that involve, for instance, musical instruments, can yield wider windows of temporal integration [10]. Windows of temporal integration observed for audiovisual speech are also found to be fairly robust to asynchrony [3,15]. Figure 1 shows an overview of a range of previously published research with windows of temporal integration for different speech stimuli, derived from different measures. This selection of work illustrates the variations in perceptual tolerance to asynchrony, not only between spoken words [3, 16], sentences [9,10,15,16] and syllables [5], but also across different experimental settings and methodologies (e.g. [9,15,17]). In general, the temporal perception of syllables appears to be less tolerant than that of full words and sentences (e.g. [10]). Furthermore, the temporal thresholds derived from temporal order judgement (TOJ) and simultaneity judgement (SJ) measures could reflect different perceptual strategies [11]; that TOJ measures involve the additional task of determining the order of two signals may make them more demanding, but the focus on precedence may also make then more sensitive [12].

Thresholds for perceived synchrony are of direct relevance when it comes to multimedia platforms (e.g., [18]). For teleconference systems the delay of an audio or video stream can have severe consequences for both the quality [19] and the intelligibility [20] of the perceived speech. Windows of temporal integration can therefore serve as guidelines for system providers to indicate the maximum misalignment that can be tolerated. As mentioned, perception of synchrony depends on the nature of the audiovisual event, but in a live conversation there are bound to be other influences and disturbances that are controlled for in experimental settings. Attention is a likely source of variation; should attention be captured by one modality or engaged by another task, larger temporal offsets may go unnoticed [21]. The current study assesses the detection of asynchrony in live conversations that take place over a teleconference platform, in order to explore the generalisability of temporal integration thresholds derived from isolated audiovisual events. By comparing the teleconference with a more controlled experimental setting, the study aims to shed light on the complexities of a real-life scenario and the consequent predicted potential of increased perceptual tolerance to asynchrony. An SJ measure is used for the experimental setting, as the task of judging synchrony versus asynchrony is more similar to asynchrony detection than to TOJ. Furthermore, by comparing the SJ derivation, with perceived synchrony thresholds at 50 % [3], to the more direct measure of asynchrony detection, the study may also provide insight on the appropriateness of SJ as a broadly applicable methodological approach.

