Universität Osnabrück

Fachbereich Humanwissenschaften

Institute of Cognitive Science

Audio-Visual Speech Processing and effects of multisensory asyncronicity

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

Universität 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:** multisensory integration, smart hearing protection, SHPD, Echo Effect, sensory asynchrony, autism spectrum disorder, multi-modal recalibration, Speech processing under temporal lag, temporal window of integration, just noticeable difference

Contents

List of Figures

1 Introduction

Speech perception inherently is a multisensory process. In any common conversation a listener will be exposed to auditory information of the uttered speech as well as some form of visual information, whether that be gestural or facial expressions. Speech information stemming from visual and auditory modalities complement each other (Kavanagh et al., 1972) and speech perception seems to be able to extract information from all of them. To conduct reseach on audiovisual asynchrony in speech it is necessary to understand how multiple sensory modalities are processed and contribute to perception of the world and speech perception specifically. I will conduct a literature review of prior findings in the field of multisensory integration, discuss those, and subsequently present my own experiment. This aims to observe the effects of delays in auditory speech signals that would occur when utilizing selective digital attenuation for background noise or distressing sounds. Such selective attenuation could be of great impact, especially for non-neurotypical people diagnosed with autism spectrum disorder (ASD), as defined in APA (2013). 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 when compared to neurotypical individuals (NT). Speculatively, this is a major reason for individuals with ASD to take longer to develop linguistic skills during childhood, in severe cases even remaining completely nonlinguistic, with prominent features of ASD being difficulties in speech and social interaction (APA, 2013). As will be pointed out, we believe that a slower and less flexible multimodal integration is responsible, (see Stevenson et al. (2014)) and there is a possible remedy in smart hearing protection devices (SHPDs). When an individual has trouble overcoming problems in speech processing due to sound defects, environmental noise and other adverse factors, it makes sense to try and eliminate the present signal defects to improve speech perception ability. Presenting a cleaner signal overall will increase performance in speech perception, as goes the idea.

TODO maybe cite lezzoum?

To achieve this, SHPDs have an inherent advantage over traditional hearing protection devices (HPDs), namely, that they are able to employ digital attenuation and filtering

methods before transmitting the sound to the wearer.

Attenuating an acoustic signal digitally introduces additional latency to the perception of that signal by the wearer. In the real world, this latency is virtually always present and scales with the distance of the sound source. Nevertheless, the temporal 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. Traditional commercially available HPDs (earplugs for example) present a physical barrier for sound, not introducing a relevant additional delay between the visual and the auditory signal. The situation is different when a digital sound attenuating device is used. Regardless of how the HPD works, (analog or digital), any attenuation requires processing and produces some auditory delay before it can transmit the filtered auditory information. Modern smart hearing protection devices (SHPD) can employ complex attenuation and noise filtering that goes beyond frequency filtering, which is not differentiating between types of sound. Advanced filtering techniques mean that the input is digitally processed in-situ, in "real-time" and it can be generalized that the processing time positively scales with complexity of the filtering mechanism applied. Thus open up many interesting research questions regarding how such a SHPD could operate and it poses rather unique and new challenges.

Research Focus My 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. Ideally, there 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 age is a relevant factor, whether children are affected by asynchrony differently compared to adults, 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, making it easily extensible to different demographics. To later be able to transfer the experimental 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 small, delay in the transmission of the auditory sensory information. For earplug-type HPDs this is typically in the $<100\,\mu s$ range, which is insignificant as stated in Lezzoum et al. (2016).

The case is different when looking at more complex attenuation processes employed in SHPDs Under conditions where the wearer has access to visual information, for example seeing the speaker's lip movements, there will be an audiovisual delay present. The effects of this delay on speech understanding are unclear, especially when looking at small, sub-perceptible delays. Since SHPDs are not perfect noise isolators, most SHPDs, like all HPDs, only attenuate the original signal around 25db using the personal attenuation rating (PAR) metric further explained in (Samelli et al., 2018). Any auditory stimulus louder will still be heard, although in an attenuated form. When that happens, the wearer will hear an inverted echo, meaning that the original attenuated signal will be perceived first and after a delay the attenuated signal will be present, potentially overlapping with the original. In a situation like this, we would then have multiple things going on: The asynchrony between the visual stimulus and the transmitted auditory signal, as well as the attenuated original auditory signal. Our experiment aims at investigating language processing under roughly the conditions a wearer of such an SHPD is under, which is why 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 an attenuated form, conflicting with the delayed stimulus, creating an echo.

