UNIVERSITÄT OSNABRÜCK

# Audio-Visual Speech Processing and effects of multisensory asyncronicity

Bachelor's Thesis

Aron Petau

aron@petau.net

967985

(Universität Osnabrück)

Supervisors:

Juliane Schwab

(Universität Osnabrück)

and

Prof. Dr. Michael Franke

(Universität Osnabrück)

October 2020

**Abstract:** In the present study, I seek to identify possible problems related to learning and speech processing in general when presented with audiovisual delays. I review literature on multimodal integration and present the current scientific status. I also examine application-specific properties such as the Echo Effect in Smart Hearing Protection Devices. I discuss possible use cases with a focus on individuals with Autism Spectrum Disorder that could benefit from increased specificity in filtering noise with a tradeoff for increased audiovisual latency. I aim to establish a relationship between audiovisual delays and speech recognition capability while trying to identify a balanced delay making complex filtering possible from an engineering perspective while ensuring that the additional harm to speech processing is minimal.

**Keywords:** Multi-sensory Integration, Smart Hearing Protection, SHPD, Echo Effect, Sensory Asynchrony, Autism Spectrum Disorder, Multi-modal Re-calibration, Speech processing under temporal lag, Temporal Window of integration, just noticeable difference

# Contents


**1 Introduction** **4**

**2 Literature Review** **5**
2.1 Introduction . . . 5
2.2 Multi-sensory integration . . . 5
2.2.1 Speech and Gestures . . . 7
2.2.2 Speech and Visual Lip Movement . . . 7
2.2.3 Multi-sensory asynchronies and the Temporal Window of integration 8
2.2.4 The McGurk Effect . . . 9
2.2.5 Just Noticeable Difference (JND) . . . 10
2.3 The Echo Effect . . . 10
2.3.1 Delayed Auditory Feedback (DAF) . . . 10
2.3.2 Tolerable Delays . . . 10
2.3.3 Smart Hearing Protection (SHPD) . . . 11
2.4 Age Effects . . . 11
2.5 Autism Spectrum Disorder . . . 12
2.5.1 Possible Differences to neurotypical individuals . . . 12
2.6 Conclusion . . . 12

**3 Timing Comparison** **13**

**4 Methods** **14**
4.1 Motivation . . . 14
4.2 Assumptions . . . 15
4.3 Hypothesis . . . 15
4.4 Procedure . . . 16
4.4.1 Participants . . . 16
4.4.2 Preparation . . . 16
4.5 Materials . . . 17
4.6 On latency effects in online studies . . . 17
4.7 The Hardware and Environment . . . 17
4.8 The Software . . . 17
4.9 Stimuli . . . 17
4.10 Variables . . . 18
4.11 Setup . . . 19

**5 Results** **19**
5.1 Statistical Analysis . . . 19
5.1.1 Spectrogram Analysis . . . 19
5.1.2 Descriptive Statistics . . . 19
5.1.3 Analysis of the Variance . . . 19

**A Appendix** **24**
A.1 Acknowledgements . . . 24
A.2 Declaration of Authorship . . . 25

# 1 Introduction 💬

It is well known that sensory modalities interact during speech comprehension. I will conduct a Literature Review of prior findings in the field, discuss those, and subsequently present my own experiment which aims to observe the effects of delays in auditory speech signals that would occur when utilizing selective digital filtering for background noise or distressing sounds, which could be of great impact, especially for non-neurotypical people diagnosed with Autism Spectrum Disorder. There are structural differences as to how individuals with ASD process stimuli, especially it will be shown that their multimodal integration seems to work differently. Speculatively, this is a major reason For Individuals with ASD to take longer for linguistic skills to develop during childhood, in severe cases even remaining completely non-linguistic. As will be pointed out, we believe that a slower and less attenuating mmltimodal integration is responsible and there is a possible remedy in smart hearing devices. When an individual has trouble overcoming problems in speech processing due to sound defects, environmental noise and other impairing factors, it makes intuitive sense to try and eliminate signal defects to improve speech perception ability.

These Smart hearing protection devices (SHPD) employ complex filtering that goes beyond frequency filtering, which is not differentiating among sounds. Advanced filtering techniques mean live digital processing of the input and it can be generalized that the processing time positively scales with filtering complexity. Thus there open up many interesting research questions regarding how such a digital filtering device could operate and it poses rather unique and new challenges. With the artificially created lag between the incoming sound and the output of the improved signal, depending on the nature of sound isolation, we have to deal with an echo, where the individual will hear the unfiltered sound first and, after some time the improved version, creating a distorted echo effect if the original stimulus was loud enough.

