



Universität Osnabrück

Fachbereich Humanwissenschaften

Institute of Cognitive Science

# Audio-Visual Speech Processing and effects of multisensory asynchronicity

aron@petau.net

967985

Bachelor's Program Cognitive Science

April 2021 - July 2021

First supervisor:     Juliane Schwab, M.Sc.  
                              Institute of Cognitive Science  
                              Universität Osnabrück

Second supervisor:   Prof. Dr. Michael Franke  
                              Institute of Cognitive Science  
                              Osnabrück

**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 usecases 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

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Literature Review</b>	<b>5</b>
2.1	Multi-sensory integration . . . . .	5
2.1.1	Speech and Gestures . . . . .	7
2.1.2	Speech and Visual Lip Movement . . . . .	8
2.1.3	The McGurk Effect . . . . .	8
2.2	Multi-sensory signal delays, asynchronies and the Temporal Window of integration . . . . .	9
2.2.1	Just Noticeable Difference (JND) . . . . .	10
2.3	The Echo Effect . . . . .	11
2.3.1	Delayed Auditory Feedback (DAF) . . . . .	11
2.3.2	Tolerable Delays . . . . .	11
2.3.3	Smart Hearing Protection (SHPD) . . . . .	11
2.3.4	On the temporal scale of audiovisual asynchrony . . . . .	12
2.4	Age Effects . . . . .	16
2.5	Autism Spectrum Disorder and possible differences to neurotypical . . . . .	17
2.6	Conclusion . . . . .	17
2.6.1	Different Values in the Literature . . . . .	18
<b>3</b>	<b>Experiment</b>	<b>20</b>
3.1	Method . . . . .	22
3.1.1	Participants . . . . .	22
3.1.2	Materials . . . . .	22
3.1.3	Procedure . . . . .	25
3.1.4	Hypothesis . . . . .	30
3.2	Results . . . . .	31
3.2.1	Statistical Analysis . . . . .	31
<b>4</b>	<b>Discussion</b>	<b>32</b>
4.1	My Results in other current research . . . . .	32
4.2	Later studies with ASD . . . . .	32
4.3	Conclusion . . . . .	32
4.4	Suggestions for further research . . . . .	32
<b>A</b>	<b>Appendix</b>	<b>37</b>
A.1	Stimuli . . . . .	37
A.1.1	Images . . . . .	37
A.1.2	Sentences . . . . .	37
A.2	Acknowledgements . . . . .	39
A.3	Declaration of Authorship . . . . .	

## List of Figures

3.1	Visible effects of modification on an example file . . . . .	24
3.2	Example of image files: Pfarrer, Pilot . . . . .	25
3.3	Temporal order of entire experiment . . . . .	26

3.4	Temporal order of presentation in main task . . . . .	27
3.5	Temporal order of presentation in adapted SJ task . . . . .	28
3.6	The stimulus presentation in the main task . . . . .	29
3.7	Presentation of questions with response indicators . . . . .	29
3.8	Example result table . . . . .	31
A.1	C. . . . .	40

# 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 multimodal 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.

Filtering a acoustic signal introduces additional latency to the auditory pathway. In a real world setting, this latency is also present and scales with the distance of the sound source. Nevertheless, the speed difference between the visual sensory information of an event and the auditory sensory information of that event is so tiny it is negligible for most intents and purposes regarding speech processing. Normal earplugs present a physical barrier for sound, not introducing a relevant additional delay. The situation is different when an auditory filtering device is used. Regardless of how the filter works, (analog or digital), any filtering requires processing and produces some auditory delay before it can transmit the filtered auditory information. Modern smart hearing protection devices (SHPD) employ complex filtering that goes beyond frequency filtering, which is not differentiating between types of sound. 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 open up many interesting research questions regarding how such a digital filtering device could operate and it poses rather unique and new challenges.

TODO: describe research focus

## 2 Literature Review

It has long been known that congruous and synchronized visual input greatly aids peoples ability to perceive audio information and to understand natural language.(Sumby and Pollack, 1954) Seeing the speakers lips especially helps in making sense of what is being talked about (Calvert et al., 1997). However, this leads to a fair amount of interesting scientific questions. I will review these questions and discuss why individuals with autism spectrum disorder (ASD) and hearing-impaired individuals can provide special insights into these topics. For that, I will introduce the large research field of Multi-sensory integration, investigate research carried out in different sensory 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 temporal 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 when compared to neurotypical individuals, <sup>1</sup>concerning multimodal Integration.

The goal is to take a look at the current state of research and investigate the basis of our own experiment, such that after the review we can settle on an experiment design that is in line with recent literature and is capable of answering some of the open questions that are not already investigated throughout this section.

### 2.1 Multi-sensory integration

Why do you have to switch off the radio when you try to park the car? So you can see better and concentrate on relevant sensory input. From our everyday experience it is already evident that different sensory pathways (modalities) are somehow linked and can at the very least influence each other and our experienced perceptual performance. This

---

<sup>1</sup>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

range of phenomena is researched since over a century ago, a notable early example being Stratton (1896), who experimented with vision-distorting glasses. Since then there was a considerable body of research conducted for a host of different modal combinations. The most prominent theory to date was put forward in 1986 Meredith and Stein (1986); Stein and Meredith (1993), who recorded single-cell neurons in several animals, finding that some neurons respond differently to specific sensory inputs. Those neurons that react to input in multiple modalities they called “multisensory”, proving that multisensory convergence is a common and essential concept in sensoric processing. In their later book, they build on that, putting forward the idea that this convergence is not restricted to a neuronal level, but instead is a global concept governing sensory processing in the entire brain. This was called Multi-sensory integration. The Idea is that redundant, overlapping and sometimes mutually exclusive sensoric information from all modalities has to be integrated by the nervous system to form the coherent picture we are used to. For us, being interested in speech perception, the most relevant multisensory integration is that between auditory and visual information. Important to answer would be whether and how powerfully it can enhance our processing capacity. One paradigmatic study was conducted by Ross et al., where speech processing was observed when participants were presented with auditory input alone and contrasted with a condition in which additional visual information on articulatory movements was available. They also manipulated the signal-to-Noise Ratio (SNR) by introducing pink noise into the auditory signal and varying the loudness of the noise portion. With a louder noise signal on top of the auditory signal, the latter becomes less intelligible. With this they were able to see whether the quality of the single inputs has any effect. Their lowest SNR was 0, achieved with both the signal and the pink noise at 50db. In their highest noise condition the noise was 24db higher, resulting in a SNR of -24. The team compared 2 main conditions where in the Audio and Video condition a word was spoken and a corresponding video of the speakers lip movements was presented, while in the audio-condition the sound of the word was presented with only a still image of the speaker. On both conditions SNR was varied. They found an increase in correctness of understanding and identifying auditorily presented words by up to three times when