TODO rework

2 Literature Review

It has long been known that congruous and synchronized visual input greatly aids people's ability to perceive audio information and to understand natural language. Sumby and Pollack (1954) already looked at the influence of visual sensory input on speech intelligibility. Seeing the speaker's lips especially helps in making sense of what is being talked about, as presented by 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 research field of multisensory 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,[1] concerning multimodal integration.

The goal is to take a look at the current state of research and provide a background on multisensory integration and what we already know about the effects of temporal asynchrony across multiple sensory modalities [2] and investigate the basis of our experiment, such that after the review we can settle on an experimental design and formulate an informed hypothesis. ~~This should be in line with the recent literature and capable of answering some of the open questions that are not already investigated and answered throughout this section.~~

2.1 Multisensory integration

The most prominent theory to date about how multiple streams of sensory information are merged into a coherent perception of the world was put forward in 1986 Meredith

[1] in studies often called TD - typically developed. For us, development is only a secondary concern. We, therefore, use the term "neurotypical" individuals to refer to the weaker notion of the current absence of neurological abnormalities

[2] a sensory modality, sometimes called stimulus modality refers to a specific type of sensory processing, e.g. the auditory modality. Commonly, we would refer to this as a sense.

and Stein (1986), 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 termed "multisensory", proving that multisensory convergence is a common and essential concept in sensoric processing. In their later book, Stein and Meredith (1993) built 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 multisensory integration. The idea is that redundant, overlapping, and sometimes mutual ly exclusive sensoric information from all modalities has to be integrated by the nervous system to form the coherent picture of our environment that we are used to. From our unitary [3] perception of the world it follows that at some point during the processing of sensory input it has to contribute to a general perception of the world.

Why do you have to switch off the radio when you try to park the car? So you can see better and concentrate on the relevant sensory input. From this small example from our abundant everyday experience of sensoric perception, it is already evident that different sensory pathways (modalities) are linked in the brain and can at the very least influence our experienced perceptual performance and each other. This range of phenomena is researched since over a century ago, a notable early example being Stratton (1896), who experimented with vision-distorting glasses, finding that he was quickly able to adapt to the sensoric discrepancy between inverted vision and haptic feedback of his environment. For us, being interested in speech perception, the most relevant multisensory interaction is that between auditory and visual information. ~~Important to answer would be whether and how powerfully multisensory integration impacts our processing capacity. Some examples of where this phenomenon can enhance our perception of the world come from our exceptional ability to synchronize to rhythmical stimuli. A study by Iversen et al. (2015) challenged the idea that our timing and synchronization abilities are bound to a specific modality and that these modalities are mostly specialized to certain tasks by comparing hearing and deaf individuals. Finding no impairment in rhythmic synchronization in the deaf~~

[3] meaning that we perceive globally: an object can have a smell and a texture, and we can relate both to the same object

~~group, 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. Should this hold, we could infer for multisensory integration that there is likely no inherent preference for unilateral sensory integration. It should be in principle true, that visual information can support auditory stimuli and vice versa. Importantly, perceived secondary input seems not simply to aid specific uni-modal~~ [4] ~~processing, but the higher-order processing seems to be largely agnostic regarding the modality of the input stimulus, as brought forward by 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 well within the framework of multisensory integration. Specifically for speech and gestures, synchronizing effects between gestures and speech have recently been demonstrated by Pouw and Dixon (2019), who used motion tracking to observe participants' hand gestures while they were either exposed to delayed auditory feedback or heard themselves normally. They found that the benefits of audio-gestural integration were the biggest under adverse conditions (DAF) where subjects heard an echo of their own voice with a delay of 150ms as a distraction. They suggest that in noisy and other environments counterproductive to speech transmission, a stronger binding by synchrony of gestures and speech follows. This indicates that synchronization is an important aid to maintain speech rhythm stability under noise perturbation. But it is not only a factor in speech production, it likely also helps in speech perception under the same circumstances, as reported by Wang et al. (2018). 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

[4] a modality here means a sensory type-specific perception mechanism, one sensory modality would be for example vision.

the brain, which would be a good explanation as to how the integration is realized.

