The 14th International Conference on Auditory-Visual Speech Processing
25-26 August 2017, Stockholm, Sweden

Perceived Audiovisual Simultaneity in Speech by Musicians and Non-musicians: Preliminary Behavioral and Event-Related Potential (ERP) Findings

Dawn M. Behne[1], *Marzieh Sorati*[1], *Magnus Alm*[1]

[1]Department of Psychology, Norwegian University of Science and Technology, Norway
dawn.behne@ntnu.no, marzieh.sorati@ntnu.no, magnus.alm@ntnu.no

Abstract

In audiovisual simultaneity perception, audio-lead and video-lead are generally considered fundamentally different, despite both being occurrences of physical misalignment. The current study pursues this difference, in a preliminary study comparing musicians with non-musicians across audio-lead, synchronous and video-lead alignments in behavioral and ERP simultaneity judgment tasks. Results to date are consistent with the conclusion that musicians are more sensitive to audiovisual asynchrony than non-musicians, in particular for audio lead and highlight the role of experience in facilitating sensory processing.

Index Terms: simultaneity judgment (SJ), event-related potential (ERP), musicians, facilitation

1. Introduction

Most of our experiences involve at least two of our senses and our ability to integrate the very distinct information from sound and light leads to benefits well beyond the individual sensory information alone.

Perception of audiovisual synchrony relies on matching temporal attributes across sensory modalities. The unity assumption suggests that when that information reaches the different senses, the more properties they have in common, such as occurring close in time, the more likely the brain is assumed to treat them as having a single source (e.g., [1]). Precision in relating audio and visual information to a mutual source decreases neural processing time [2] and gives a perceptual benefit (e.g., [3], [4]).

The relative alignment of the audio and video signals has differing effects: to be perceived as asynchronous, a greater physical synchrony of video preceding audio is needed compared to audio preceding video, where a smaller audiovideo misalignment is needed for the asynchrony to be perceived (e.g., [5], [6]). The asymmetry of subjective perception around the point of audiovisual synchrony has generally been ascribed to perceptual accommodation to differences in the propagation speeds of sound and light and their corresponding neural processing times for the different senses (e.g., [7]). Furthermore, since articulatory movement precedes the speech signal [8], the visual information may provide the perceiver with a predictor for the auditory signal [9].

Previous research has shown individual differences in the perceived simultaneity in audiovisual speech perception (e.g., [3]) and that effects from training can be lasting (e.g., [10]). In particular music experience has been shown to shape temporal binding of auditory and visual signals [11], and to specifically lead to a great sensitivity to audio-lead [12]

It is generally acknowledged that the fundamental difference in how natural audio-lead and visual lead are generated inevitably result in diverse consequences for how availability of visual information and its timing potentially facilitates auditory processing. In the current study, highly skilled musicians were compared with non-musicians in behavioral and ERP audiovisual simultaneity judgment tasks to study the facilitation on audio processing. Musicians are expected to be more sensitive to audiovisual asynchrony relative to non-musicians, inn particular, have an increased sensitivity to an audio signal preceding the video.

2. Method

In an on-going study of audiovisual synchrony perception, skilled musicians were compared with non-musicians across audio-lead, synchronous and video-lead alignments. For both groups behavioral responses were logged in an audiovisual simultaneity task to evaluate participants´ perceived audiovisual synchrony judgments and in a separate task, EEG was recorded to compare auditory N1 across groups for audio-lead, audiovisual synchronous and video-lead speech materials.

2.1. Participants

Two groups of young adult NTNU students participated in the study: 6 musicians (M=21 yrs, 4 males) and 8 non-musicians (M=23 yrs, 7 males). All were native-speakers of Norwegian, right-handed based on a variant of the Edinburgh Handedness Inventory [13] had normal to corrected visual acuity (Snellen test) and normal bilateral hearing acuity (≥20 dB audiometric thresholds for 250 to 4000 Hz, [14].

2.1.1. Musicians

The musicians were all students at the Department of Music, NTNU, having met strict criteria on theoretical and practical musical evaluations in addition to advanced musical skills on a primary and secondary musical instrument. None of the musicians were singers or dancers.