compared to the audio-only condition. The team observed that the integration seems to work best with medium SNRs (-12), 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 investigated the same phenomenon while recording neurophysiological activity through EEG. They extend on the findings by Ross et al. by examining continuous speech versus single syllables, providing a more naturalistic framework. They report 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)

### **2.1.1 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 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 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.

### **2.1.2 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 pseudo speech 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.

### **2.1.3 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 (McGurk and MacDonald, 1976). To produce the effect, they took a video of a speaker speaking a syllable of the structure consonant-vowel and replaced the phoneme in the auditory canal of the video with a different one. The replacement and the original form an auditory pair, one example

would be "ba" and "ga". If done correctly, an incredibly robust fusion occurs, where the visual information of the speaker's 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. For the previous example, the fusion product would be "da". When presented with a dubbed video, where the visual information is taken from the "ba"-video and the auditory information from the "ga"-video, most people consistently hear the speaker in the artificial video saying "da". 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. 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 <sup>2</sup> task, effectively showing that Multi-sensory Integration happens automatically and we cannot just disregard one modality stream of information in processing. 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.

## **2.2 Multi-sensory signal delays, 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 integration fails? 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 the 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.

---

<sup>2</sup>explain

A popular term here is the Temporal Window of Integration (TWIN), which specifies the timeframe within which Multisensory integration performs optimally. Outside of this window, integration effects are weaker and speech perception suffers.

van Wassenhove et al. (2007) and Team performed a classic simultaneity judgment (SJ) and an Identification Task in a separate experiment. In an 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 conclude that audiovisual (AV) integration works optimally 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 quite an unnatural situation and light can travel quite a bit faster than sound, usually arriving earlier at the individual. <sup>3</sup> 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.

### **2.2.1 Just Noticeable Difference (JND)**

Attunement? Closely related to the question about the size of TWIN 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 for us to realize there might be a problem? This is an interesting topic of research because this point does not seem to be fixed, it can vary, depending on the needs of the situation. This amazing ability is called Attunement. What's the minimum delay people can notice?

What about Sub-Noticeable Delays? Any Studies?

Klockgether and van de Par (2016)

---

<sup>3</sup>a common example would be how in an approaching storm the lightning is perceived sometimes seconds before the thunder.

## **2.3 The Echo Effect**

### **2.3.1 Delayed Auditory Feedback (DAF)**

Delayed auditory feedback classically occurs when 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,<sup>4</sup> 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

---

<sup>4</sup>for example, by applying a high- or low-pass filter to make the mid-range frequencies, which contain speech more present

for the hearing individual. For an interesting in-depth discussion of this trade-off see Lezzoum et al. (2016)

#### **2.3.4 On the temporal scale of audiovisual asynchrony**

With regards to our study examining temporal asynchronies, an essential question to ask is what asynchronies have been used and can we make any prior claims about certain ranges?

TODO: transition

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 a 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 an 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 an unambiguous impact on speech reception performance. Several TWIN studies suggest that the speech-specific audiovisual TWIN is asymmetric being larger for visual-leading stimuli over auditory-leading stimuli. A likely explanation is the prevalence of this type of asynchronies in nature, due to light travelling faster than sound. (van Wassenhove et al., 2007; Maier et al., 2011) Here, the optimally performing temporal window is estimated to be around 200ms, from -30ms to 170ms. At larger delays, the integrative capacity is still present but gradually declines. Our value should have clear effects, so we want to remove compensation effects from a possibly strongly interfering TWIN, resulting in our value having to be

larger than the optimal performing TWIN. Ideally, we set it quite a bit larger since even for delays larger than 200ms we can still expect some multisensory integration, specifically audiovisual integration happening although likely with less efficiency. We can assume this through the loosely gaussian-shaped response patterns in typical SJ Tasks (Maier et al., 2011). A more recent audiovisual delay study Li et al. (2021) noted that in a standard audiovisual simultaneity judgement (SJ) Task with stepped delays from -400 to 400ms delay, roughly 50 per cent of the participants incorrectly judged the 200ms delayed stimulus to be synchronous. Even in the 400ms condition, around 10 per cent of the 27 participants still judged the stimulus as being synchronous. This leads us to think that the temporal corrective capacity of some underlying sensory integration mechanism is surprisingly strong when it can in some cases correct for up to 400ms delay. This effect seems to be slightly stronger even when both the auditive and visual part of the stimulus are causally related and therefore more predictable. They also looked at conditions where this causal link was impaired by either blurring the video or the audio and found that for the causally less related conditions, they received less "synchronous" responses, suggesting that also in our study, when interference (the artificial echo) is present, we would expect a more accurate performance of the participants. Similarly, a temporal order judgement (TOJ) and SJ task setup Maier et al. (2011) provided evidence that stimuli with a subjective auditory lag in the range of up to 200ms are still highly likely to be judged synchronous, corresponding roughly to the previously defined "optimal" TWIN performance. For larger audiovisual delays they measured up to 267ms visual leading to subjective delay (TODO: subjective?) where still less than 80% of the participants were able to identify the stimulus as asynchronous. They also investigated spectrally rotated and temporally reversed speech, reporting that the TWIN in these conditions got larger, resulting in a worse performance of the participants in the SJ task. This points at a highly specified recognition system for speech that is not purely dependent on causal correlations but hints at some specialized statistical recognizer for natural language also being present. This provides further evidence that, to create a condition in which the majority of participants clearly can identify a temporal lag between the visual and auditory stimulus, the lag between them would have

to be around 400ms.