2.1.2 Speech and Visual Lip Movement

Another strong demonstration of multimodal 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. The study, being conducted on normally hearing participants with fMRI, showed that access to visual lip-reading information only was enough to specifically activate areas that are known to be involved in auditory language processing. This suggests that some multimodal integration has to occur where some mechanism processing the visual information can acquire additional resources from other areas and thereby adapt to more advanced tasks. Additionally, a counter-check with pseudo-speech and non-linguistic facial movements 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 mouthing real words or language-like pseudowords were presented. For nonlinguistic stimuli, no activation was present. This suggests that the measured activation is specific to language-related processing and routinely utilizes multisensory integration.

Another paradigmatic study was conducted by Ross et al. (2007), 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 an SNR of -24. ~~The team compared 2 main conditions wherein the audio and video condition a word was spoken and a corresponding video of the speaker's lip movements was presented, while in the audio condition the sound of the word was presented with only a still image of the~~

~~speaker~~. In both conditions the SNR was varied. They found an increase in the 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 adversarial noises occurring at almost all times. A more recent study was conducted by Crosse et al. (2015) investigating the same phenomenon while recording neurophysiological activity through EEG. They extend ~~on~~ the findings by Ross et al. (2007) 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.

2.1.3 The McGurk effect

Also essential in the context of multimodal 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 uttering a syllable of the structure consonant-vowel and replaced the phoneme in the auditory canal of the video clip with a different phoneme. The replacement and the original form an auditory pair [5], 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 that the third phoneme cannot be real. (Macdonald and McGurk,

[5] an auditory pair is formed when both syllables share some articulatory features, like ending on the same vocal.

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 [6] task, effectively showing that multisensory integration happens automatically and we cannot just disregard one modality stream of information in processing.

However, some research suggests that it is not a fine-grained enough measure to accurately assess audio-visual integration and may hinder research regarding the automaticity of integration 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 Multisensory signal delay, asynchrony, and temporal window of integration

Based on the framework of multisensory integration that was introduced in 2.1, a sensible question might be the limits of integration. 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. 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, the integration effects are weaker and speech perception suffers. van Wassenhove et al. (2007) performed a ~~classic~~ simultaneity judgment (SJ) ~~and an identification task [7] in a separate experiment~~.

[6] This paradigm is based on the idea that if two dimensions of a stimulus can be attended to independently, then irrelevant variations along one of the stimulus dimensions will not affect response latencies in a discrimination task regarding the other dimension (a classical example is color and shape). The reverse also holds: If two dimensions are perceptually dependent, then a distractor in one dimension will produce interference in the other dimension, one example would be color and hue.

[7] an identification task is any task where a participant has to discriminate a stimulus or tries to recognize it.

With this, they were trying to recreate the original findings by Sumby and Pollack (1954), who investigated audiovisual integration and the potential of one modality to enhance the other. In an SJ Task utilizing auditory and visual speech, the participant is presented with two stimuli temporally close together and has to decide whether those stimuli occurred simultaneously or not. With their findings they 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 temporal order of the modal information seems to matter. The authors find that TWIN is decidedly asymmetric in temporal lag and substantially larger when auditory stimuli were trailing 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. [8]

Further research suggesting that tolerance for visual-leading asynchronies is bigger can be found in Maier et al. (2011), who conducted a study where they compared different temporal offsets of the components in audiovisual stimuli. ~~Humans seem to be much more sensitive overall towards auditory-leading stimuli, which is likely explained by the relatively minor statistical occurrence in nature.~~ Investigating the difference in TWIN for speech and nonspeech stimuli using audiovisual simultaneity judgment (SJ) [9] and temporal order judgment (TOJ) [10] tasks, they found that this especially holds for speech perception specifically. This was indicated by a more narrow and asymmetric TWIN for unmodified speech stimuli in contrast to distorted or garbled speech, suggesting that, through the constant experience of speech, humans recognizing speech rely on a learned relationship between visual temporal cues and the auditory information, resulting in a tightened TWIN. When talking about perceived synchrony, it would be of interest whether we can quantify just how small a temporal difference can be noticed and whether there is a threshold of detectable asynchrony.