On average the musicians started playing a musical instrument when they were 8 years old (range: 6-11 yrs), had at the time of the study been playing regularly for an average of 13 years (range: 9-16 yrs), and were playing a musical instrument approximately 12 hours (range: 4-30 hours) per week. The musicians also self-reported a very strong interest in music, with an average of 9.0 points on a 10-point scale.

10.21437/AVSP.2017-19

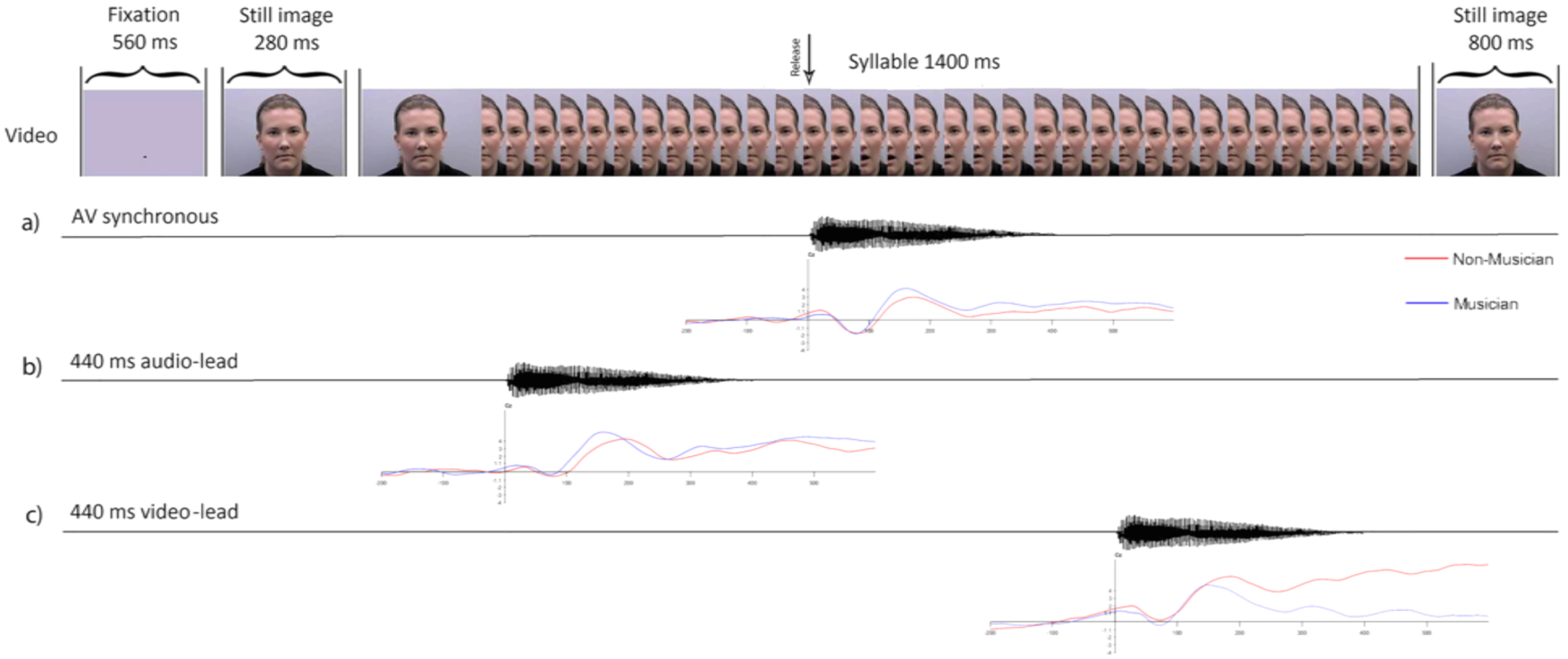

Figure 1: The trial timeline for the video stream shows a fixation cross followed by a still image and the speakers face during articulation of the syllable /ba/. Below, the audio stream and corresponding ERP are shown for the (a) AV synchronous condition, (b) 440ms audio-lead condition, and (c) 440ms video-lead condition.

