Is Text the New Lipreading? Audiovisual Integration Effects of Speech and Syllables

Joëlle Mouwen

SNR: 2023739

Bachelor of Psychology

Department of Cognitive Neuropsychology, Tilburg University

Dr. Martijn Baart and S. Faezeh PourHashemi

Dr. Marcel Bastiaansen

30th of June, 2021

Abstract

Previous studies showed that the N1 and P2 components get attenuated and sped up when auditory information is accompanied by concordant visual lipreading information. The present study aimed to find out whether similar audiovisual integration effects would occur when replacing the lipreading cues with text and in what amount stimulus asynchrony onset affects the neural responses by adding predictive value. The main expectation of this research was that the attenuation and speeding effect of the N1 and P2 components will also be present with written text, provided that text has some predictive value about the upcoming speech sound. Electroencephalography (EEG) and the recording of Event Related Potentials (ERPs) were used to investigate this hypothesis. Twenty-three Dutch students (5 male, 18 female) from 18 to 35 years old participated. The experiment consisted of visual-only (V), auditory-only (A) and audiovisual (AVtext) stimuli, which were presented in Stimulus Onset Asynchrony (SOA) conditions of 0 and 300 ms. Brain Vision Analyzer was used to preprocess the EEG data. Repeated measures ANOVA revealed a significant enhancement of the P2 in the 0 ms SOA and a significant attenuated amplitude in the 300 ms SOA, while no significant effects were found for the N1 nor latency. The enhanced response in the P2 is not in line with the main expectation, but can be explained by supra-additivity, while the attenuated response in the 300 ms SOA did partly confirm the main hypothesis for finding a similar audiovisual interaction effect with text as with lipreading.

Keywords: audiovisual integration, speech, text, Electroencephalography, Event Related Potentials, Stimulus Onset Asynchrony

Is Text the new Lipreading? Audiovisual Integration Effects of Speech and Syllables

Due to the Covid-19 pandemic and the wide use of face masks, we tend to have more difficulty with interpersonal communication than ever. Muffled voices do not seem to be sufficient for interpersonal comprehension and we do not have the possibility to seek support in the articulatory mouth movements of the speaker (Mheidli et al., 2020). Especially individuals with hearing loss experience serious difficulties in comprehending others due to the lack of lip reading cues (Chodosh et al., 2020). This can be explained by the fact that multisensory integration in general is an important feature of our perceptual system. Common findings in related studies are that multimodal presented objects tend to get identificated and recognised faster than unimodal perceived objects (Hershenson, 1962) and that lip reading is not only crucial for the hard of hearing, but also plays an essential role during speech perception in general. Especially in noisy listening conditions, seeing the moving mouth of a speaker drastically improves intelligibility (Sumby & Pollack, 1954). This means that visual information even has an impact on perception of non-distorted speech, without us even noticing (Callan & Jones, 2003). All in all, it has become clear that lip reading contributes to comprehension of speech and identification of words and syllables (Calvert et al., 1997).

Sams et al. (1991) already showed that the visual information from articulatory movements activates the auditory cortex and in this way has the ability to affect the processing of speech sounds. An example would be the McGurk effect, which provides effective evidence that lip reading has the ability to change our auditory perception (McGurk & MacDonald, 1976). Calvert et al. (1997) investigated the activation of the brain in combination with silent lip reading, which led to the finding that visible pseudo speech does activate the auditory cortex, but closed-mouth movements do not. Another example of this is the ventriloquist illusion, also called ‘perceptual fusion’, which makes us tend to attribute auditory speech to a particular source, even when this is not the actual producer of the sound.

Mouth movements that match the auditory speech temporally and spatially can trick us into perceiving otherwise (Bertelson & Radeau, 1981). According to Calvert et al. (2002), the superior temporal sulcus (STS) plays the largest role in audiovisual speech integration. The inferior frontal regions, premotor cortex, right superior temporal gyrus and anterior cingulate gyrus also showed significant responses.

Involvement of the motor system has been further emphasized by Skipper et al. (2007), who suggested that even the conventional mirror system participates in audiovisual speech perception. Mirror neurons are a specific kind of neurons, found in macaques, that fire when observing someone else's movements and when performing similar movements themselves. The mirroring functionality of mirror neurons would be present in the motor system (Rizzolatti and Craighero, 2004). The theory is that automatic mirroring functions could get activated in multiple motor areas during audiovisual speech perception, like lip reading. Observing someone else's mouth movements would carry most responsibility to this effect, compared to observing auditory information only (Skipper et al., 2007).