For the smaller value, the situation is more unclear: 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.

However, we still need to assess whether this holds for language-specific stimuli and whether there are more findings regarding the specific combination of senses involved in our experiment: audiovisual integration.

It has to be acknowledged that our study being browser-based has technical limitations being discussed Bridges et al. (2020). The authors, which are the same team developing PsychoPy (Peirce et al., 2019). In their timing study, they specifically looked at lag, taken to mean a constant error, and variance, representing an unpredictable error occurring more or less randomly. Looking at the experiment package paired with the software setup we estimate to be prevalent among participants, Psychopy via pavlovia.org executed within Chrome browser on a Windows 10 machine, we can expect on average a variance in reaction time (RT) of 0.39ms and variance of audiovisual synchronicity of 3.01ms. These values would be slightly higher for Edge users and even larger, but still slightly under 6ms for Firefox users. Concurring with the authors, we mostly disregard the lag, since a constant error will not affect the significance results between conditions. Further, differences in internet speed should be disregarded since all resources should be loaded and read from the disk at the time of the RT measurement. Another significant factor could be the screen resolution of the participant as drawing more pixels will take more time. We are recording the screen resolution the experiment is conducted on and will be able to tell whether it interacts significantly with RT after experimenting. Regarding Hardware, the experiment makes no use of the computer mouse, eliminating errors from different types of input devices. From a standard keyboard, where we record the responses, we expect a rather constant lag of around 20-40ms (Bridges et al., 2020), which we should also be able to disregard. All this leaves us with roughly 4ms of variance and no control over the type of graphics rendering device used. Taken together this results in us expecting the smallest

meaningful results at an audiovisual asynchrony of at least 10ms.

The literature seems quite divided on the question of what temporal difference the subjects can reliably detect. 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 delays 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 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) The team of Goehring et al. (2018) did not only look at audiovisual DAF but took also external voices into account. They looked at 20 NH and 20 HI participants and presented modified sound signals to them via circumaural headphones asking for their subjective annoyance rating. Divided into three conditions, they investigated delayed own voice (DAF), as well as unattenuated external voice and 20db attenuated external voice. The tolerance for external voices is much more interesting for us, due since this more accurately reflects general speech perception in the real world. They found slightly elevated annoyance ratings in the unattenuated condition for the NH participants, which interestingly disappeared in the attenuated condition, showing the first notable increase in annoyance between 20 and 30ms. Since we attenuated the echo in our conditions where an auditive echo is present, we should expect a similar timeframe. At large, HI Participants were more tolerant towards auditory delay, with the authors suggesting that experience in using hearing aids likely enlarges the delay tolerance in participants. They also note that the delay tolerance in their setup linearly scaled with hearing loss severity. The team of Lezzoum et al. (2016) looked at simulated echoes found that the smallest speech-related echo was detected by at least 20 per cent of the participants at 16ms delay.

TODO: background



For simple non-speech stimuli, the asynchrony 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 a 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, will later apply to HI subjects too. TODO: cite

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.

## 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 for impaired speech perception.

## 2.5 Autism Spectrum Disorder and possible differences to neurotypical

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.

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 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) TODO Elaborate

## 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 an SHPD are critically missing.

### 2.6.1 Different Values in the Literature

**Table 2.1: Caption above table.**

Reference	What was Measured	Measurement	Range	Participants	Setup	Device
Lezzoum et al. (2016)	Echo Threshold (20 % noticed echo)	16ms (shallow) 28ms (deep)	8ms (bell) - 68ms (noisy speech)	20 NH	asked people to identify echo threshold (manipulate slider)	
Stone and Moore (2002)	DAF disturbance and speech production rate	15ms clean, 20ms noisy, 30ms rate affected	7 - 43ms (4 discrete steps)	32 NH	Clean vs noisy background	
Maier et al. (2011)	TWIN synchronicity (own/other voice)	67ms highest SJ response/most balanced TOJ	-333ms - 333ms	9 NH	SJ and TOJ task	
van Wassenhove et al. (2007)	AV TWIN via McGurk Fusion	-30 ms to +170 ms	-467 ms - 467 ms	43 NH	McGurk Effect and SJ Task	
Zakis et al. (2012)	Delay detection in music	3-4ms not reliably detected	1.4ms - 3.4 ms	12 HI musicians	blind paired comparisons, preference rating	behind-the-ear, open-canal hearing aids
Agnew and Thornton (2000)	DAF	3-5ms noticeable, 30ms objectionable	noticeable effects ranged 2.15ms - 7.04ms	18 NH	delay slider manipulation	DSP hearing aid
Li et al. (2021)	audio-visual onset asynchrony (AVOA)	50 % judged the 200ms delay as synchronous, 400ms	Five SOAs (-400, -200, 0, 200, and 400ms)	27 NH	SJ Task low/high causality via blurring	Speakers
Vroomen and Keetels (2010)	Meta-study, AV temporal asynchrony	i 20ms usually unnoticed	-	-	SJ and TOJ	-
Petrini et al. (2009)	TWIN for AV Drumming	112 ms (TWIN), 80ms highest SJ	-266ms - 266ms	34 NH, 17 were expert	SJ Task	Speakers

### 3 Experiment

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 specifically and language processing in general. Many interesting questions will remain unaddressed, for example, whether a very small latency affects children differently, especially while learning language capacity 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. To later be able to transfer the experiment paradigm to participants on the autism spectrum, which may not have fully developed linguistic capabilities, 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 delay stemming from the processing time necessary for complex digital filtering methods. Any SHPD will have some amount of processing time, before relaying the attenuated original speech signal to the wearer. Therefore there will be some more or less constant delay in the auditive sensory information. Under conditions where the wearer has access to visual information, for example is seeing the speakers lip movements, there will be an audiovisual delay present. Since SHPDs are not perfect noise isolators, most filter only around 20db **TODO cite**, any auditive stimulus louder than that will still be heard, although in attenuated form. When that happens, the wearer will hear an inverted echo, meaning that the original but attenuated signal will be heard first and after a delay the filtered signal will be present, potentially overlapping with the original. We also want to examine this scenario by simulating this specific echo. To simulate, we will introduce additional conditions where the original auditory signal is present in attenuated form, conflicting with the delayed stimulus, creating an echo. 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 (RT) as the operating variable, with the assumption that RT provides a direct index of the time that

was needed to sufficiently process the linguistic signal, whether it is primarily auditory, visual or both in order to respond to the task. We also record the accuracy to be able to detect any secondary effects. The task was a referent identification task, in which the participant has to choose which entity is modified by the target adjective. To prevent inferential problem solving, we introduce filler trials where the target is misleading and not modifying any of the referents.