2.1.2. *Non-musicians*

Non-musicians had no more than the one year of weekly musical experience required in Norwegian elementary schools. They had not learned to play a musical instrument or, for example, sung in a choir, or participated in similar musical activities. On a 10-point scale of music interest, non-musicians self-reported a neutral (average 6 points) interest in music.

2.2. Simultaneity judgment behavioral task

2.2.1. *Stimuli*

The stimulus set for the behavioral experiment used the audiovisual syllable /ba/ and consisted of one audiovisual synchronous stimulus and 22 audiovisual asynchronous stimuli. The audiovisual asynchronous stimuli were created by keeping the video signal constant while moving the auditory signal so that the visual consonant release would either precede or follow the auditory consonant release, thus creating 22 audiovisual asynchronous stimuli ranging in 40ms increments from 440ms audio-lead to 440ms video-lead.

2.2.2. *Procedure*

The behavioral experiment was carried out on an iMac 11.3 (1920x 2000 pixels) in the Speech Lab, Department of Psychology, NTNU. Alignment of the audio and video streams was measured using Black Box Toolkit (Black Box Toolkit Ltd., England) and this difference was accommodated when preparing the stimuli.

Trial presentation and data collection were carried out using Superlab (v.5). The 23 stimuli differing in audiovisual alignment were randomly presented once in each of four blocks, for a total of 92 trials.

During the experiment the participant was seated ca 70 cm from the computer where the visual stream was presented in the center of the screen. The audio stream was presented binaurally over AKG K273 studio headphones at 68 ±1 dBA. The participant´s task was to keep focused on the center of the monitor and indicate as quickly as possible, using a Cedrus RB730 response box, whether the perceived the audio and visual components of the stimulus to synchronized or asynchronous. Between blocks participants had two 30s breaks and one 60s break midway in the experiment.

Two versions of the experiment were prepared to counter-balance the left-right placement of the two response buttons.

2.3. Simultaneity judgment ERP task

2.3.1. *Stimuli*

The stimuli in the ERP experiment included a sample of the stimuli used in the behavioural experiment. In addition to an AV synchronous stimulus, the sample included 440ms audio-lead and 440ms video-lead.

2.3.2. *Procedure*

The experiment took place in a dimly lit sound-attenuated room in the NTNU Speech Lab. A participant sat with her/his head supported by a chinrest positioned approximately 190 cm from a 40-in Samsung SyncMaster 400DX-2 (1720x1200 pixel) monitor via which the visual stream (MPEG4) was presented. The audio stream was presented via Etymotic Research ER1 insert earphones at ca 60dBA. Alignment of the audio and video streams was measured using an EGI Audio/video timing device (Electrical Geodesics, Oregon, USA) which then formed the basis for compensatory adjustment prior to data analyses.

Stimuli were presented using Psychtoolbox-3 where each of the three audiovisual stimuli (synchronous, 440ms audio-lead and 440msvideo-lead) was randomly presented 10 times in each of 15 blocks, for a total of 150 presentations of each stimulus. Participants had 30s breaks between blocks and two 3min breaks in the course of the experiment. As illustrated by the trial timeline in Figure 1, each trial included a 1400ms audiovisual syllable, preceded by a 560ms grey screen with a fixation cross and a 280ms still image of the speaker's face. The same still image was also presented for 800ms after the audiovisual syllable. The fixation cross on the grey background was located at the position that the mouth of the speaker would later emerge. The still image was the same as the first frame of the AV syllable segment. The total duration of each trail was 3040ms.

The participant´s task was to focus on the fixation cross and then listen and watch the syllable. Participants gave no response.

During the experiment EEG was recorded at 1000 Hz sampling rate with 128 channel dense array EEG system (Electrical Geodesics, Oregon, US). No online filter was applied and Cz was the reference.