[8] a common example would be how in an approaching storm the lightning is perceived sometimes seconds before the thunder.

[9] participants are usually presented with 2 stimuli in different modalities that are either synchronous or follow each other and have to report whether they perceive those to be synchronous or not

[10] a TOJ task is similar to an SJ Task with the difference that the participant now has to report which stimulus was perceived first. Usually, there is no option to declare them as synchronous.

2.2.1 Just Noticeable Difference (JND)

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. Take the movie subtitle example again, how many milliseconds 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 ability is called Attunement.

What is the minimum delay people can notice? Vatakis and Spence (2006) looked at the sensitivity of NH participants towards audiovisual asynchrony for speech and nonspeech stimuli and found that the JND for speech is lower ~~(better)~~ than for other tested ~~musical (nonspeech)~~ stimuli and found the detectable threshold on average to be around 100ms in a TOJ task. Importantly, they used short video clips of single spoken syllables and reported a lower JND than studies using continuous speech (Grant et al., 2004) and larger JNDs than studies using short flashes as stimuli. Zampini et al. (2003) ~~for example, used LED flashes of just 9ms as visual stimuli and 9ms white noise bursts as auditory stimuli and got results of 80-90ms in their TOJ task.~~ This would suggest that the JND highly depends on the type of situation a person is perceiving speech in, wherein a continued discussion the JND would be much higher than for a short alarm signal, for example. Eg et al. (2015) They also stress that the audiovisual JND is highly dependent on the context and content of the conveyed information, but they also find that for speech-related stimuli the JND is smaller.

After having looked at the size of the perceptual threshold, we turn towards the effects of sub-perceptible asynchrony. Van der Burg et al. (2018) argue with the results of their study for the notion that that rapid temporal recalibration is determined by the physical timing of the preceding events, not by prior perceptual decisions. If this holds, sub-perceptual temporal lags would indeed influence our speech perception without a detectable change in the percept decisions taken in a test set.