# 2 Literature Review

## ~~2.1 Introduction~~

It has long been known that congruous and synchronized Visual Input greatly aids peoples ability to perceive audio information and to understand natural language. Seeing the speakers lips especially helps in making sense of what is being talked about. However, this leads to a fair amount of interesting scientific questions.[1] I will review these questions and more and discuss why individuals with Autism Spectrum Disorder (ASD) and hearing-impaired individuals can provide special insights into these questions. For that, I will introduce the large research field of Multi-sensory integration, talk about research carried out in different modalities and present the concept of a temporal Window of Integration (TWIN). Then I will continue to deal with questions about the ability to detect asynchronies between modalities and discuss several ideas concerning echos in hearing. Finally, I will have a look at research on individuals with ASD and explain what we know about the differences towards neurotypical individuals,[2] concerning multimodal Integration. The goal is to inform about the current state of research and identify possible open questions worth more research.

## 2.2 Multi-sensory integration

The idea of multi-sensory integration goes back over a century, a notable early example being Stratton (1896), who experimented with vision-distorting glasses. Since then there is a considerable body of research conducted. Why do you have to switch off the radio when you try to park the car? Why is it harder to understand people when you cannot see their face? The most prominent theory to date was put forward in 1986 by Meredith and Stein (1986), where they found, via observing single cell neurons in several animals that some neurons respond differently to specific sensory inputs. Those neurons that react to input

[1] e.g. How can visual information help us hearing? Are these modalities integrated into one stream of information or are they processed separately? What happens if this process is dysfunctional? What happens when the actual sensory information is corrupted? As in, say, a video call with lagging audio?

[2] in studies often called TD - typically developed, for us, development is only a secondary concern, we will take neurotypical individuals to extend only to the weaker notion of the current absence of neurological abnormalities

in multiple modalities they called "multisensory", proving that multisensory convergence is a common and essential concept in sensoric processing. Later, in Stein and Meredith (1993) they build on that, putting forward the idea that this convergence is not restricted to a neuronal level, but ~~it~~ is a global concept governing sensory processing in the entire brain. This was called Multi-sensory integration ~~and is still discussed in recent articles.~~ Integration was since often modeled via statistical neuronal networks. (Nakamura, 2002)

Seeing that integration seems to be a common phenomenon, we might ask which purpose it fulfills, whether and how powerfully it can enhance our processing capacity. One paradigmatic study was conducted in 2006 by Ross et al., where speech processing was observed when participants were presented with auditory input alone and contrasted with additional visual information of articulatory movements. They also manipulated the signal-to-Noise Ratio (SNR) in the auditory signals to see whether the quality of the single inputs has any effect. They found a performance increase in understanding up to three times when compared to the uni-modal condition, and observed that the integration seems to work best with medium SNRs, meaning that our system might be best attuned to only partly corrupted inputs, corresponding best with a real world scenario, with all kinds of interference noises occurring at almost all times. (Ross et al., 2007)

A more recent study observed the same phenomenon, ~~–~~while having a closer look at the neuronal behaviour in the actual human brain through scalp recordings via EEG. They extend on the findings by Ross et al. by examining continuous speech versus single syllables, providing a more naturalistic framework. ~~Furthermore, –~~they show an increase in performance even for noise-free congruent situations, once more demonstrating that temporally congruent audiovisual (AV) stimuli (as occurring in natural face-to-face conversation) greatly aid in processing and understanding speech.(Crosse et al., 2015)

Related, and similarly striking is our exceptional ability to synchronize to rhythmical stimuli. A 2015 study by Iversen et al. challenged the idea that our timing and synchronization abilities are bound to a specific modality by comparing hearing and deaf individuals. Finding no impairment in rhythmic synchronization in the deaf group when presented with rhythmic visual stimuli, when compared to the hearing group with auditory

stimuli, they proposed the existence of an amodal timing system responsible for integration. In support, there was no accuracy difference for the hearing and deaf groups for visual synchronization tasks, hinting towards this timing system not being predetermined and adaptive in nature. (Iversen et al., 2015)

### 2.2.1 Speech and Gestures

Another well-established field of research is Audio-gestural integration. The idea that we constantly incorporate information about facial expressions, body language, and hand gestures into our processing of speech fits within the framework of Multi-modal integration. Importantly, secondary input seems not simply to aid a specific uni-modal processing, but the whole processing pipeline seems to be amodal in nature, or agnostic to the modality. Specifically for speech and gestures, synchronizing effects have again recently been demonstrated by Pouw and Dixon (2019), where again the benefits of integration were the biggest under suboptimal conditions where subjects heard a slight echo of 150ms as a distraction. In another EEG study by Biau et al. (2015) it has been put forward that rhythmically congruent hand gestures, so-called "beat gestures" have a significant "tuning" effect on the low frequency oscillatory bands in the brain, which would be a good explanation as to how the integration is realized.