EEG data was exported to MATLAB R2016b with ERPLAB v.6.1.3 extension [15] and re-referenced off-line to the average reference and bandpass filtered (0.1-30 Hz, 12 dB/octave). The EEG data were segmented to 800ms epochs (-200, 600) and the baseline was corrected based on the -200ms prestimulus period. ERPs were time-locked to the onset of the audio signal (Figure 1). A step function was used for artifact detection [16]. ERPs were averaged separately for the three stimuli (synchronous, 440ms audio-lead and 440ms video-lead). Visual inspection of the ERP scape map showed that the auditory N1 had most activity in the region near Cz, which was therefore used for the analysis. Auditory N1 peak latency was scored in the 55-93ms window for musicians and the 65-111ms window for non-musicians.

3. Results

3.1. Behavioral responses

Data from each participant were plotted with the percentage of responses as a function of audiovisual alignment using Sigma Plot (v.12). A Gaussian curve was fit to the data where the point of subjective simultaneity (PSS), audio-lead threshold (ALT), video-lead threshold, and the full width of the half maximum (FWHM) were identified ([5][6]]. The PSS is defined as the x-value at the peak of the Gaussian curve. ALT is the x-value for audio-lead where the y-value is 50%, and VLT is the corresponding value for video-lead. The FWHM is the absolute value of the summation of ALT and VLT. Figure 2 shows curves based on data from the musicians and nonmusicans.

For PSS, ALT, VLT and FWHM an independent-samples t-test was carried out comparing musicians and non-musicians. With so few participants, not surprisingly, no reliable differences were observed for any of the four dependent variables. However, results for ALT suggest the expected

pattern, with musicians (M=-169m) having a mean ALT closer to physical synchrony than non-musicians (M=-196ms).

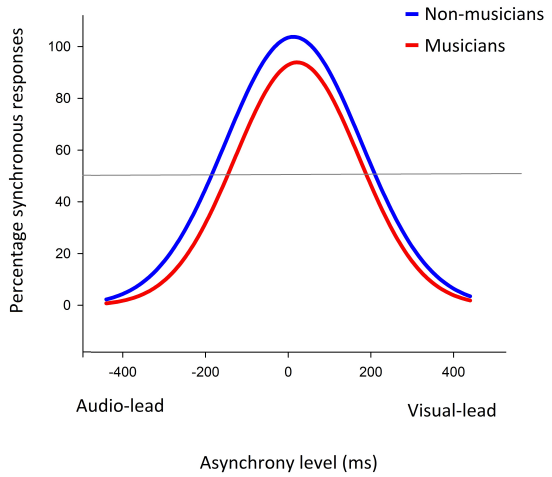

Figure 2: *Normalized simultaneity responses for musicians and non-musicians and plotted as a function of audiovisual alignment.*

3.2. Auditory N1 peak latency

For the 440ms audio-lead, synchonous, and 440ms video-lead conditions, an independent-samples t-test compared the auditory N1 peak latency for musicians and non-musicians. Figure 1 shows the ERPs for each of the three conditions, respectively in (a), (b) and (c). Means are presented in Figure 3.

As can be seen in Figure 3, the auditory N1 peak latency for nonmusicans is the same for the 440ms audio-lead and synchronous conditions whereas for musicians auditory N1 peak latency is smaller the 440ms audio-lead condition. Results for the 440ms audio-lead condition show a significantly shorter auditory N1 peak latency for musicians (M=75ms, SD=10) compared to non-musicians (85ms, SD=11) [t(12)=0.048]. No differences in auditory N1 latency based on musical experience were observed for the synchronous condition or for the 440ms video-lead condition.

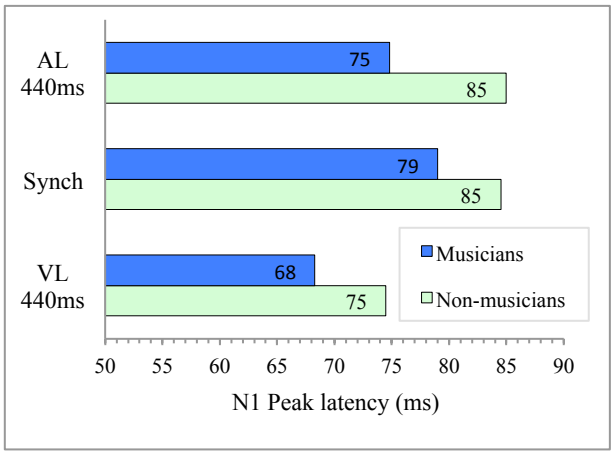

Figure 3: *Mean auditory N1 peak latency for musicians and non-musicians in the 440ms audio-lead (AL), synchronous, and 440ms video-lead (VL) conditions.*

4. Discussion and Conclusions

Behavioral and ERP simultaneity judgment tasks were carried out to study the facilitation on audio processing.