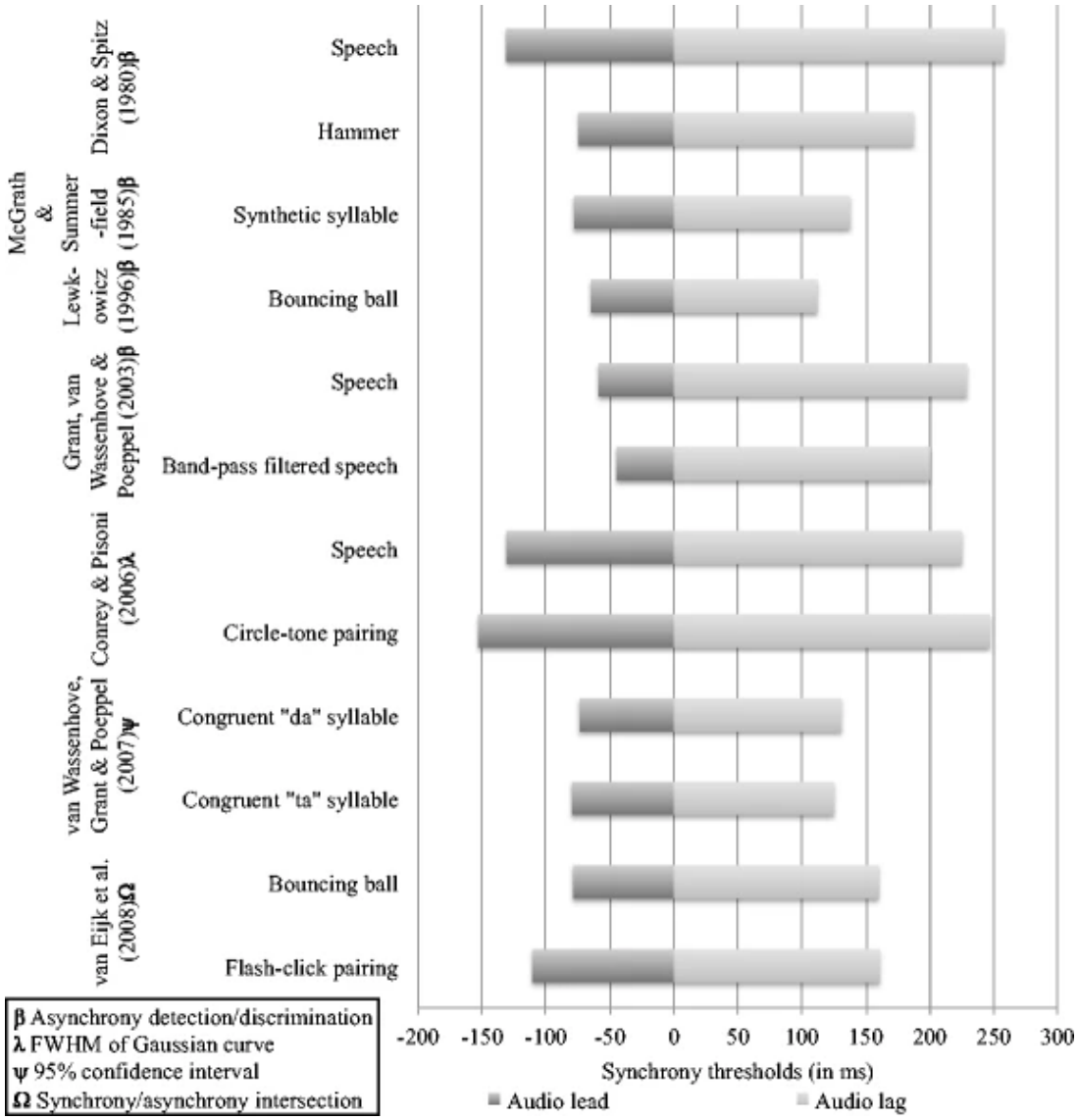


Figure 2.1: Overview of SJ thresholds taken from Eg et al. (2015)

2.3 The echo effect

2.3.1 Delayed Auditory Feedback (DAF)

Delayed auditory feedback classically occurs when a speaker hears her voice in a slightly delayed manner, which has been shown to induce stress, see 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 Smart hearing protection

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, [11] 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 in-depth discussion of this trade-off see Lezzoum et al. (2016)

2.3.3 On the temporal scale of audiovisual asynchrony

Regarding 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?

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

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

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 traveling faster than sound.~~ (van Wassenhove et al., 2007; Maier et al., 2011) Here, the optimally performing temporal window for integration is estimated to be around 200ms, from -30ms to 170ms. At larger delays, the integrative capacity is still present but gradually declines.

To minimize the effect of audiovisual integration in our results, we want to remove compensation effects from a possibly strongly interfering TWIN. Therefore, our choice of value for the large asynchrony condition should 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 found by Maier et al. (2011). A more recent audiovisual delay study Li et al. (2021) noted that in a standard audiovisual simultaneity judgment (SJ) Task with stepped delays from -400 to 400ms delay, roughly 50 percent of the participants incorrectly judged the 200ms delayed stimulus to be synchronous. Even in the 400ms condition, around 10 percent 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 auditory and visual parts 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 less causally related conditions, they received less

synchronous" responses, suggesting that also in our study, when interference (attenuated echo) is present, we would expect a more accurate performance of the participants.

Similarly, with a 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. This corresponds roughly to the previously defined "optimal" TWIN performance. For larger audiovisual delays they measured up to 267ms subjective delay with the visual stimulus leading, where still less than 80% of the participants were able to correctly 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 among different modalities below 20 msec are usually unnoticed, they put forward that this is due to a strong natural tendency to reduce errors and adaptive temporal recalibration. 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 auditory 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 differences 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 voices. Agnew and Thornton (2000) using delayed auditory feedback (DAF) report people noticing a delay as small as 3-5ms, Stone and Moore (2002) report the smallest noticeable DAF rather be around 15ms under optimal conditions. Studies looking at DAF cannot be applied at face value here, the detection threshold for own voice recordings consistently seems a lot lower than for external voices. Both teams demonstrate findings that auditory lag with DAF applied is already clearly annoying and speech production performance decreasing to the speaker at around at 20-30ms. 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) and unattenuated external