### 2.2.2 Speech and Visual Lip Movement

Another powerful demonstration of Multi-modal integration comes from an oft-cited paper by Calvert et al. (1997), Where they specifically looked at the phenomenon of lip-reading which amounts to trying to assess auditory information visually. In normally hearing participants, lip information being available will lead to a major speech perception increase. The study being conducted with fMRI clearly showed that the Visual Lip-Reading Information only was enough to activate areas in the auditory cortex, suggesting that these stimuli were processed as if they were of auditory nature. Additionally, a counter-check for pseudospeech showed that the activation patterns in the auditory cortex are more than random excitement reactions to face movement, as the activation specifically only occurred

when faces actually mouthing real words or language-like pseudowords were presented. For non-linguistic stimuli, no activation was present. For an excellent in-depth review, see Stilp (2020), where several speech configurations are opened up and cases are neatly separated between forward effects, where the context precedes the target, and backwards effects, with the opposite occurring. Especially interesting for us are the backward proximal effects, which would include echo and other typical speech effects.

### 2.2.3 Multi-sensory asynchronies and the Temporal Window of integration

Based on the framework of Multi-sensory integration that was already introduced, a very sensible question might be what the limits of integration are. Some research about properly functioning integration was already presented, but what about situations where it does not? In a naturally occurring dialogue that may not be the first thing that comes to mind, but in an ever-increasing digital world of indirectly transmitted speech, we come to note that the temporal alignment of visual information and auditory input is of essence here. Think of the mild annoyance when the subtitles are slightly off, or even gross misunderstandings during an online Video Conference caused by temporal misalignment. A popular term here is the Temporal Window of Integration (TWIN) and it tries to capture some amount of minimal synchronicity that needs to be present to ensure speech comprehension. A famous first try at identifying the temporal breaking point of integration comes from Sumby and Pollack (1954). van Wassenhove et al. (2007) ~~and Team~~ performed a classic simultaneity judgment (SJ) and an Identification Task in a separate experiment, to replicate these rather ancient results with more accurate measurements. In a SJ Task, the participant is presented with two stimuli temporally close together and has to decide whether those stimuli occurred simultaneously or not. Their findings are surprisingly close to the original ones made by Pollack and conclude that specifically audiovisual (AV) integration successfully occurs within a frame of about 200ms, making AV bi-modal integration relatively resilient against temporal asynchronies. Another important finding for us is that the Modal Order seems to matter. Integration was overall better when auditory stimuli were training the visual stimuli, making sense in so far that hearing the

sound before seeing the source is a quite an unnatural situation and light can travel quite a bit faster than sound, usually arriving earlier at the individual. [3] Hay-McCutcheon et al. (2009) Further research suggesting that tolerance for visual-leading asynchronies is bigger can be found in Maier et al. (2011). Humans seem to be much more sensitive overall towards auditory-leading stimuli, which is likely explained by the relative minor statistical occurrence in nature.

### 2.2.4 The McGurk Effect

Also essential in the context of Multi-modal integration is a classical illusion dubbed the McGurk effect after the first team to note its existence. In order to produce the effect, they took a video of a speaker and replaced the phoneme in the auditory canal of the video with a different one. If done correctly, an incredibly robust "Fusion" occurs, where the visual information of the speakers lips together with the auditory information of a conflicting phoneme get "merged" and form a third phoneme that can be distinctly heard, without being present in any of the stimuli. The effect persists even when the subject is presented with the uni-modal presentations of the phonemes separately and therefore knows the third phoneme cannot be real. Macdonald and McGurk (1978) This rather astonishing effect has been serving as a paradigmatic test for audiovisual integration, for a compelling analysis of why it has to be considered outdated, see Rosenblum (2019). The case is being made, that the McGurk Effect is not fine-grained enough to properly assess multimodal integration in general and may hinder research regarding automaticity of integration.

Soto-Faraco et al. (2004) used the McGurk effect in an interesting manner, where they produced the effect in the Independent Dimension in a speeded classification [4] task, effectively showing that Multi-sensory Integration happens automatically and we cannot just disregard one modality stream of information in processing.