Electroencephalography (EEG) studies and the recording of Event Related Potentials (ERPs) have shown that the neural activity of speech sound processing gets attenuated and sped up when a perceived spoken word is accompanied by concordant lipreading information. ERPs are EEG changes that occur in response to specific sensory, motor or cognitive events (Blackwood & Muir, 1990). This suppressing and speeding effect is visible in the N1 and P2 components of auditory evoked potentials (AEP). The N1 or N100 component refers to a negative peak between 90 and 200 milliseconds (ms) after the stimulus onset. This component occurs when an unexpected stimulus is observed. The P2 or P200 component is a positive peak between 100 and 250 ms after the stimulus onset (Blackwood & Muir, 1990).

The attenuation and speeding up of the N1 and P2 seem to occur because visual articulatory information precedes the auditory information. This anticipatory motion provides

predictability about the auditory stimulus onset (Wassenhove et al., 2005). Stekelenburg and Vroomen (2007) also showed that audiovisual interaction was present when preceded by anticipatory motion. However, Baart (2016) explained that the attenuations and speeding effects of N1 and P2 are not always observed and reported. In addition, Van Wassenhove et al. (2005) described supra-additivity as a principle mechanism for audiovisual integration, which refers to an increased response to simultaneously presented events instead. Other recent findings, however, have suggested that sensory-specific brain regions are responsive to input presented through different modalities. Hereby, Giard & Peronnet (1999) distinguished between subjects who were better at identifying objects based on visual cues and based on auditory attributes. They showed that the addition of auditory cues to visual stimuli led to enhanced responses in the visual cortex with subjects who are better in auditory tasks and the addition of a visual cue to auditory stimuli led to enhanced responses in the auditory cortex with subjects who are better in visual tasks. In short, multimodal integration seems to induce increased neural responses in the brain area of the non-dominant sensory modality.

Massaro et al. (1996) have already shown that we are naturally tolerant to visual-first asynchronies in speech, while we are specifically sensitive to auditory-first asynchronies. Because preceding auditory information is not representative for real-life speech perception, since anticipatory motion naturally precedes auditory information, we tend to notice this difference instantly (Czap, 2011). Although the main assumption refers to multimodal integration as an automatic process, Alsius et al. (2005) found that attention is actually necessary for multisensory binding.

According to Baart (2016) the N1 component seems to be especially dependent on the temporal predictability of the preceding auditory stimulus onset and the P2 component is not. The N1 is also unaffected by congruence of audiovisual stimuli, while the suppression of the

P2 component was larger with incongruent stimuli than with congruent ones (Stekelenburg & Vroomen, 2007).

However, lip reading is obviously crucial in everyday communication, but this is not the only visual stimulus that can be involved in audiovisual speech perception. Stekelenburg and Vroomen (2007) showed that audiovisual integration is not speech specific at all. Their research about audiovisual speech perception used multiple different non-speech stimuli like pictures of objects and videos of voluntary actions. One non-speech stimulus that has not been addressed much in research when combined with auditory speech, is text. Just like lip reading, the perception of text involves the processing of visual stimuli. The ventral as well as the dorsal visual streams are necessary in order to read words. The ventral stream is largely responsible for turning written words into mental representations, whereas the dorsal stream plays a large part in the conversion of letters and words into sounds and adding semantical value (Borowski et al., 2006). Raij et al. (2002) conducted research about audiovisual letter/speech integration and found that, besides the sensory and auditory projection areas, the superior temporal sulcus is mainly responsible for audiovisual integration. The fact that the STS also showed stronger interactions with congruent than with incongruent letters, supports this.

As discussed earlier, the suppression and speeding up in the N1 component occurs when visual stimuli precede auditory stimuli and cause the possibility to predict the auditory stimulus (Stekelenburg & Vroomen, 2007). According to Stekelenburg and Vroomen (2010) the effect in N1 did specifically depend on the presence of anticipatory motion. When it comes to text, the research of Froyen et al. (2008) describes that single letter speech integration is also highly affected by stimulus onset asynchrony (SOA). Specifically, a 300 ms SOA leads to a decrease in ERP amplitude (in that case, it was the Mismatch negativity, or MMN), just like it does with lipreading.