ce and 20db attenuated external voice. The tolerance for external voices is much more interesting for us 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 auditory 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 with the same attenuating function that we are testing and found that the smallest speech-related echo was detected by at least 20 percent of the participants at 16 - 22ms delay. In their setup participants were able to tune the temporal asynchrony between auditory and visual stimulus between 0 and 1000ms. Testing two different types of fit of the SHPD, a shallow and a deep one, participants were listening to a french sentence approximately 2000ms long. Testing the uncorrupted speech signal versus modified noisy versions of the same sentences, they report that participants have different asynchrony detection thresholds depending on the quality of speech and fit of the device. They found that the size of the echo threshold depends on the presence of background noise: with noise the threshold increases. With clean speech stimuli the median echo threshold was 38 ms, while when speech is corrupted by noise, the median echo threshold was found to be at 96 ms. Compatible with other findings, they also state that the echo threshold scales with the duration of the signal: for a short 8ms non-speech bell signal the threshold is much smaller. The team also stresses that detection thresholds depend on the attenuation function: The higher the attenuation is, the higher the AV delay perceivable. Testing different types of background noise (babble speech vs. factory noise) yielded the conclusion that stable background noises impact the threshold less than dynamic noise like speech.

For simple non-speech stimuli, the asynchrony detection threshold is smaller, Lezzoum

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 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. They also report that there is a clear tendency for NH-Participants to be less tolerant towards temporal delay than HI-Participants. Even more, 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.

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. Taking the literature on detectable thresholds into account, using complex speech stimuli, we should be reasonably sure that a vast majority of our participants will be able to detect neither the AV asynchrony nor the simulated attenuated echo. Should we still find any speech performance impact in these conditions, this should be a good indication for ~~strong~~ multimodal subconscious mechanisms involved in speech perception, ultimately preventing the use of any higher-order filter in SHPD, at least in environments where speech understanding is critical.

2.3.4 Different values in the literature

Table 2.1: Collected findings on speech related AV asynchrony

Reference	What was Measured	Measurement	Range	Participants	Setup
Lezzoum et al. (2016)	Attenuated AV 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 in clean speech condition, 20ms noisy, 30ms with DAF, rate affected	7, 16, 27, 43ms DAF lag	32 NH	clean vs noisy environment
Maier et al. (2011)	TWIN synchronicity, DAF, (own/other voice)	mainly false synchronous responses between 0ms and 200ms AV lag	-333ms - 333ms	9 NH	SJ and TOJ task
van Wassenhove et al. (2007)	AV TWIN via indirect McGurk Fusion	Optimal TWIN spans -30 ms to +170 ms	-467 ms - 467 ms	43 NH	SJ Task with McGurk combination in components
Agnew and Thornton (2000)	DAF	3-5ms noticeable, 30ms objectionable	noticeable effects ranged 2.15ms - 7.04ms	18 NH	delay slider manipulation with a DSP hearing aid
Li et al. (2021)	audio-visual onset asynchrony (AVOA)	50 % judged the 200ms delay as synchronous,	Five SOAs (-400, -200, 0, 200, and 400ms)	27 NH	SJ Task low/high causality via blurring on Speakers
Vroomen and Keetels (2010)	Meta-study, AV temporal asynchrony	¡ 20ms usually unnoticed	–	–	SJ and TOJ

Table 2.2: Other, non-speech related findings

Reference	What was Measured	Measurement	Range	Participants	Setup
Petrini et al. (2009)	TWIN for AV Drumming	112 ms (TWIN), 80ms highest SJ	-266ms - 266ms	34 NH, 17 were expert	SJ Task on Speakers
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 with open-canal hearing aids

2.4 Age Effects

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 towards neurotypicals

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. One rather well-established processing difference lies in recalibration speed, or maybe even the overall capacity for re-calibration. As well explained in Turi et al. (2016), TD individuals exhibit rapid re-calibration, often shown via SJ tasks. The skew of the temporal asynchrony of the preceding trials partially determines the judgment in the current trial. 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 optimize adverse speech perception situations. This would also explain why individuals with ASD typically start to speak later and under-perform in language reproduction. In Brandwein et al. (2013) this is discussed and extended to more general, basic nonspeech and nonsocial stimuli, suggesting this to be a rather consistent effect even present in relatively early stages of information processing. The team puts forward that there is a general deficit in AV integration present in ASD and that it is likely responsible for the communicative deficits

exhibited in ASD. More information on a comparison with still-developing children can be found in Noel et al. (2017), who compared the ability to rapidly recalibrate in TD and ASD participants aged 7-17. They demonstrated a significant difference in performance in an SJ task, but not in all stimulus categories. While the ASD participants were found to recalibrate on a trial-by-trial basis similar to the TD participants for speech stimuli, they presented a significant underperformance in nonlinguistic stimuli. This is the opposite of general findings for adults and suggests that speech integration processes drastically change with age and throughout development. For a concise overview see Stevenson et al. (2014), who state that atypical sensory binding [12] is likely the underlying cause for many traits typically associated with ASD, like impairments in social and communicative skills.