Prior studies are reluctant to come up with a concrete number for the audiovisual delay that can be utilized with engineering in mind, partly because there is a lack of clarity on different terms. The earlier introduced "just noticeable difference" relies on self-report and as such is hard to measure indirectly. This experiment enables us to compare different audiovisual delay conditions without reliance on subjective feedback of whether a delay was perceived or not. This means that we can measure how processing is affected without requiring explicit judgments on the nature of the signal from participants. Due to the uncoupling of conscious experience and speech perception performance, we can now gain insight on very small audiovisual delays and attenuated echoes without the need for the participant to perceive and report a delay, effectively eliminating a lower boundary of testing present in JND paradigms.

## Assumptions

- There is a universal underlying mechanism of multimodal 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.
- The auditory noise present in the videos due to recording quality has no significant effect.
- Hardware differences, as well as resulting visual and auditory artifacts are consistent.

## 3.1 Method

### 3.1.1 Participants


The experiment recruited 60 participants via the university’s internal mailing list targeting cognitive science and psychology students. All participants were native German-speaking adults with normal or corrected to normal vision and normal hearing. The participants had a male to female ratio of 1:1, with a mean age of 25.

TODO upload graphs

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. The experiment was approved by the ethics committee of Osnabrück University. Participants could receive Experiment Hours (VP-Stunden) as compensation, ~~a mandatory part to finish the Cognitive Science program.~~ No other compensation was granted. We required our participants to wear wired 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 optionally recorded prior psychological diagnoses or diagnoses with ASD.

### 3.1.2 Materials

**Media files** Video and corresponding audio files are used with friendly permission by the original creators of the OLAKS Corpus Uslar et al. (2013); Rosemann and Thiel (2018). These are full HD recordings of a male German native speaker reading German sentences centered onto his lips. 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 selectively use 80, all of which follow an SVO structure, where both the subject and the object is modified respectively. One example sentence taken from the corpus is

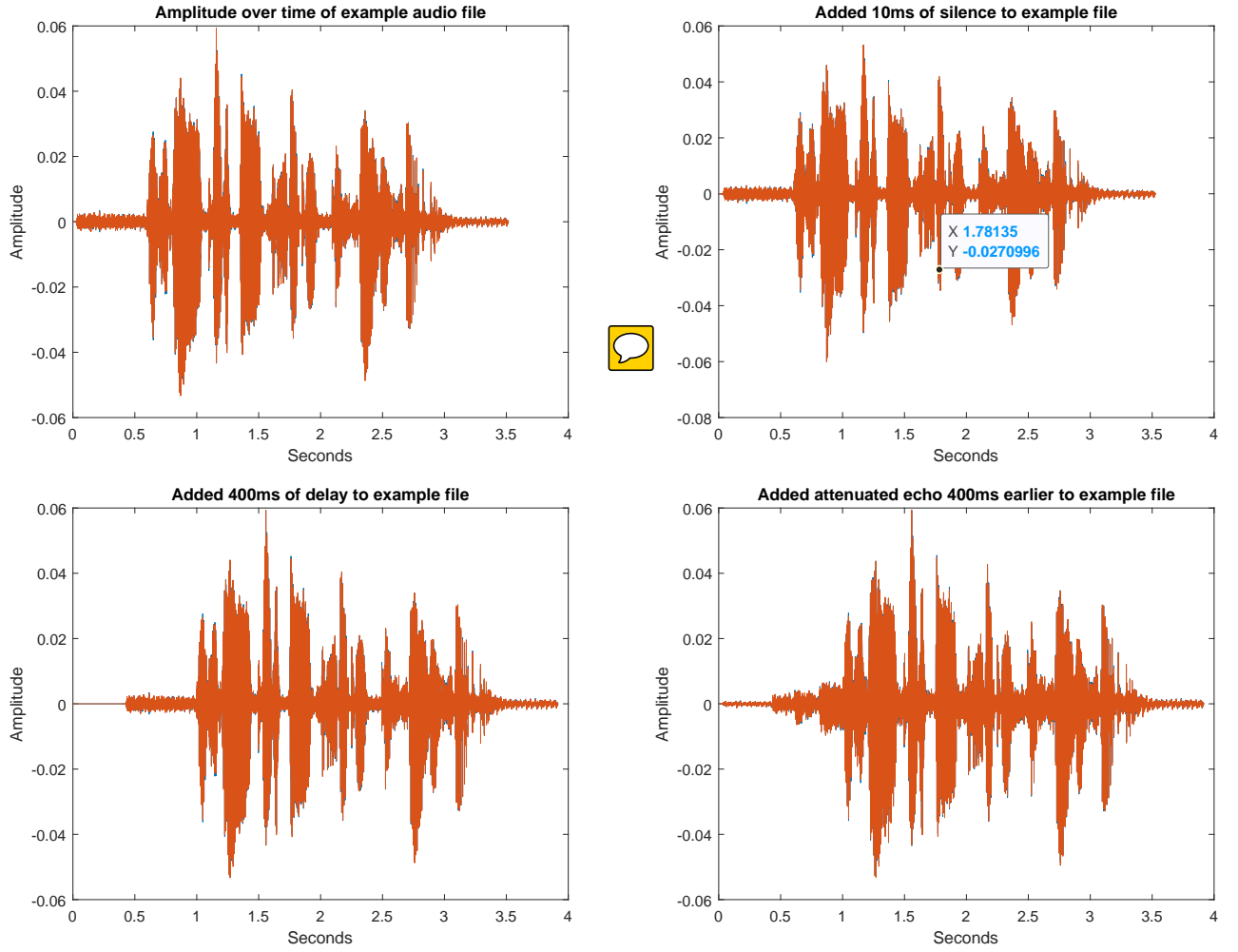
“Den alten Pfarrer grüßt der kluge Pilot.” 