The findings of Froyen et al. (2008) indicate a possibility that lipreading and text both show similar behavioral effects. However, it is yet unknown whether the present audiovisual integration effect on the N1 and P2 components can also be induced using text. Hereby, it would be new and interesting to replace the more common visual lip reading stimulus by the written text of a spoken syllable. This leads to the following research question: does audiovisual speech perception with a written syllable as visual stimulus induce similar audiovisual integration effects in the N1 and P2 components as it does with lipreading stimuli? Being able to answer this question would extend our knowledge from the edge of single letter speech integration and clear out a path for future research in the field of textual audiovisual integration.

Based on the studies mentioned earlier, the main expectation of this research is that the attenuation and speeding effect of the N1 component will be present with written text, provided that text has some predictive value about the upcoming sound. As Stekelenburg and Vroomen (2007) also showed, the P2 component seems to be content-dependent due to its larger attenuation with incongruent stimuli. Therefore the second expectation is that P2 will get attenuated and speeded up as well, but it would not show a remarkably strong suppression since the audiovisual stimuli in the experiment will be congruent only. The fact that single letter speech integration is also influenced by SOA (Froyen et al., 2008), leads to the third expectation of a similar outcome in the present study. To examine these hypotheses and explore the temporal features of multisensory speech perception, an EEG study will be executed. Spoken and written syllables will be used as auditory and visual stimuli and the results will be compared to findings that are obtained using lipreading stimuli (but are not collected in current experiment). According to Baart and Samuel (2015) lipreading and lexical context operate simultaneously, but function separately. Hereby, the present study will be using non-lexical syllables only. Anticipatory motion will be simulated by presenting the

visual stimulus earlier than the auditory stimulus (in a 300 ms stimulus onset asynchrony). This will be compared to a condition in which the auditory and visual stimuli will be presented simultaneously, where text has no (temporal) predictive value about the sound.

Methods

Participants

Twenty-three Dutch listeners (5 male, 18 female) with regular hearing and normal or corrected-to-normal vision participated in exchange for two participant credits. They were all students at Tilburg University, with an age range from 18 to 35 and a mean of 20,17. All the participants gave their written informed consent in advance of the experiment. The study is approved by the local ethics committee of Tilburg University.