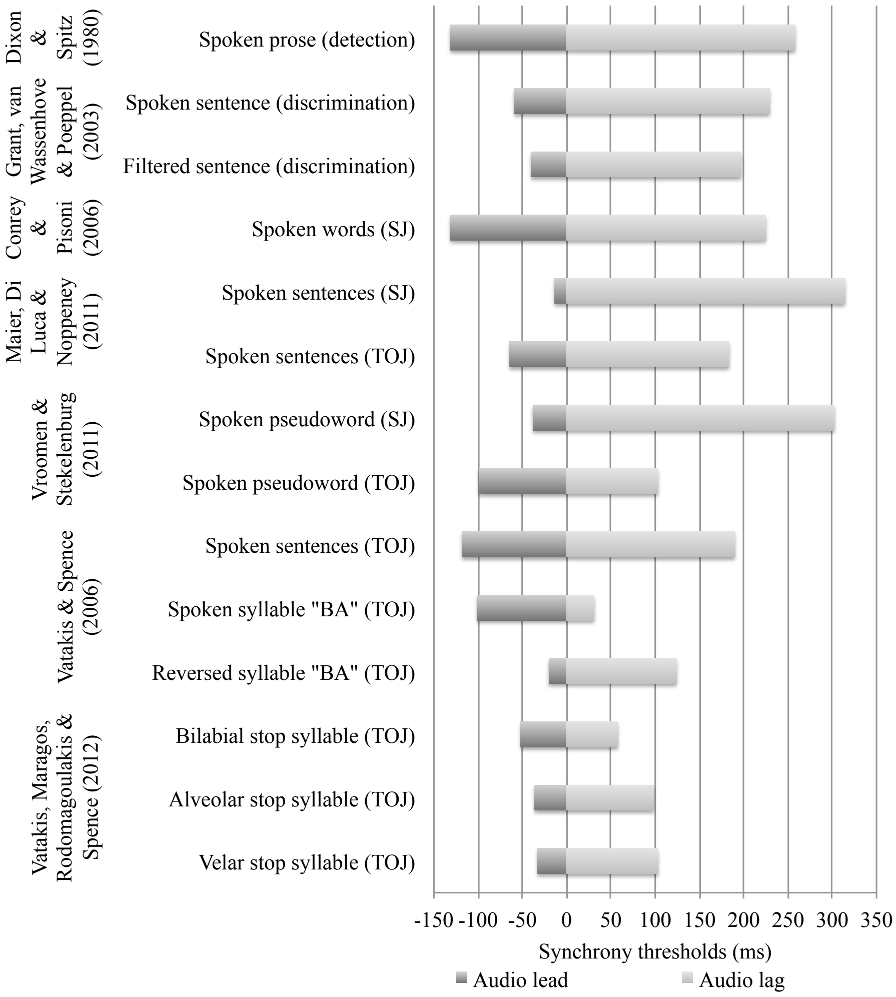

Figure 1: *Windows of temporal integration from earlier studies, established with different experimental methods and for speech stimuli of different complexities [3,5,9,10,15,20,21].*

2. Method

The current study was planned and run as two experiments. The *TelCo* experiment was carried out as a live teleconference between two people engaged in a conversation, whereas the SJ experiment was conducted in a laboratory setting with two television broadcasts used as speech stimuli.

2.1. Participants

The TelCo experiment included 6 male and 4 female adults under the age of 50. They were all Cisco employees who had been asked and agreed to participate. Participants for the SJ experiment were recruited at the University of Oslo, and in total they included 9 females and 11 males between 20 and 38 years old (*M*=25.60, *SD*=4.35).

2.2. Stimuli and procedure

Due to the differences between the two experimental methods, the implemented stimuli and procedures are presented separately.

2.2.1. TelCo asynchrony detection

The TelCo experiment was run like a teleconference, with two volunteers participating in a question game carried out in two of Cisco's teleconference rooms at Lysaker, Norway. One participant, the respondent, sat in the smaller of two rooms, and would draw a ticket that stated the name of a person, animal, object, or place. The other participant, the correspondent, sat in the larger and would try to guess what was printed on the ticket by asking a series of yes/no questions. In addition to the respondent´s task in the conversation, the respondant also focused on the experimental task of assessing the synchrony throughout. Halfway through, the two participants switched rooms and consequently roles.

Two experimenters sat in the same room as the correspondent in order to continuously manipulate the temporal offset between the audio and the video. Audio lead or audio lag asynchrony was introduced in a random order determined beforehand. The offsets increased in steps of 30 ms and the experimenters took care to start each step during a pause in the conversation. The respondent was instructed to raise a hand the moment s/he detected asynchrony. The teleconference system then needed to be re-set with a new dial-up for the next round of gradually introduced asynchrony.

Participants completed eight rounds, four audio-lead and four audio-lag repetitions, before switching roles. Control measurements with a flash- and tone-device were completed twice for every participant, once for audio lead and once for audio lag. This was done to make sure that the introduced offsets corresponded to the audiovisual asynchrony coming through to the small conference room.

2.2.2. Simultaneity judgements for speech

For the SJ experiment, two speech sequences, *News* and *P.M.,* were selected from previously aired television broadcasts on the basis of their differences in shooting angle and movement of the speaker. The *News* sequence shows a female news anchor filmed in studio, while *P.M.* is an excerpt with the Norwegian Prime Minister in a current issues show. Video playback duration was set to 13 seconds, so that the coherence of both sequences was maintained. Audio editing was done with Audacity (2.0.1) [22] and Praat [23], with average audio intensity at 70 dB. Videos were edited in Final Cut Pro (10.0.8), with 1024x576 pixel resolution. Temporal offsets were based on our earlier experiments [e.g., 3] and introduced by displacing the audio track relative to the video track. Audio lead asynchrony was presented at 50 ms, 100 ms, 150 ms, and 200 ms, while audio lag asynchrony was set to 100 ms, 200 ms, 300, and 400 ms. The two ranges of asynchrony levels reflect the asymmetry in perceptual sensitivity to lead and lag asynchrony [6].

The experiment was conducted in a meeting room at the University of Oslo, with videos presented using the Superlab software running on a 2.53 GHz MacBook Pro with a 15” monitor (1440x900 pixel resolution). Two Logitech Z4i satellite speakers (8.5 watts each, >92 dB S/N, 35 Hz - 20 kHz frequency response) were placed on either side of the monitor and participants sat at a distance of approximately 70 cm.