Each sentence contains two entities, each modified by an adjective and a verb con-

necting the two entities. The entities are either animals, professions or other simple and unambiguous terms. Audio and video stream are separated, the audio stream is then modified using Matlab (MATLAB, 2020), adding the necessary delay and transforming and adding the attuned echo with proprietary code supplied by the CRITIAS Lab (Lezzoum et al., 2016). TODO: Details about code

The results obtained by modification are visualized here:






**Figure 3.1: Visible effects of modification on an example file**

TODO fix image

The video and audio streams are then merged and compressed using FFmpeg (Tomar, 2006) into h.264 mpeg4 format, which is compatible with most modern browsers. The audio stream is left as-is, repackaged into an aac mp4 format with a sampling rate of 48kHz, 32bits/sample, which corresponds to the original. Due to browser playback issues during testing, the videos are compressed using built-in FFmpeg compression for h.264 and resized to 1280x720px resolution. The original frame rate of 25fps is left as-is to leave synchrony intact. For each condition, a separate file is generated resulting in  $5 \times 80 = 400$  stimuli. The merging is done prior to the experiment to minimize av-synchrony issues resulting from different media playback handling in different browsers.



**Figure 3.2: Example of image files: Pfarrer, Pilot**

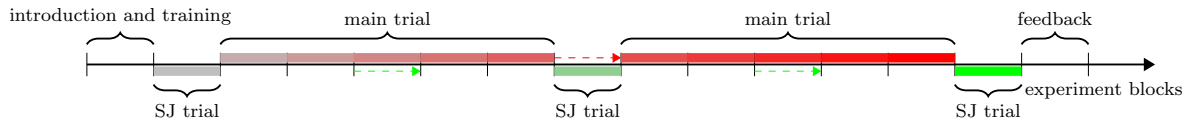
**Images** The majority of corresponding images are taken from the internationally tested MultiPic Corpus (Duñabeitia et al., 2018), a set of hand-drawn colored files in .png format with available data for measured complexity and percentage of correct recognition in a German speaking population. These are supplemented with a range of copyright-free images to cover each subject appearing in the text corpus not having a counterpart in the MultiPic corpus. All of these are then manipulated using GIMP 2.10.22, centered on a quadratic canvas with transparent background, all resolutions ranging from 500 to 1200 pixels. 

**Words** The presented target words are extracted from the sentences in the media files and then stored in a dictionary table and presented via the internal functions of the experiment software.

TODO: Image of target presentation

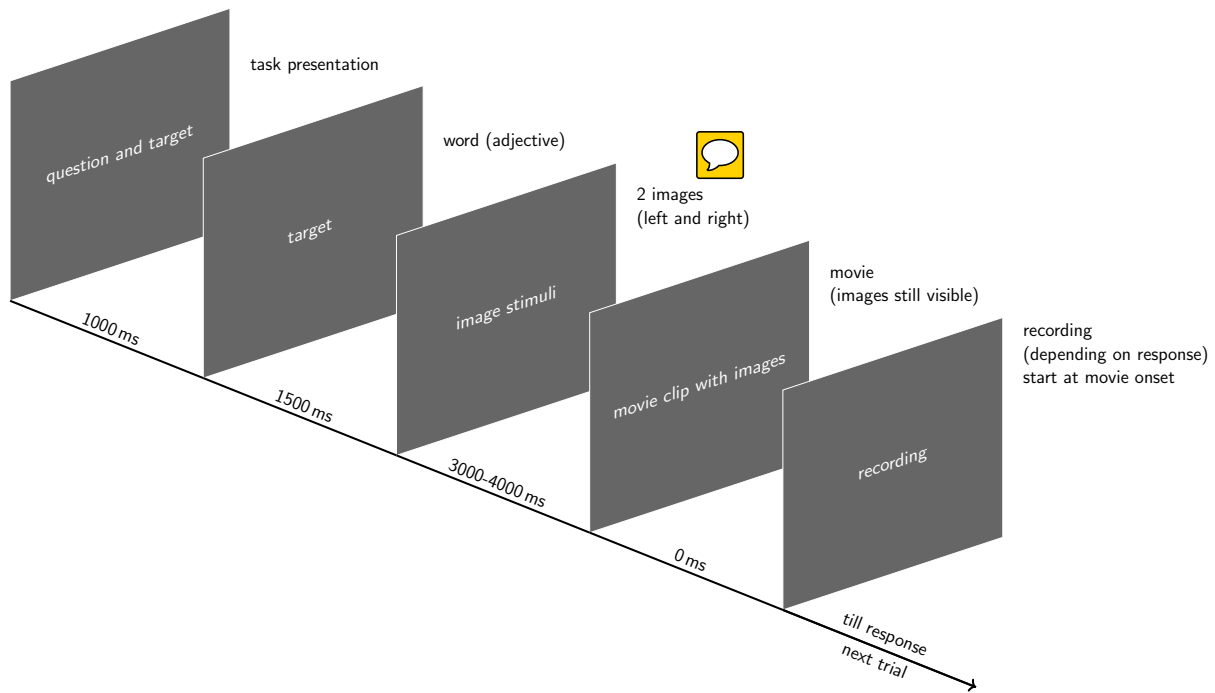
### 3.1.3 Procedure

Participants are lead via browser link to an introduction page, explaining the tasks, listing the requirements, and explaining the general purpose of the experiment. Here, consent is collected and participants are instructed on how to abort the experiment and withdraw