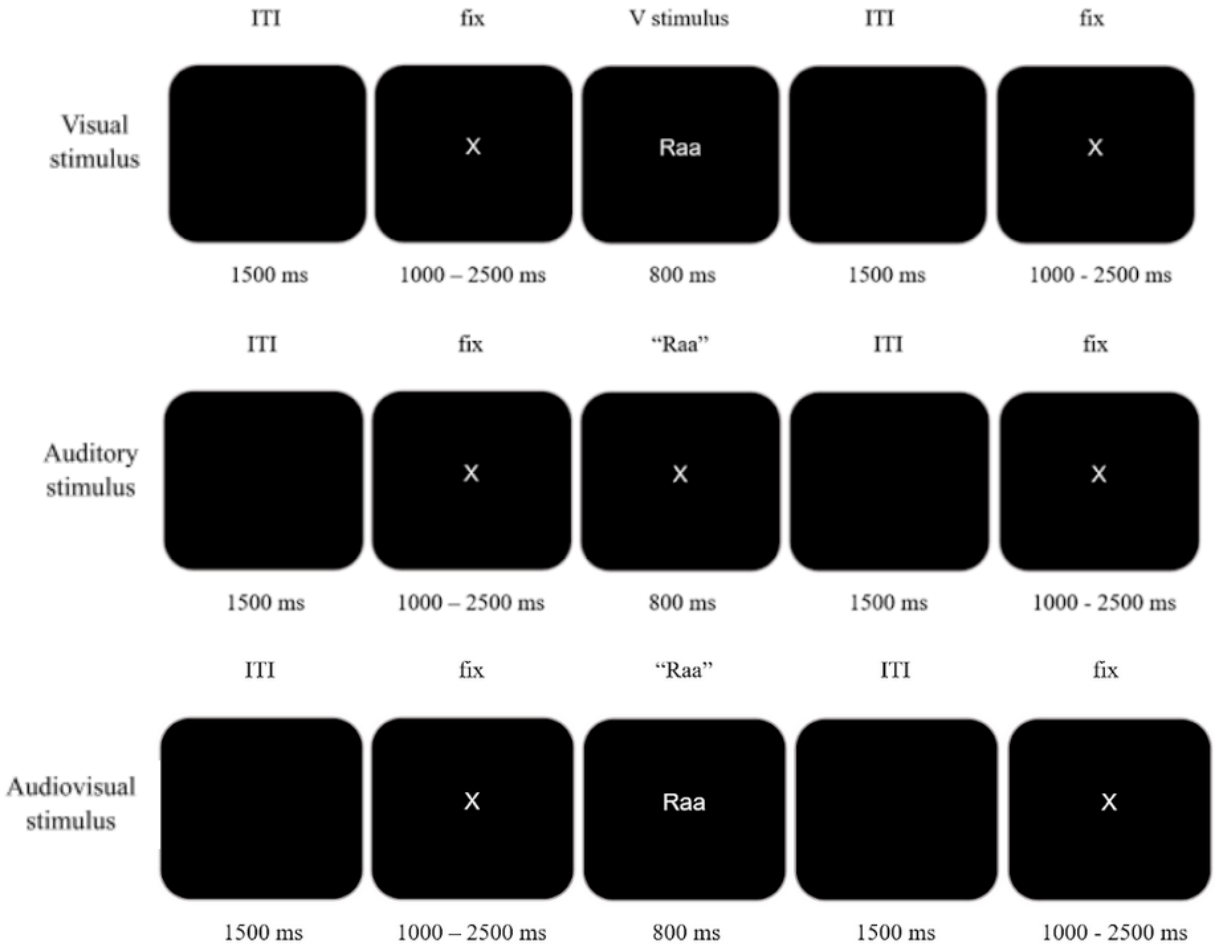
Measures

The experiment consisted of visual-only (V), auditory-only (A) and audiovisual (AV) stimuli. The stimuli involved seven different spoken and written syllables (/Daa/, /Faa/, /Kaa/, /Laa/, /Maa/, /Paa/, /Raa/) which were presented in all conditions (A/V/AV). The task of the participant was to push a random button after seeing or hearing /Raa/.

The visual text (Arial font, size 60) was centered on the screen and presented in white on a black background. As also shown in Figure 1, a fixation point (Arial font, size 40) was presented preceding the visual stimuli, with a randomly varying duration of 1000 – 2500 ms. During the auditory-only stimulus, the fixation point was visible for 800 ms. The Inter Trial Interval (ITI) was set at 1500 ms. The text was presented on a 25-in monitor (BenQ Zowie XL 2540, 240 Hz refresh rate), which was positioned at eye-level and approximately 70 cm from the participant. The spoken syllables (recorded by a male, native Dutch speaker) were presented at about 65 dBA through a speaker positioned directly beneath the monitor. The audiovisual stimuli were all congruent.

Figure 1

Experimental Design



Note: The three conditions of the experiment. In the visual condition, a visual-only stimulus was presented on screen for 800 ms. In the auditory condition, the participant heard a spoken syllable while looking at a fixation point. In the audiovisual condition, both auditory and visual stimuli were presented. These were shown simultaneously in the 0 ms SOA and visual stimulus preceding auditory stimulus with 300 ms in the 300 ms SOA.

The experiment was divided into four separate blocks, containing 126 trials each. Two blocks contained 0 ms SOA AV stimuli (the visual and auditory stimuli occurred at the same time), and the other two blocks contained 300 ms SOA AV stimuli (the visual stimulus preceded the auditory stimulus by 300 ms). The blocks were alternately presented two times per participant. Half of the participants thus received 0-SOA as their first block and the other half received 300-SOA as their first block (after practicing). In total, there were six conditions(0-SOA/AV, 300-SOA/AV, 0-SOA/V, 300-SOA/V, 0-SOA/A and 300-SOA/A) and 504 trials. The experiment was programmed in E-Prime 3.

EEG and ERP were recorded using 32 Ag – AgCI electrodes which were placed according the international 10 – 20 system. Two electrodes served as reference (Common Mode Sense; CMS) and ground electrode (Driven Right Leg; DRL). Additional electrodes were placed on both mastoids, above and below the right eye to measure the vertical electro oculogram (EOG) and on the outer canthi to measure the horizontal electro oculogram.