[3] just think of how in an approaching storm the lightning occurs before the thunder.
[4] explain

### 2.2.5 Just Noticeable Difference (JND)

Attunement? Closely related to the question of how large TWIN is, is the concept of the Just Noticeable Difference (JND). While TWIN looks at the breaking point of successful integration, we now talk about a presumed point where integration is still possible, but we already notice the temporal misalignment. Think again Movie subtitles. how many ms do they have to be off-sync in order for us to realize there might be a problem? This is an interesting topic of research, because this point does not seem to be fix, it can vary, depending on the needs of the situation. This amazing ability is called Attunement. Whats the minimum delay people can notice? Quené (2007) Do not use! its for tempo in speech, not for asynchrony. Find other sources

What about Sub-Noticeable Delays? Any Studies?
Klockgether and van de Par (2016)

## 2.3 The Echo Effect

### 2.3.1 Delayed Auditory Feedback (DAF)

Delayed auditory feedback classically occurs when an a speaker hears her own voice in a slightly delayed manner, which has been shown to induce stress. Badian et al. (1979) Usually this occurs when the speaker is wearing hearing aids, but a microphone connected to a speaker with some latency for karaoke is another easy example where DAF could occur.

In a rather recent replication of a classic study McNeill (1992) on Gestural Synchronicity, Pouw and Dixon (2019) found a reliable entrainment effect by Introducing a 150ms DAF and analyzing subsequent performance.

### 2.3.2 Tolerable Delays

Also utilizing DAF, Stone and Moore (2002) looked at the permissible delays in hearing aids and identified that for regular speech, no disturbance is noticed under 30 ms. This means that any hearing aid processor, to be helpful and not actually detrimental, should

ideally relay auditory information faster than this threshold.

### 2.3.3 Smart Hearing Protection (SHPD)

This becomes especially interesting when confronted with the emerging option of smart hearing devices. Whereas it is nowadays efficiently and fast possible to filter out auditory frequencies, [5] this does have annoying side effects as filtering by frequency completely disregards the nature of the auditory input. With the use of modern digital microcontrollers, it becomes possible to preprocess the audio signal to decide before relaying on to the integrated speakers, what type of audio is presented. Based on the result, it would become possible to apply a different set of filters, specifically tailored for the incoming signal. This type of advanced filtering comes with a substantial trade-off. Generally, the more complex and advanced a filter becomes, the more processing time is added, introducing more delay for the hearing individual. For an interesting in-depth discussion of this trade-off see Lezzoum et al. (2016)

## 2.4 Age Effects

How much does development and age influence this integration capability?

Going in the opposite direction, looking at age-related hearing loss, Rosemann and Thiel (2018) brought forward strong fMRI data to suggest that with increased hearing loss, the AV integration gets stronger. This would suggest that there likely is no linear relationship between hearing capacity and integration and it supports other claims discussed earlier that integration works best under moderately adverse conditions (such as mild hearing loss). Du et al. (2016) suggest that increased multimodal integration seems to be a common and effective way to compensate impaired speech perception.

[5] for example, by applying a high- or low-pass filter to make the mid-range frequencies, which contain speech more present

## 2.5 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) often presents itself in social interaction and communication deficits and often goes along with atypical processing of sensory information. (APA, 2013). There have been established consistent findings from a multitude of studies regarding regularities in the atypical sensory processing across individuals with ASD. In Brandwein et al. (2013) this is discussed and extended to more general, basic nonspeech stimuli, suggesting this to be a rather consistent effect.

### 2.5.1 Possible Differences to neurotypical individuals

One rather well-established processing difference lies in re-calibration speed, or maybe even the overall capacity for re-calibration. As very well explained in Turi et al. (2016), TD individuals exhibit rapid re-calibration, often shown via SJ Tasks, where the skew of the preceding runs partially determines the judgement in the current run, the individual gets "attuned" to temporal discrepancies. This finding is particularly well demonstrated in Bertelson et al. (2003), using hearing individuals. This rapid re-calibration is very diminished in ASD individuals, one consequence being a lower susceptibility to the McGurk Effect. Another, probably more important one is the the reduced ability to optimise sub-optimal speech perception situations. This would also explain very well, why ASD typically start to speak faster and under-perform in language reproduction. More on a comparison with still developing Children can be found in Noel et al. (2017).