Although results from the behavioral task did not show a significant difference between musicians and non-musicians for ALT, the number of participants in each group was small compared to our previous study [12]. Nevertheless pattern of results consistent with Behne et al, 2013 [12] emerged, showing musicians to have an ALT closer to physical synchrony than non-musicians. A larger sample will determine the reliability of this observation.

Previous ERP research has shown an auditory facilitation with audiovisual stimuli [2]. The current study demonstrates an auditory facilitation of a different kind; in this case, rather than facilitation from other sensory information, we observe facilitation based on experience. Findings from the ERP experiment using the same small number of musicians and non-musicians, revealed a reliable difference in auditory N1 peak latency. Results show faster early audio processing by musicians than non-musicians for the audio-lead condition, but no difference when the audio and video streams are physically synchronous or for the video-lead condition.

These findings are consistent with musicians having greater sensitivity to audio-lead than non-musicians. They findings highlight the difference in sensitivity to audio-lead and video-lead (e.g., [17]) and reinforce that audio-lead and video-lead have different consequences for audio processing.

5. Acknowledgements

We thank the willing participants in the study and the Department of Psychology, NTNU which provided funding for the project (2016, 2017).

6. References

[1] Welch R.B., Warren D.H. (1980). Immediate perceptual response to intersensory discrepancy. Psychological Bulletin, 88, 638– 667.

[2] van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. PNAS, 102(4):1181–1186.

[3] Grant, K., and Seitz, P. (1998). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. 104, 2438–2450.

[4] Sumby, W. H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. 26, 212–215.

[5] Conrey, B., and Pisoni, D. B. (2006). "Auditory-visual speech perception and synchrony detection for speech and nonspeech signals," J. Acoust. Soc. Am. 119, 4065-4073.

[6] Hay-McCutcheon, M., Pisoni, D., and Hunt, K. (2009). "Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis," Int. J. Audiol. 48, 321-333.

[7] Keetels, M., and Vroomen, J. (2012). "Perception of synchrony between the senses," in Frontiers in the Neural Bases of Multisensory Processes, edited by M. M. Murray and M. T. Wallace, (Taylor and Francis Group, London), pp. 147-177.

[8] Smeele, P. M. T., Sittig, A., van Heuven, V. (1994): "Temporal organization of bimodal speech information", In *ICSLP-1994*, 1431-1434.

[9] Grant, K., Greenberg, S., Poeppel, D., and van Wassenhove, V. (2004). "Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing," Semin. Hear. 3, 241–255

[10] Powers, A. R. 3rd., Hillock, A. R., and Wallace, M. T. (2009). “Perceptual training narrows the temporal window of multisensory binding,” *Journal of Neuroscience, 29*, 12265–12274.
[11] Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses, P. Natl Acad Sci USA, 108, 1441-1450.
[12] Behne, Dawn Marie; Alm, Magnus; Berg, Aleksander; Engell, Thomas; Foyn, Camilla; Johnsen, Canutte; Srigaran, Thulasy; Torsdottir, Ane E.. (2013) Effects of musical experience on perception of audiovisual synchrony for speech and music. Journal of the Acoustical Society of America. vol. 133.
[13] Oldfield, R. (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9, 97-113.
[14] British Society of Audiology. (2004). “Recommended procedure: Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels,” http://www.thebsa.org.uk/docs/RecPro/PTA.pdf. Last viewed June 3, 2010.
[15] Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. Frontiers in human neuroscience, 8, 213.
[16] Luck, S. J. (2014). An introduction to the event-related potential technique. MIT press.
[17] Alm, Magnus; Behne, Dawn Marie. (2013) Audio-visual speech experience with age influences perceived audio-visual asynchrony in speech. Journal of the Acoustical Society of America. vol. 134 (4).