Preprocessing

The EEG data were preprocessed using Brain Vision Analyzer (BVA) and digitised at a sample rate of 512 Hz. The data were re-referenced off-line to an average of the two mastoid electrodes and were filtered with a high-pass filter of 0,5 Hz, a low-pass filter of 30 Hz and a 50 Hz notch filter to remove the 50 Hz interference. After ocular correction (Gratton and Coles), the data were segmented in 1300 ms epochs with a 500 ms prestimulus baseline. An artifact rate of 150 Hz has been applied. There were two participants with a remarkably small number of remaining segments (less than 50 segments in every condition) after applying the artifact rate, which is why they were excluded from the data. An AV-V condition was created by subtracting the visual-only ERPs from the audiovisual ERPs, so it could be compared to the audio-only modality for assessing the effect of audiovisual integration.

Procedure

The participants were asked in advance to wash their hair and not to wear make-up on the day of the experiment. Due to Covid-19 and the increased safety measures it entails, every participant received a clean face mask on arrival and had their temperatures checked before entering the laboratory. They were asked to leave their phones, smartwatches and other communicative devices in order to make sure there were no distractions for the participant. After placing the face sensors and preparing the cap with the plugged electrodes, the participant took a seat in a dimly-lit, noise-cancelled and electrically shielded cabin. Each experiment took about 80 minutes, EEG preparations included. After clear instructions ('Push a random key when you see or hear Raa') and a short practice block (9 trials) in which the participants could get familiar with their task, the actual experiment started. The participant was allowed to take a short break after 60 trials into the block. After each block, the participant's welfare was confirmed by a short checkup through the microphone.

Statistical analysis

After preprocessing, the N1 and P2 peak amplitude and latencies were exported to SPSS and submitted to a repeated measures analysis of variance (ANOVA), using the latency and amplitude scores of the Cz electrode. Paired T-Test were used to follow-up interaction effects. Besides the two removed outliers, data was missing for three participants. These were also excluded from the statistical analysis.

Results

Repeated measures ANOVA on the N1 latency showed no significant main effect of modality ($F(1, 21) = .005$, $p = .943$) and no significant main effect of SOA ($F(1,21) = 0,73$, $p = .402$). Also no significant interaction effect was observed ($F(1, 21) = 0,16$, $p = .692$). Analysis on the N1 amplitude also showed no significant main effect of modality [$F(1, 21) =$

0,35, p = .561] nor SOA (F(1, 21) = 2,17, p = .155). No significant interaction effect was observed either (F(1, 21) = 3,39, p = .080).

Repeated measures ANOVA on the P2 latency showed no significant effect of modality (F(1, 21) = 4,17, p = .054), nor SOA (F(1, 21) = .559, p = .463). No significant interaction effect was found (F(1, 21) = 2,37, p = .138). Analysis on the P2 amplitude showed no significant effect of modality (F(1, 21) = 2,84, p = .107), but it did show a significant effect of SOA (F(1,21) = 15,45, p = .001). This significant main effect, however, can not be interpreted due to the significant interaction effect (F(1, 21) = 33,24, p < .001). Paired t-tests confirmed a significant simple effect of modality on both SOA's (SOA-0: [t(21) = -5.264, p < .001] SOA-300: [t(21) = 2.626, p = .016]) and a significant effect of SOA on the AV-V modality (t(21) = 6.138, p < .001). Paired t-tests showed no significant effect of SOA on the A-modality (t(21) = .256, p = .801).

In short: The only two significant effects are observed in the P2 and show a larger amplitude in the AV-V modality (M = 12,44, SD = 4,85) than in the A modality (M = 9,28, SD = 3,90) with a SOA of 0 ms and a decrease in amplitude in the AV-V modality (M = 7,44, SD = 2,92) than in the A modality (M = 9,09, SD = 4,41) with the 300 ms SOA. The mean differences are presented in Table 1 and representations of the N1 and P2 peaks in both SOA conditions are presented in Figure 1 and Figure 2.

Table 1

Mean Differences between A and AV-V Modalities (N1 and P2).