For a concise overview see Stevenson et al. (2014)

## 2.6 Conclusion

As could be seen earlier, some of these phenomena are overwhelmingly well researched, while others are still largely open. Even though we know the noticeable latency boundary for a smart hearing protection device is somewhere around 30ms, this refers to self-reported variables, it does not strictly have to coincide with a latency boundary for good performance. It is also an open question whether these boundaries are generally similar for TD and ASD populations. Furthermore, although the DAF is well represented in the research, other

Echo-configurations that are imaginable with a SHPD are critically missing.

TODO: Write better (longer) Conclusions, make it a mini-discussion

# 3 Timing Comparison

The objective of the experiment at hand is clear, but what variable size to use? For better comparability, the auditory lags in the delay and the echo condition should be of the same size. In our simple setup we settle on 3 conditions: a 0-condition, to get a benchmark result, a condition with small difference (echo or delay respectively), and a condition with a large, obvious difference, where we expect to obtain clear results and which will enable us to verify our general hypothesis, that speech processing ability is indeed positively dependent on synchronicity within our specific setup. Analogous, we expect performance to suffer more when the simulated echo is present compared to conditions without echo. Following that, the large value should be chosen in a range where literature suggests that we can expect a clear performance impact. Slightly more complicated is the choice of the smaller value, since ideally we want this condition to impact the performance slightly without necessarily being noticeable to the participant. Upon reviewing the literature with this specific question in mind for the larger value we settled on 400ms. This is estimated to be distinctly noticeable, with unambiguous impact on speech reception performance.

Several TWIN studies suggest that the speech-specific audiovisual TWIN is asymmetric with preference to visual-leading stimuli. (van Wassenhove et al., 2007; Petrini et al., 2009; Maier et al., 2011) Here, the largest Temporal window is estimated to be around 200ms, from -30ms to 170ms. Our value should have clear effects, so we want to remove compensation effects from TWIN, resulting in our value having to be larger. In a more recent study, Li et al. (2021) noted that in a standard Simultaneity Judgement (SJ) Task with stepped delays from -400 to 400ms delay, 50 percent of the participants judged the 200ms delayed stimulus to be synchronous. Similarly, Maier et al. (2011) provided evidence that stimuli with an auditory lag in the range of 67ms up to 267ms are more likely to be judged synchronous than even actual synchronous stimuli. This provides further evidence that, in order to create a condition in which the majority clearly is able to identify a

temporal lag, the lag itself would have to be at least 300ms large.

For the smaller value, it has to be acknowledged that our study being a browser-based study has technical limitations being discussed in Bridges et al. (2020). The authors, which are the same team developing PsychoPy (Peirce et al., 2019) Due to a host of reasons ranging from variance in internet speed over browser compatibility and hardware limitations, we need the value to be at least 10ms in order to reliably separate any detected effect from these interfering effects. The literature seems quite divided on the question at what temporal difference subjects can reliably detect change. What seems clear is that this ability is highly dependent on the type of auditory signal used. People are generally very capable of detecting temporal delay in their own voice. Some studies using the delayed auditory feedback (DAF) report people noticing a delay as small as 3-5ms (Agnew and Thornton, 2000), others report the smallest noticeable DAF rather to be around 15ms under optimal conditions (Stone and Moore, 2002) Studies looking at DAF cannot be applied at face value here, since the detection threshold for own voice recordings consistently seems a lot lower than for external voices, most studies demonstrate consistent findings that auditory lag in DAF is already clearly annoying and performance decreasing to the speaker at 20-30ms. (Stone and Moore, 2002; Agnew and Thornton, 2000), with Goehring et al. (2018) even claiming elevated annoyance at around 10ms for normal hearing (NH) participants. The tolerance for external voices is much more interesting, due to that being more relevant to the question of general speech perception. The team of Lezzoum et al. (2016) looked at simulated echoes, found that the smallest speech-related echo was detected by at least 20 percent of the participants at 16ms delay. In a review of intersensory synchrony, Vroomen and Keetels (2010) concluded that

> Temporal lags below 20 msec are usually unnoticed, probably because of hard-wired limitations on the resolution power of the individual senses.

For non-speech stimuli, generally the asynchronity detection threshold is smaller, Lezzoum et al. (2016) measuring a bell signal with delayed echo to be detectable at 8ms, Zakis et al. (2012) estimating experts to be able to detect delay in music already down at 3-5m. Analogous, the TWIN for non-speech stimuli is smaller, Petrini et al. (2009)

measuring a 112ms Window in an audiovisual SJ task with drumming sounds.

Furthermore, there is a clear tendency for NH-Participants to be less tolerant towards temporal delay than HI-Participants. Also, the tolerance seems to scale linearly with hearing impairments, suggesting that HI-people have one or several compensating mechanisms in place that are resilient against temporal delay. For us this means that designing the experiment with NH people in mind, it will later be applicable to HI subjects too.