**Figure 3.3: Temporal order of entire experiment**

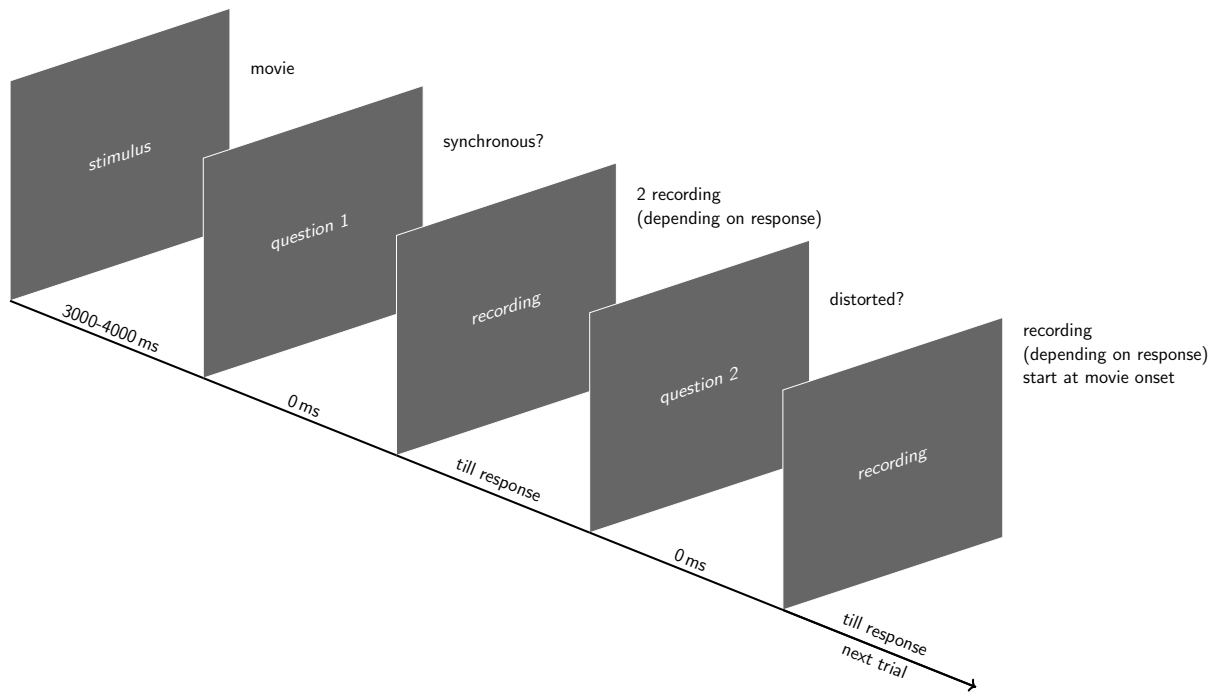
their consent. After consent, pseudonymous data is collected, such as age, gender, vision, and hearing capacity. We request the participant to be alone in the room and eliminate possible noise interference. We ask participants to complete the experiment on a laptop or computer, ideally sitting on a desk in a fixed position, roughly 60 cm away from the screen. ~~After the chance to correct missing requirements, such as switching to a desktop computer~~ or enabling headphones and adjusting the sound level to a comfortable level comparable to a face-to-face conversation, participants are redirected to another browser window playing in fullscreen, which contains the entire experiment. After the instructions and repeating the task, they receive 5 trial runs to get familiar with the nature of the main task. Participants are then reminded to answer as fast as they possibly can, using only the middle and the index finger of their dominant hand. They are also instructed to ensure consistent viewing conditions: no direct sunlight and subjectively adequate brightness of the screen. The entire experiment is conducted in one browser session requiring internet access, a keyboard, wired headphones, and a display. Between each main trial block breaks are inserted and not time-restricted, the participant could choose for how long to take each break. The experiment is inaccessible from a mobile device and records the participants' operating system, the frame rate, resolution, and the browser used. All stimuli of the experiment are downloaded before starting to prevent and mitigate download speed, performance, or playback issues. The main trials are organized in pseudorandom blocks with constraints preventing identical targets, conditions, images, and sentences in sequence.



**Figure 3.4: Temporal order of presentation in main task**

**Main Task**  The main task consists of 10 blocks with 16 trials each, where each sentence is used twice: once with the target being the first entity, once with the target being the last entity. Each trial is divided into target presentation, stimulus presentation and records the response starting with the onset of the video stimulus. The target, either one of the two adjectives or a randomly chosen adjective that is present in another trial, is flashed for 2500ms. Then, in the stimulus phase, 2 images are shown, corresponding to the left or right answer option indicated by the position of the images and helping arrows. An upward arrow is presented alongside to remind the participant to press the upper arrow when no image fits. After another 2500ms the audiovisual stimulus is presented, after being primed by a fixation cross. The video clip is presented centered in the upper half of the screen, alongside the images in the lower half. The trial ends with a keypress registration of the answer, there is no hard upper response time limit. The background is white throughout the entire experiment. In the inter-trial break, trial progress is presented, and there is no minimum time between blocks.

**sj task**  The modified SJ task is performed 3 times with 10 random trials, 2 repetitions of each of the 5 conditions. The blocks are distributed before, in the middle, and after the



**Figure 3.5: Temporal order of presentation in adapted SJ task**

main task. Each block consists of the same 10 sentences also taken from the OLAKS set, but not presented in the main task. The participant, after watching one video per trial, is then asked whether the audio was perceived to be synchronous and whether any auditive distortion was perceived. These questions are asked in sequence after it has finished playing on the same stimulus. For each question, correctness and RT are measured. The answer is recorded via a keyboard press with separate buttons for yes and no. The buttons are not changed throughout the experiment, the mapping stays invariant.