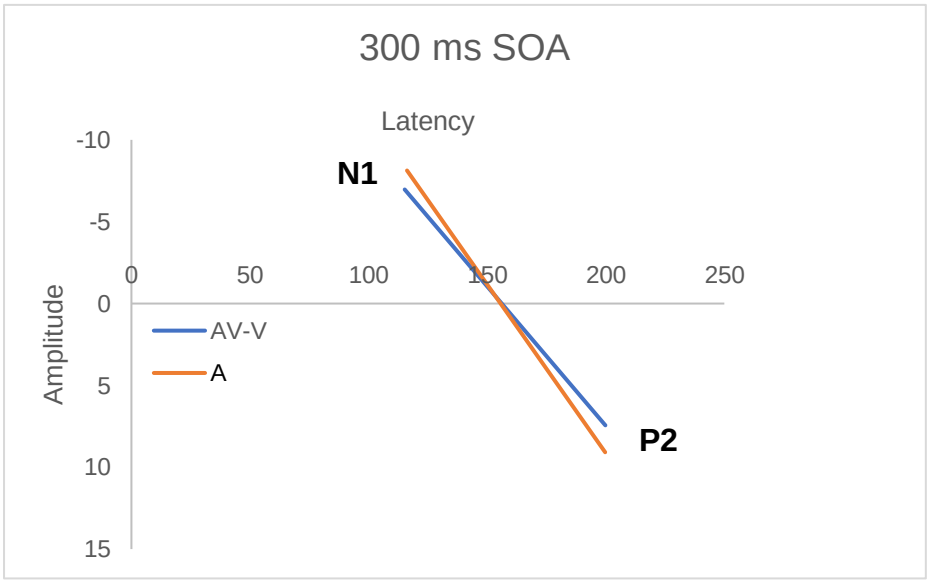
	N1		P2	
SOA	300 ms	0 ms	300 ms	0 ms
Amplitude (µV)	-1,16	0,64	-1,65*	3,16*
Latency (ms)	-0,98	0,80	-0,09	-8,61

*Note: *p < .05*

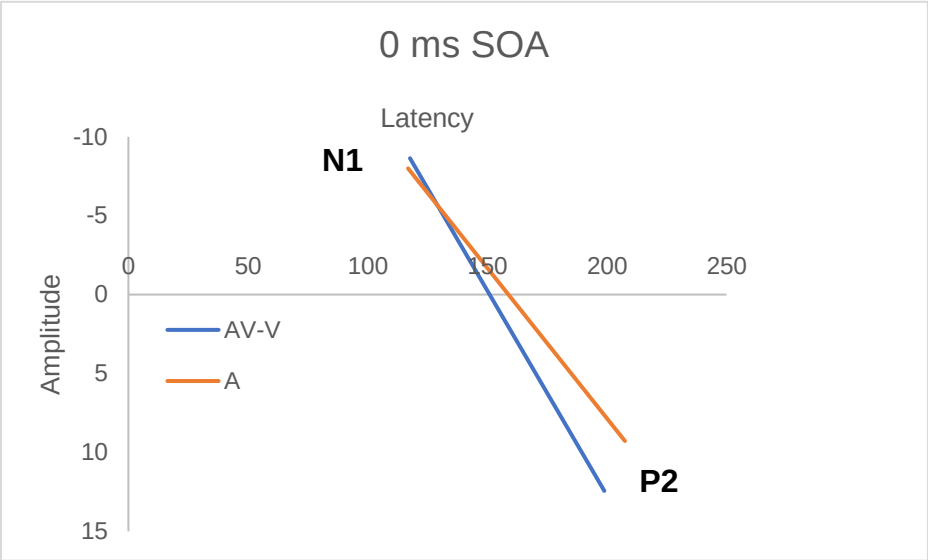
Figure 2

Representations of N1 and P2

A



B



Discussion

The present study aimed to determine whether text as a visual stimulus would cause attenuations in amplitude and an increased latency in the N1 and P2 when accompanied by concurrent auditory stimuli. To do so, text was either presented 300 ms before the onset of the auditory stimulus (aiming to provide the visual signal with predictive value with respect to sound onset), or simultaneously with the auditory stimulus (diminishing the predictive value of the visual signal). Based on research in which the visual signal comprised lip-read information rather than text, , the first main expectation was that the N1 component would show the largest attenuation and speeding effects when the visual text stimulus precedes the auditory stimulus, just as it does with lip reading. The second expectation was that P2 would show less extreme attenuations, since the current study only presents congruent audiovisual stimuli and P2 seems to show stronger attenuations in response to incongruent stimuli according to previous research (Vroomen & Stekelenburg, 2007). However, the results show that text induced an attenuation of the auditory ERP only for the P2 in the 300 ms SOA condition. Meanwhile in the 0 ms condition, there was a significant increase of amplitude

instead of an attenuation. There was no main effect of latency in any of the ERP components and also no significant decrease of amplitude in N1.