To comply both with the technical limitations of a browser-based online study and the need to make the AV-lag small enough to be unnoticed by most of our participants, we chose 10ms for both the delay and the echo condition. We argue that specifically for external speech stimuli this threshold should be well below the participants capacity to detect neither a pure auditory lag nor our simulated echo. should we still find any speech performance impact in these conditions, it should be an indication for strong multimodal subconscious mechanisms involved in speech perception, ultimately preventing the use of any higher order filter in SHPD.

# 4 Methods

## ~~4.1 Motivation~~

My own experiment aims at extending the field of research such that some concrete recommendations for the latency of a smart hearing protection device can be made, where we can be reasonably sure that the additionally introduced latency will not carry negative consequences for speech recognition. Many interesting questions will remain unaddressed, for example, whether a very small latency affects children differently, especially while learning language capacities in school. The debate of whether this latency affects individuals with ASD in a different fashion is also not the focus here. The intention is to set a baseline with adult TD participants and establish a protocol that is repeatable and comparable, easily extensible to different demographics. In order to not be restricted to fully linguistic subjects, we chose a picture-dependent task, where no reading skill is essential. Concretely, we are interested in the possible consequences of an additionally introduced audiovisual

indirectly. In this experimental setup, we want to establish a valid indirect measure of speech comprehension performance when presented with audiovisual delay. We choose reaction time as the operating variable, with the assumption that a higher cognitive load, as present in adverse conditions will represent the language comprehension performance. We also record the accuracy to be able to detect any secondary effects. We intend to keep the experiment as simple as possible to minimize interaction effects. We are further interested in the echo effect unique to the situation of the subject wearing SHPDs and the noise isolation not being sufficient, where the filtered audio conflicts with the earlier original, unfiltered, but dimmed noise. To simulate, we will introduce additional conditions where the original audio is not completely removed, conflicting with the delayed stimulus, creating an echo.

## 4.2 Assumptions

There is a universal underlying mechanism of multi-modal integration for speech perception. Reaction time is indicative of cognitive effort spent on speech perception. The time difference between subjects recognizing the images is negligible. All Stimuli are free from ambiguities, it is always clear what the proper name for the image is. TODO: not sure how much sense a section like this makes, got to rethink it, but would love to have it for transparency

## 4.3 Hypothesis

When presented with greater temporal dis-alignment, speech perception ability decreases. Speech perception ability is operationalized through reaction time (RT), meaning that we expect a bigger reaction time with increasing adversity of the speech stimulus. Concretely, RTs will be shortest in the base condition, 0ms latency, 0ms echo, and the RTs will positively correlate with both latency and echo conditions. We expect some inverse correlation with accuracy, and accuracy will be inversely correlated with recognizability of the image stimulus. Generally, we expect the effect to be stronger in all conditions with echo, because we think this is the harder task overall.

## ~~4.4 Procedure~~

### 4.4.1 Participants

The experiment recruited 50 participants via the university internal mailing list targeting cognitive science students. All participants were native German-speaking adults with normal or corrected to normal vision and had normal hearing. The participants had a male to female ratio of 1:1, with a mean age of 25, in a range of 20 to 34. TODO: real numbers Participation was completely voluntary and written consent was obtained from all participants. They could leave at any time without penalty leading to the destruction of the collected data. Participants received Experiment Hours (VP-Stunden) as compensation, a mandatory part to finish the Cognitive Science program. No other compensation was granted. We require our participants to wear headphones in an attempt to minimize distracting environmental noises beyond our control. The participants are asked to perform the experiment exclusively with 2 fingers of their dominant hand to reduce possible differences between the dominant and non-dominant hand reflecting in the RTs. We exclude any participants reporting a prior psychological diagnosis or reporting diagnoses with ASD.

### 4.4.2 Preparation

We ask participants to complete the experiment on a laptop or computer, ideally sitting on a desk in a fixed position. The participants are instructed via audio instructions while simultaneously seeing written instructions to ensure a rigid comparability to all participating demographics in the future. After the instructions, they receive a few trial runs to get familiar with the nature of the task. ~~This will help us ensure more consistent RTs and fewer outliers, so we hope~~. Participants are then reminded to answer as fast as they possibly can, using only the middle and the index finger of their dominant hand. We ask them to keep a stable posture to lessen the variance in the distance to the screen. They are also instructed to ensure viable viewing conditions: no direct sunlight, subjectively adequate brightness of the screen.