Participants were asked to attend to both the audio and the video and make decisions on whether they perceived them to be synchronous or asynchronous. Responses were collected with a Cedrus RB-530 response-box. The experiment was divided into blocks, in which single instances of stimulus conditions were presented in a random order. As responses could be given at any time, the duration of the experiment varied between individuals, with an upper restriction of 90 minutes including breaks between blocks. Thus the total number of blocks, and thereby also repetitions, varied between 6 and 8 blocks, depending on the rate of progression.

3. Results

Detection times for the asynchrony introduced in the TelCo experiment were averaged across repetitions to establish audio lead and lag thresholds. The point of subjective simultaneity (PSS) was calculated as the mean of the lead and lag thresholds, whereas the window of temporal integration (TI) spans the two thresholds. For the SJ experiment, responses were scored as synchrony match or non-match and proportions were calculated across repetitions of each stimulus condition. Gaussian curves were then fitted individually across the range of temporal offsets. The TI is represented by the full-width at half-maximum, which was calculated from the standard deviation of each curve, while the PSS corresponds to the mean of the curve. From these statistics, the audio lead and audio lag thresholds were also established.

An initial one-way ANOVA assessed the effect of the order of participant roles for the TelCo experiment on the detection of lead and lag asynchrony, as well as the derived TI and PSS. None of the measures were found to differ significantly between the first and second respondents, indicating that the order of experimental tasks did not impact participants’ detection of asynchrony.

The PSS, TI, and lead and lag thresholds for the TelCo and SJ experiments are derived from different procedures, different measures, different participants, and different calculations; the statistical analyses are therefore carried out purely exploratively. To gain some insight into possible variations in the perception of synchrony for the different audiovisual conditions, we ran one-way ANOVAs for lead and lag thresholds, the TI, and the PSS, to compare TelCo, News and P.M. Main effects for PSS [$F(2,47)=16.07$, $p<.001$], audio lead thresholds [$F(2,47)=29.68$, $p<.001$], and audio lag thresholds [$F(2,47)=4.48$, $p<.02$], indicate that temporal perception might not be consistent across the three contexts. The differences between thresholds are plotted in Figure 2; however, Figure 2 also illustrates the similarities in the TI widths. The corresponding lack of an effect for TI [$F(2,47)=1.02$, *ns*] suggests that the overall tolerance for asynchrony is similar across scenarios, although with a directionality that may depend on characteristics of the speech events. The variations in the perception of lead and lag asynchrony are demonstrated by the different thresholds for TelCo, News and P.M, presented in Figure 2. The graph shows that the differences are particularly prominent for lead asynchrony, and that the temporal offsets required for detection are quite long for the TelCo scenario. The PSS plotted in Figure 3 also highlights how subjective synchrony is closest to objective synchrony for TelCo compared to the video sequences. The main effects were further explored with Dunnett C post-hoc analyses, with significant contrasts represented by black arrows in Figures 2 and 3. The greater variance between participants in the SJ experiment, as compared to the TelCo experiment, may explain why the audio lag threshold does not differ significantly between News and P.M. Still, the significant differences between all audio lead thresholds, and two of the PSS contrasts, emphasize the notion that the asymmetry in temporal tolerance may be the major source of variation in responses.

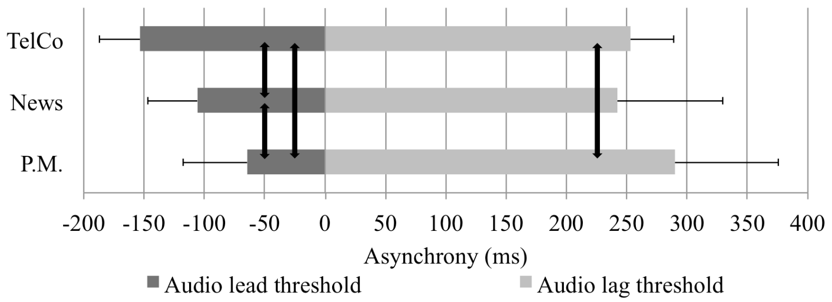

Figure 2: *Windows of temporal integration with audio lead and audio lag thresholds. Thresholds for the TelCo scenario are established from the points of asynchrony detection, whereas thresholds for News and P.M. are calculated from the fitted Gaussian distributions and correspond to synchrony perceived at chance level. The black arrows indicate audio lead and audio lag thresholds that differ significantly from the others, while error bars represent standard deviations. Windows of temporal integration for the three conditions were not significantly different.*