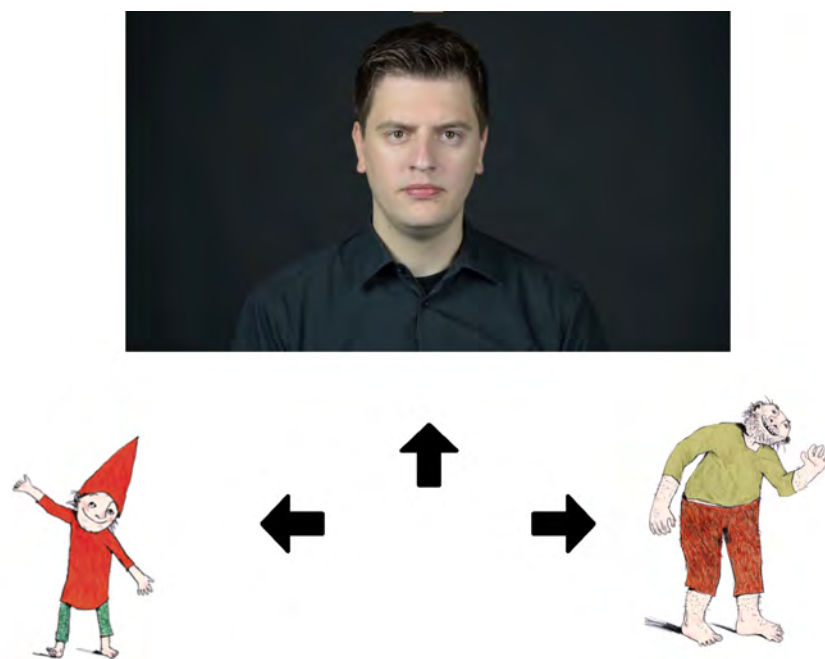


Figure 3.6: The stimulus presentation in the main task

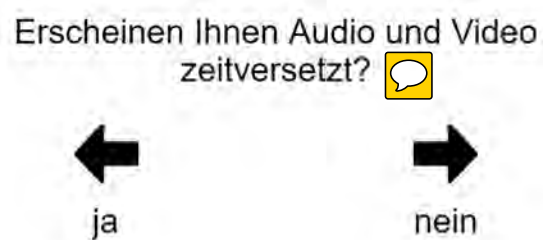


Figure 3.7: Presentation of questions with response indicators

### 3.1.4 Hypothesis

**Table 3.1: Dependent Variables (DV)**

Symbol	Variable	Measurement
rt	reaction time	measured from the onset of the stimulus video
acc	accuracy	registered as either correct or incorrect



TODO fix that shitty table alignment

**Table 3.2: Independent Variables (IV)**

Symbol	Variable	Values
		0ms latency (no latency)
lat	latency of audio to visual stimulus	0ms latency, 400ms latency,
		0ms echo (no echo),
ech	Attenuated inverse Echo	0ms echo, 400ms echo

As the main effect, we expect that when presented with greater temporal dis-alignment of sensory input, a degraded multisensory integration will result in more time needed to process the linguistic signal. Processing of the linguistic signal 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 the shortest in the base condition, 0ms latency, 0ms echo, and the RTs will be higher for latency and echo conditions, respectively. With more interference for the processing by either temporal disalignment or the attenuated echo, effectively presenting degraded input signals, we also expect the accuracy in responses to be lower. In short: linguistic processing will be both slower and less accurate under our artificial adverse conditions.

Since we introduced both a large modification and a small modification, we expect the small modification (10ms delay, 10ms echo) to be below conscious detection thresholds, reflected in incorrect responses in the secondary trial. The intent of the large modification

(400ms delay, 400ms echo) is to verify that linguistic processing is indeed linearly dependent on synchrony and noninterference. Therefore, we expect largely correct identification in the secondary task. Since in the secondary task linguistic processing is not directly verified and participants are asked only to detect asynchrony and auditory distortion, the large modification conditions are easier to solve and will present lower RT.

## 3.2 Results

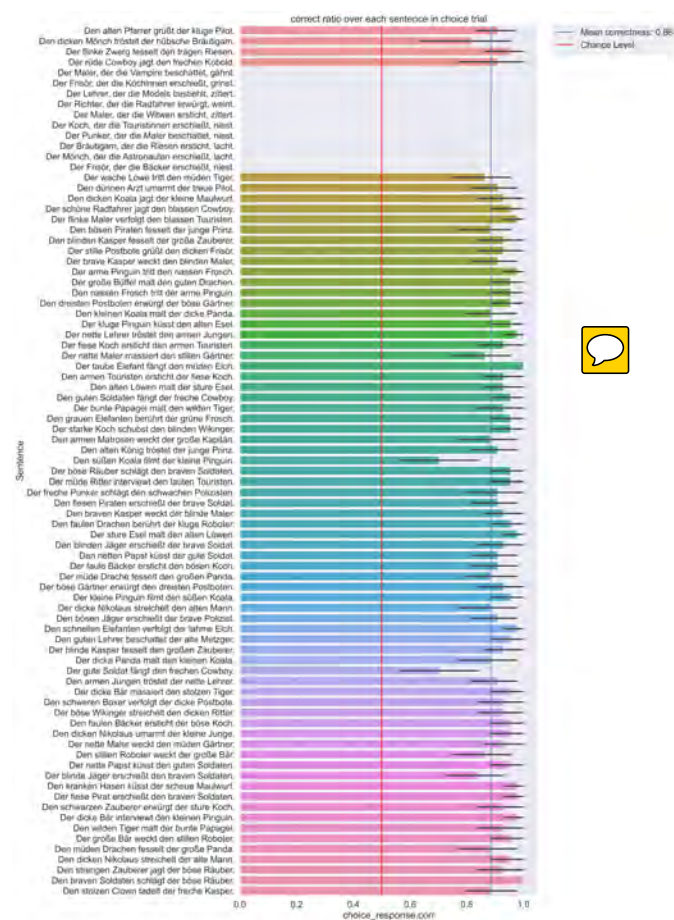


Figure 3.8: Example result table

### 3.2.1 Statistical Analysis

RStudio Team (2020)

~~Spectrogram Analysis~~

Descriptive Statistics



~~Analysis of the Variance~~

## 4 Discussion

4.1 My Results in other current research

4.2 Later studies with ASD

4.3 Conclusion

4.4 Suggestions for further research

## 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.
- 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.

- Goehring, T., Chapman, J. L., Bleeck, S., and Monaghan\*, J. J. (2018). Tolerable delay for speech production and perception: effects of hearing ability and experience with hearing aids. *International journal of audiology*, 57(1):61–68.
- 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.
- Li, S., Ding, Q., Yuan, Y., and Yue, Z. (2021). Audio-visual causality and stimulus reliability affect audio-visual synchrony perception. *Frontiers in Psychology*, 12:395.
- 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.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- 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.
- Noel, J.-P., De Nier, 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.
- 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.
- Vroomen, J. and Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, perception psychophysics*, 72:871–84.
- Zakis, J. A., Fulton, B., and Steele, B. R. (2012). Preferred delay and phase-frequency response of open-canal hearing aids with music at low insertion gain. *International Journal of Audiology*, 51(12):906–913.