While the expectation was to find attenuations in especially N1, it is surprising that the only significant attenuation was located in P2. Another expectation was to find a significant decrease in latency, but those were absent as well. No significant effect was found in the N1. Not only do the results seem contradictory to the hypotheses, but also to multiple scientific theories. According to Stekelenburg & Vroomen (2010) the N1 component seems to be especially dependent on the predictability of the preceding visual stimulus and the P2 component is not. Also would the P2 component be especially sensitive for incongruent stimuli, which causes it to show more extreme attenuations to incongruent stimuli than for congruent ones. No incongruent stimuli were presented during the experiment, but the only significant attenuations are still in P2. The current findings also contradict the letter/speech integration research of Froyen et al. (2008), which found that an increasing SOA also leads to decreasing amplitude with textual visual stimuli. In present study, SOA only significantly affected P2 amplitude.

There are several reasons that may have led to finding results that do not support the hypothesis. As a first, the increased response of the P2 in the 0 ms SOA is not an unusual finding. Wassenhove et al. (2005) already described that enhanced responses occur regularly due to supra-additivity. It could also be explained by the findings of Giard and Peronnet (1999), which suggested that increased responses may be dependent of the subject's dominance for auditory or visual attributes. The difference in dominance for visual or auditory attributes was not taken into account in the present study, which could have led to the current finding and the fact that it contradicts the main expectation. For future research, it would create more insight into the ERP responses to pay attention to the subjects dominance for visual or auditory attributes.

As a second, many participants have described the experiment as “incredibly boring”. The experiment took a long time and required a high level of attention, which made participation mentally intensive. This could have caused decreasing levels of attention within the participant. Although audiovisual integration operates mainly automatically, findings of Alsius et al. (2005) suggest that multimodal binding is subject to attentional demands, which supports that a lack of attention could have affected this.

When looking at earlier mentioned literature, an explanation could be found in the fact that the simulated anticipatory motion in this research is not actually motion. When comparing the involved brain areas that have been distinguished in the single letter research of Froyen et al. (2008) and the audiovisual speech perception study of Calvert et al. (2002), both studies describe the superior temporal sulcus as most important brain area when it comes to audiovisual integration. One important difference between these studies is that Calvert et al. (2002) reports involvement of the premotor cortex and Froyen et al. (2008) does not report involvement of any motor area at all. It could be a possibility that actual physical movement has a different effect on ERP components than stimulus onset asynchrony only. This might be interesting to take into account for future research. It would be clarifying to conduct EEG research with mixed textual- and lipreading visual stimuli, so direct comparison would be possible. Interesting would be to assess the involvement of motor areas with both kinds of stimuli and then compare the ERP components. After specifying the possible explanations for the contradictory results, the findings are not illogical after all.

On the other hand, the present study has a solid theoretical base for well executed research and also offers an extensive amount of possibilities for replication. It took a leap in the unknown by replacing lipreading cues with text as new visual stimuli and would be an excellent stepping stone towards follow-up research. As mentioned earlier, the P2 attenuation in the 300 ms SOA was the only significant increased amplitude that confirmed the main

expectation. Wassenhove et al (2005) already explained that these attenuations are expected because of the temporal predicting value of preceding visual information, which was simulated by the 300 ms SOA condition. However, now that we used text as visual stimulus instead of lipreading, the question could be asked whether the attenuated amplitude is actually caused by temporal predicting value since text could also add phonetical predicting value. Since Froyen et al. (2008) did not report any motor areas to be involved in audiovisual speech integration with textual stimuli as visual cue, because there is no motion involved with text, it might be possible that P2 would actually be responding to phonetic instead of temporal predictability. This is an interesting point to take into account for future research.

To conclude, EEG has been conducted to investigate whether the auditory N1 and P2 components show similar audiovisual integration effects with text as with lip reading, while anticipatory movement was simulated using SOA conditions of 0 and 300 ms. Although strong attenuations and increased latency were expected in the N1 peak, no significant effects were found. Mild, but significant, attenuations were expected in P2 and these were also found in the 300 ms SOA, which could be explained by temporal or phonetic predictability. Enhanced amplitude of the P2, however, was found in the 0 ms SOA. This was not in line with the main expectation, but based on previous literature (Wassenhove et al., 2005) this is not an unusual finding. Significant speeding effects in P2 were also expected, but there were no significant latency effects whatsoever. Although not all findings were in line with the main expectations, variability in the results does not particularly have to be a problem since Baart (2016) also described that significant N1 and P2 effects are, even with lipreading cues, not always observed.