## 4.5 Materials

## 4.6 On latency effects in online studies

Overall, the variance of RTs measured in online experiments seems to be comparable to that of lab-based experiments. (Bridges et al., 2020)

## 4.7 The Hardware and Environment

The entire experiment is conducted in one single browser session requiring only a keyboard, headphones and a display. Breaks in between blocks are possible and not controlled.

~~Due to the experiment having to be executed as an online experiment, naturally, these variables are hard to control.~~ We ensured that it is impossible to participate in the experiment from a mobile device and requested each participant prior to the experiment to ensure an adequate environment: To sit down comfortably in front of a stationary monitor with at least full hd resolution and connect stereo headphones. We establish a procedure to ensure beforehand that the participants computer audio is in a comfortable setting, where both the sentence and the echo can be heard. All resources of the browser-based experiment are downloaded prior to the experiment to prevent download speed and performance issues. We request the participant to be alone in the room and eliminate possible noise interference insofar possible.

## ~~4.8 The Software~~

~~RStudio Team (2020)~~

~~Audacity (2020)~~

~~puredata (2020)~~

~~Tomar (2006)~~

~~MATLAB (2020)~~

## 4.9 Stimuli

TODO:Figure sources

Participants are shown short video clips from the OLACS Corpus (Uslar et al., 2013; Rosemann and Thiel, 2018), which are full HD Recordings of a speaker reading German sentences centered onto his mouth. They are extensively controlled for speech reception thresholds (SRTs) as well as response latencies within adult native german speakers with full hearing capacity. Of the full 160 sentences we settled on using only the first 80, as all of them are grammatically structured comparably One example sentence taken from the corpus is "Der schlaue Kaspar beschattet den faulen Vater." Although the syntactic composition varies, each sentence contains two entities, each attributed with an adjective and either an active or a passive verb connecting the two entities. The corpus contains 160 sentences, of which we selected 80 with best fitting and readily recognizable images. The corresponding images are taken from the internationally tested MultiPic Corpus (Duñabeitia et al., 2018), a set of handdrawn colored files in .png format with available data for measured complexity and percentage of correct recognition in a german speaking population. These are supplemented with a range of copyrightfree images from various online sites to cover each subject appearing in the Text corpus.

All of these are then manipulated using GIMP 2.10.22, centered on a quadratic canvas with transparent background, all resolutions ranging from 500x500 to 1200x1200 pixels.

One example:

## ~~4.10 Variables~~

~~Dependent Variable (DV) : Reaction Time (in ms) and accuracy (correct or incorrect), measured from the onset of the video. Independent Variables (IV) : The Latency of the audio; 0ms latency, 50ms latency, 200ms latency, The echo: 0ms echo (no echo), 30ms echo, 100ms echo Since both base conditions should behave identical, this setup means we have a slightly simpler 2 by 3 setup with only 5 conditions in total. TODO: Justify Values, bigger big latency~~

## 4.11 Setup

Participants are presented with a series of sentence videos from the OLAK corpus, organized in blocks with 8 examples each, after which follows a ISI period of 2000ms. The subject completes a total of 8 blocks.

The task, performed after seeing each individual video clip is constant: the participants have to answer whether the image they are flashed afterwards refers to an entity that was present in the sentence.

The video is always shown on the left half of the screen, the corresponding image stimulus is always visible on the right. The image is visible from the onset of the video, no instruction on where to look is given. The 2000ms ISI indirectly focuses the view via presenting a cross-hair at the center of the screen. The background is kept at the same neutral grey at all times.

While seeing and hearing the videoclip on a pseudo-randomly selected sentence, the participant is presented with an image representing an entity and has to respond via a button press to the stable question of whether this entity was contained in the last sentence she heard. Because it is an image, this cannot be solved simply with auditory memory, as the subject has to form the basic semantic information of the sentence to connect the image with the prior auditory information. The answer is recorded via a keyboard press with separate buttons for yes and no.

TODO: Add images of setup, task etc.

# 5 Results

## 5.1 Statistical Analysis

### 5.1.1 Spectrogram Analysis

### 5.1.2 Descriptive Statistics

### 5.1.3 Analysis of the Variance

# References


Agnew, J. and Thornton, J. (2000). Just noticeable and objectionable group delays in digital hearing aids. *Journal of the American Academy of Audiology*, 11 6:330–6.

APA, A. P. A. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

Audacity (2020). Audacity® software is copyright © 1999-2020 audacity team.web site: https://audacityteam.org/. it is free softwaredistributed under the terms of the gnu general public license.the name audacity® is a registered trademark of dominic mazzoni.