# A Appendix

## A.1 Stimuli

### A.1.1 Images

### A.1.2 Sentences

Here is a full list of all the sentences taken from the OLAKS Corpus and appearing in the experiment.

1. Der schlaue Kasper beschattet den faulen Vater.
2. Der blinde Jäger erschießt den braven Soldaten.
3. Der fiese Pirat erschießt den braven Soldaten.
4. Der faule Bäcker ersticht den bösen Koch.
5. Der fiese Koch ersticht den armen Touristen.
6. Der böse Gärtner erwürgt den dreisten Postboten.
7. Der taube Elefant fängt den müden Elch.
8. Der gute Soldat fängt den frechen Cowboy.
9. Der blinde Kasper fesselt den großen Zauberer.
10. Der müde Drache fesselt den großen Panda.
11. Der flinke Zwerg fesselt den trägen Riesen.
12. Der kleine Pinguin filmt den süßen Koala.
13. Der stille Postbote grüßt den dicken Frisör.
14. Der müde Ritter interviewt den lauten Touristen.
15. Der dicke Bär interviewt den kleinen Pinguin.
16. Der rüde Cowboy jagt den frechen Kobold.
17. Der schöne Radfahrer jagt den blassen Cowboy.
18. Der süße Junge küsst den lieben Vater.
19. Der nette Papst küsst den guten Soldaten.
20. Der kluge Pinguin küsst den alten Esel.
21. Der dicke Panda malt den kleinen Koala.
22. Der sture Esel malt den alten Löwen.
23. Der große Büffel malt den guten Drachen.
24. Der bunte Papagei malt den wilden Tiger.
25. Der dicke Bär massiert den stolzen Tiger.
26. Der nette Maler massiert den stillen Gärtner.
27. Der böse Räuber schlägt den braven Soldaten.
28. Der freche Punker schlägt den schwachen Polizisten.
29. Der starke Koch schubst den blinden Wikinger.
30. Der dicke Nikolaus streichelt den alten Mann.
31. Der böse Wikinger streichelt den dicken Ritter.
32. Der böse Zauberer tadelt den frechen Kobold.
33. Der arme Pinguin tritt den nassen Frosch.
34. Der wache Löwe tritt den müden Tiger.
35. Der nette Lehrer tröstet den armen Jungen.
36. Der flinke Maler verfolgt den blassen Touristen.
37. Der nette Maler weckt den müden Gärtner.
38. Der große Bär weckt den stillen Roboter.
39. Der brave Kasper weckt den blinden Maler.

40. Den faulen Drachen berührt der kluge Roboter.
41. Den grauen Elefanten berührt der grüne Frosch.
42. Den guten Lehrer beschattet der alte Metzger.
43. Den blinden Jäger erschießt der brave Soldat.
44. Den bösen Jäger erschießt der brave Polizist.
45. Den fiesen Piraten erschießt der brave Soldat.
46. Den faulen Bäcker ersticht der böse Koch.
47. Den armen Touristen ersticht der fiese Koch.
48. Den dreisten Postboten erwürgt der böse Gärtner.
49. Den schwarzen Zauberer erwürgt der sture Koch.
50. Den guten Soldaten fängt der freche Cowboy.
51. Den blinden Kasper fesselt der große Zauberer.
52. Den bösen Piraten fesselt der junge Prinz.
53. Den müden Drachen fesselt der große Panda.
54. Den süßen Koala filmt der kleine Pinguin.
55. Den alten Pfarrer grüßt der kluge Pilot.
56. Den strengen Zauberer jagt der böse Räuber.
57. Den frechen Kobold jagt der rüde Cowboy.
58. Den dicken Koala jagt der kleine Maulwurf.
59. Den netten Papst küsst der gute Soldat.
60. Den kranken Hasen küsst der scheue Maulwurf.
61. Den kleinen Koala malt der dicke Panda.
62. Den alten Löwen malt der sture Esel.
63. Den wilden Tiger malt der bunte Papagei.
64. Den braven Soldaten schlägt der böse Räuber.
65. Den starken Touristen schubst der lahme Bauer.
66. Den dicken Nikolaus streichelt der alte Mann.
67. Den stolzen Clown tadelt der freche Kasper.
68. Den frechen Kobold tadelt der böse Zauberer.
69. Den nassen Frosch tritt der arme Pinguin.
70. Den alten König tröstet der junge Prinz.
71. Den armen Jungen tröstet der nette Lehrer.
72. Den dicken Mönch tröstet der hübsche Bräutigam.
73. Den dünnen Arzt umarmt der treue Pilot.
74. Den dicken Nikolaus umarmt der kleine Junge.
75. Den schweren Boxer verfolgt der dicke Postbote.
76. Den schnellen Elefanten verfolgt der lahme Elch.
77. Den stillen Roboter weckt der große Bär.
78. Den braven Kasper weckt der blinde Maler.
79. Den armen Matrosen weckt der große Kapitän.
80. Der grobe Riese ersticht den scheuen Piloten.
81. Der Papst, der die Detektive berührt, gähnt.
82. Der Punker, der die Maler beschattet, niest.
83. Der Maler, der die Vampire beschattet, gähnt.
84. Der Lehrer, der die Models bestiehlt, zittert.
85. Der Mönch, der die Astronauten erschießt, lacht.
86. Der Frisör, der die Bäcker erschießt, niest.
87. Der Frisör, der die Köchinnen erschießt, grinst.
88. Der Koch, der die Touristinnen erschießt, niest.
89. Der Bräutigam, der die Riesen ersticht, lacht.
90. Der Maler, der die Witwen ersticht, zittert.
91. Der Richter, der die Radfahrer erwürgt, weint.
92. Der Bauer, der die Ärztinnen fängt, lächelt.

## A.2 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.3 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.

Osnabrück, April 28, 2021

A handwritten signature in black ink that reads "Aron Petau". The letters are cursive and fluid.

Aron Petau

---

city, date

---

signature