4. Discussion

Earlier studies into the perception of synchrony between auditory and visual events have contributed to a field of research with widely varying thresholds of temporal integration (e.g., [9,10,20]). Considering the differences found when comparing methodologies and stimulus complexities [12], and the assumed contribution of attention [21], the results from this study are surprisingly consistent across conditions. The windows of temporal integration did not differ significantly across the three experimental scenarios in the current study, implying that the perceptual tolerance to temporal offsets is more constant than we originally expected. On the other hand, significant differences between all audio lead thresholds point to a bias in the direction of temporal integration of different audiovisual events. This skew is also reflected in the difference between audio lag thresholds when comparing TelCo and P.M., and in the PSS contrasts between TelCo and P.M., as well as between News and P.M. In other words, the overall tolerance to audiovisual asynchrony is more or less the same for the teleconference and the two speech sequences, but the directionality sets them apart.

The distinctly greater tolerance to audio lead asynchrony observed for the teleconference scenario, compared to the speech sequences, could possibly reflect the predicted complexities of a real-life scenario. The impact from the added attentional demand of the question game might only be manifested for asynchrony where the margins for detection were already narrow. Given that temporal perception is especially sensitive to auditory signals that precede visual signals [6], the disturbances attributable to the teleconference could have contributed to a greater perceptual tolerance in this direction. If so, we deduce that temporal integration may be even more robust in the busy surroundings of everyday life than is generally found in experimental settings.

As for the two speech sequences from the SJ experiment, we surmise that the intentional choice of video content with different speaker characteristics and shot angles may have affected the perceptual sensitivity to lead and lag asynchrony. While one speaker is positioned face forward, the other is facing sideways and viewed at an angle. The viewing angle of a speaker can indeed influence performance on speech recognition tasks, at least in the presence of distractions [25]. Thus temporal speech cues are also likely affected by the visibility of a speaker's face. Moreover, the clarity of the acoustical phonemes, along with the prominence of the speech movements, is also likely to contribute to the accuracy of participants' simultaneity judgements.

Overall, the temporal integration of audiovisual speech shows a remarkable resilience to asynchrony. With the narrowest window of temporal integration approaching 350 ms, the potential leeway available for teleconference platforms is quite remarkable. Our results correspond to previously published thresholds established for continuous speech stimuli [9,15,20]. Although TOJ measures tend to yield more narrow windows of temporal integration [10], particularly for short speech segments [5], SJ and detection tasks can be argued to provide more ecologically valid measures of synchrony perception [12]. Based on the results from the current study and related works, our most conservative recommendations to developers of teleconference platforms, and similar systems, would therefore be to ensure that thresholds do not exceed 50 ms for audio lead asynchrony and 200 ms for audio lag asynchrony. Within this window of temporal integration, perception can compensate for the temporal misalignment and asynchrony is unlikely to be noticed.

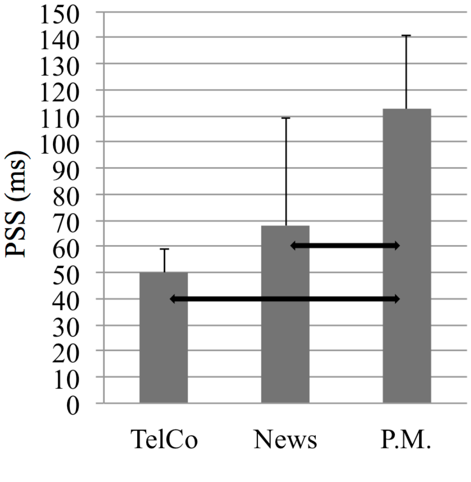

Figure 3: *PSS for the three speech scenarios, calculated as the mid-point between lead- and lag-thresholds for TelCo and represented by the mean of the Gaussian distribution for News and P.M. Significant differences in PSS are highlighted by the black arrows and standard deviations are shown as error bars.*

5. Acknowledgements

The authors would like to thank the team at Cisco Norway for the rewarding collaboration.