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual Integration of Speech Falters under High Attention Demands. *Current Biology*, *15*, 839–843. https://doi.org/10.1016/j.cub.2005.03.046

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology*, *53*(9), 1295–1306. https://doi.org/https://doi.org/10.1111/psyp.12683

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. https://doi.org/10.1016/j.neuropsychologia.2013.11.011

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters*, *471*(2), 100–103. https://doi.org/10.1016/j.neulet.2010.01.019

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, *29*(6), 578–584. https://doi.org/10.3758/BF03214277

Blackwood, D. H. R., & Muir, W. J. (1990). Cognitive brain potentials and their application. *British Journal of Psychiatry*, *157*(9), 96–101. https://doi.org/10.1192/s0007125000291897

Blustein, J., Weinstein, B. E., & Chodosh, J. (2018). Tackling hearing loss to improve the care of older adults. *BMJ*, *360*, 9–10. https://doi.org/10.1136/bmj.k21

Bonath, B., Noesselt, T., Krauel, K., Tyll, S., Tempelmann, C., & Hillyard, S. A. (2014). Audio-visual synchrony modulates the ventriloquist illusion and its neural/spatial representation in the auditory cortex. *NeuroImage*, *98*, 425–434. https://doi.org/10.1016/j.neuroimage.2014.04.077

Borowsky, R., Cummine, J., Owen, W. J., Friesen, C. K., Shih, F., & Sarty, G. E. (2006). fMRI of ventral and dorsal processing streams in basic reading processes: Insular sensitivity to phonology. *Brain Topography*, *18*(4), 233–239. https://doi.org/10.1007/s10548-006-0001-2

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649–657. doi:10.1016/S0960-9822(00)00513-3

Czap, L. (2011). On the audiovisual asynchrony of speech. *Auditory-Visual Speech Processing*, 137–140. Retrieved from https://www.researchgate.net/publication/288981827_On_the_Audiovisual_Asynchrony_of_Speech

Froyen, D., Atteveldt, N. Van, Bonte, M., & Blomert, L. (2008). *Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds*. *430*, 23–28. https://doi.org/10.1016/j.neulet.2007.10.014

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63*(3), 289–293. https://doi.org/10.1037/h0055703

Jones, J. A., & Callan, D. E. (2003). Brain activity during audiovisual speech perception: An fMRI study of the McGurk effect. *NeuroReport, 14*(8), 1129–1133. https://doi.org/10.1097/00001756-200306110-00006

López Zunini, R. A., Baart, M., Samuel, A. G., & Armstrong, B. C. (2020). Lexical access versus lexical decision processes for auditory, visual, and audiovisual items: Insights from behavioral and neural measures. *Neuropsychologia*, *137*, https://doi:10.1016.org/j.neuropsychologia.2019.107305

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Acoustical Society of America, 100*(3), 1777–1786. https://doi.org/10.1121/1.417342

McGurk, H., & MacDonald, J. (1976). Hearing Lips and Seeing Voices. *Nature, 264*, 746 – 748. https://doi.org/10.1038/264746a0

Mheidly, N., Fares, M. Y., Zalzale, H., & Fares, J. (2020). Effect of Face Masks on Interpersonal Communication During the COVID-19 Pandemic. *Frontiers in Public Health*, *8*, 1–6. https://doi.org/10.3389/fpubh.2020.582191

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron*, *28*(2), 617–625. https://doi.org/10.1016/S0896-6273(00)00138-0

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., & Simola, J. (1991). Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*(1), 141–145. https://doi.org/10.1016/0304-3940(91)90914-F

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*(10), 2387–2399. https://doi.org/10.1093/cercor/bhl147

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964 – 1973. https://doi.org/10.1162/jocn.2007.19.12.1964

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26*(2), 212–215. https://doi:10.1121.org/1.1907309

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*(7), 1583–1596. https://doi.org/10.1162/jocn.2009.21308

Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *PNAS, 102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102