Badian, M., Appel, E., Palm, D., Rupp, W., Sittig, W., and Taeuber, K. (1979). Standardized mental stress in healthy volunteers induced by delayed auditory feedback (daf). *European journal of clinical pharmacology*, 16(3):171–176.

Bertelson, P., Vroomen, J., and De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, 14(6):592–597.

Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., and Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, 68:76–85.

Brandwein, A. B., Foxe, J. J., Butler, J. S., Russo, N. N., Altschuler, T. S., Gomes, H., and Molholm, S. (2013). The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cerebral Cortex*, 23(6):1329–1341.

Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8:e9414.

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

Crosse, M. J., Butler, J. S., and Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, 35(42):14195–14204.

Du, Y., Buchsbaum, B. R., Grady, C. L., and Alain, C. (2016). Increased activity in frontal motor cortex compensates impaired speech perception in older adults. *Nature communications*, 7(1):1–12.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., and Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, 71(4):808–816.

Hay-McCutcheon, M. J., Pisoni, D. B., and Hunt, K. K. (2009). Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis. *International Journal of Audiology*, 48(6):321–333.

Iversen, J. R., Patel, A. D., Nicodemus, B., and Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals. *Cognition*, 134:232–244.

Klockgether, S. and van de Par, S. (2016). Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *The Journal of the Acoustical Society of America*, 140(4):EL352–EL357.

Lezzoum, N., Gagnon, G., and Voix, J. (2016). Echo threshold between passive and electro-acoustic transmission paths in digital hearing protection devices. *International Journal of Industrial Ergonomics*, 53:372–379.

Macdonald, J. and McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3):253–257. cited By 284.

Maier, J., Di Luca, M., and Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of experimental psychology. Human perception and performance*, 37:245–56.

MATLAB (2020). *9.9.0.1592791 (R2020b) Update 5)*. The MathWorks Inc., Natick, Massachusetts.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Meredith, M. A. and Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, 56(3):640–662.

Nakamura, S. (2002). Statistical multimodal integration for audio-visual speech processing. *IEEE Transactions on Neural Networks*, 13(4):854–866.

Noel, J.-P., De Niear, M. A., Stevenson, R., Alais, D., and Wallace, M. T. (2017). Atypical rapid audio-visual temporal recalibration in autism spectrum disorders. *Autism Research*, 10(1):121–129.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1):195–203.

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C., Avanzini, F., Puce, A., and Pollick, F. (2009). Multisensory integration of drumming actions: Musical expertise affects perceived audiovisual asynchrony. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, 198:339–52.

Pouw, W. and Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science*, 43(3):e12721.

puredata (2020). puredata.info.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35(3):353–362.

Rosemann, S. and Thiel, C. M. (2018). Audio-visual speech processing in age-related hearing loss: Stronger integration and increased frontal lobe recruitment. *NeuroImage*, 175:425–437.

Rosenblum, L. D. (2019). Audiovisual speech perception and the mcgurk effect. *Oxford Research Encyclopedia of Linguistics*.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what i am saying? exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex*, 17(5):1147–1153.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition*, 92(3):B13–B23.

Stein, B. E. and Meredith, M. A. (1993). *The merging of the senses.* The MIT Press.

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., and Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Frontiers in Psychology*, 5:379.

Stilp, C. (2020). Acoustic context effects in speech perception. *WIREs Cognitive Science*, 11(1):e1517.

Stone, M. A. and Moore, B. C. (2002). Tolerable hearing aid delays. ii. estimation of limits imposed during speech production. *Ear and Hearing*, 23(4):325–338.

Stratton, G. M. (1896). Some preliminary experiments on vision without inversion of the retinal image. *Psychological review*, 3(6):611–617.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2):212–215.

Tomar, S. (2006). Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Turi, M., Karaminis, T., Pellicano, E., and Burr, D. (2016). No rapid audiovisual recalibration in adults on the autism spectrum. *Scientific reports*, 6:21756.

Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., and Kollmeier, B. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America*, 134(4):3039–3056.

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607. Advances in Multisensory Processes.

# A Appendix

## A.1 Acknowledgements

I would like to thank

Danielle Benesch

(NSERC-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS),

Université du Québec (ÉTS)) for guiding me through the entire process and always providing quick helpful tips and feedback. and the whole research team at the NSERC-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS) for providing useful code for simulating the echo effect.

## A.2 Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.

Aron Petau

Osnabrück, March 30, 2021