6. References

[1] I. J. Hirsh and C. E. Sherrick, “Perceived order in different sense modalities,” *Journal of Experimental Psychology*, vol. 62, no. 5, pp. 423–432, Nov. 1961.

[2] J. A. J. Roufs, “Perception lag as a function of stimulus luminance,” *Vision Research*, vol. 3, pp. 81–91, 1963.

[3] B. Conrey and D. B. Pisoni, “Auditory-visual speech perception and synchrony detection for speech and nonspeech signals,” *The Journal of the Acoustical Society of America*, vol. 119, no. 6, pp. 4065–4073, 2006.

[4] K. W. Grant, V. Wassenhove, and D. Poeppel, “Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony,” *Speech Communication*, vol. 44, no. 1–4, pp. 43–53, Oct. 2004.

[5] A. Vatakis, P. Maragos, I. Rodomagoulakis, and C. Spence, “Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual temporal perception,” *Frontiers in Integrative Neuroscience*, vol. 6, no. 71, pp. 1–18, Jan. 2012.

[6] J. Vroomen and M. Keetels, “Perception of intersensory synchrony: A tutorial review,” *Attention, Perception & Psychophysics*, vol. 72, no. 4, pp. 871–884, 2010.

[7] C. E. Jack and W. R. Thurlow, “Effects of degree of visual association and angle of displacement on the ‘ventriloquism’ effect,” *Perceptual and Motor Skills*, vol. 37, pp. 967–979, 1973.

[8] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[9] N. F. Dixon and L. Spitz, “The detection of auditory visual desynchrony,” *Perception*, vol. 9, pp. 719–721, 1980.

[10] A. Vatakis and C. Spence, “Audiovisual synchrony perception for music, speech, and object actions,” *Brain Research*, vol. 1111, no. 1, pp. 134–142, Sep. 2006.

[11] A. Vatakis, J. Navarra, S. Soto-Faraco, and C. Spence, “Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments,” *Experimental Brain Research*, vol. 185, no. 3, pp. 521–529, Mar. 2008.

[12] R. L. J. van Eijk, A. Kohlrausch, J. F. Juola, and S. van de Par, “Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type,” *Perception & Psychophysics*, vol. 70, no. 6, pp. 955–968, Aug. 2008.

[13] K. Petrini, M. Russell, and F. Pollick, “When knowing can replace seeing in audiovisual integration of actions.,” *Cognition*, vol. 110, no. 3, pp. 432–439, Mar. 2009.

[14] J. Navarra, A. Alsius, I. Velasco, S. Soto-Faraco, and C. Spence, “Perception of audiovisual speech synchrony for native and non-native language.,” *Brain Research*, vol. 1323, pp. 84–93, Apr. 2010.

[15] K. W. Grant, V. van Wassenhove, and D. Poeppel, “Discrimination of auditory-visual synchrony,” in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2003, pp. 31–35.

[16] J. Vroomen and J. J. Stekelenburg, “Perception of intersensory synchrony in audiovisual speech: Not that special,” *Cognition*, vol. 118, pp. 75–83, 2011.

[17] J. X. Maier, M. Di Luca, and U. Noppeney, “Audiovisual asynchrony detection in human speech.,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 1, pp. 245–256, Feb. 2011.

[18] N. Miner and T. Caudell, “Computational requirements and synchronization issues for virtual acoustic displays,” *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 4, pp. 396–409, Aug. 1998.

[19] R. Steinmetz, “Human perception of jitter and media synchronization,” *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61–72, 1996.

[20] K. W. Grant and S. Greenberg, “Speech intelligibility derived from asynchronous processing of auditory-visual information,” in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2001, no. 1, pp. 132–137.

[21] L. B. Stelmach and C. M. Herdman, “Directed attention and perception of temporal order,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, no. 2, pp. 539–550, 1991.

[22] Audacity Team, “Audacity.” 2012.

[23] P. Boersma and D. Weenink, “Praat: doing phonetics by computer.” 2012.

[24] R. Eg and D. M. Behne, “Short and sweet, or long and complex? Perceiving temporal synchrony in audiovisual events,” *Seeing and Perceiving*, vol. 25, p. 96, Jan. 2012.

[25] T. R. Jordan and S. M. Thomas, “Effects of horizontal viewing angle on visual and audiovisual speech recognition,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 6, pp. 1386–1403, 2001.