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 often 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.

3 Experiment

To reiterate, the goal of the experiment is to discover the effect of small audiovisual asynchronies on speech perception. We are interested specifically in sub-perceptible delays and the specific echo that occurs when the unattenuated auditory signal is partly blocked by the SHPD and the wearer perceives both the attenuated original signal and the delayed output signal of the SHPD.

[12] binding refers here to the conceptual mapping and integration of modal sensory input

TODO look at the transition here

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 is needed to sufficiently process the linguistic signal, whether it is primarily auditory, visual, or both to respond to the task. We also record the accuracy to be able to detect any secondary effects, as lower accuracy could index a deterioration in language processing and comprehension.

Due to the large subjective differences in audiovisual sensory intake and the direct manner of reporting requiring conscious perception, it was hard for prior studies to come up with a concrete number for the audiovisual delay that can be utilized with engineering in mind. This is 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 the 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 the resulting visual and auditory artifacts, are consistent.

3.1 Method

3.1.1 Participants

The experiment recruited

TODO

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

TODO,

with a mean age of

TODO.

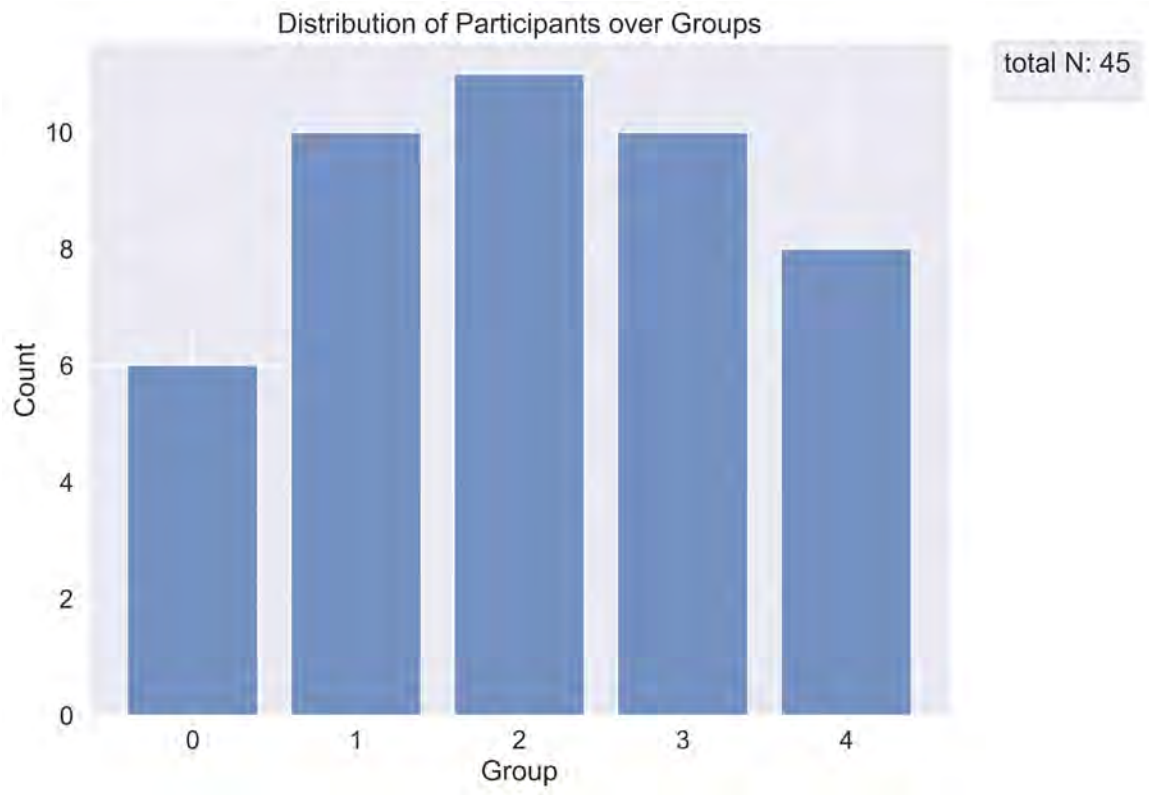


Figure 3.1: Distribution of participants over in-between groups

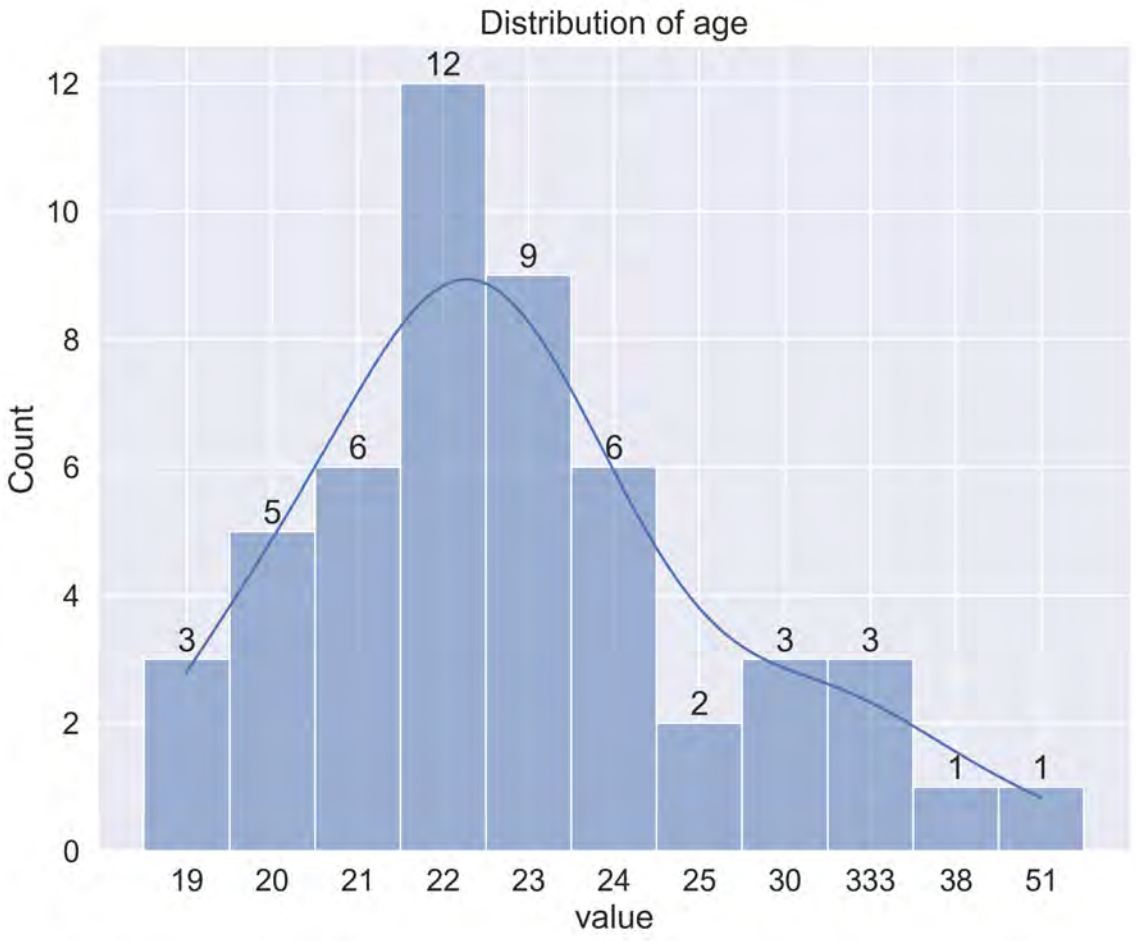


Figure 3.2: Age distribution of participants

Participation was completely voluntary and written informed consent was obtained from all participants. They could abort the experiment 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 partial course credit (VP-Stunden) as compensation. No other compensation was granted. We asked our participants to wear wired headphones in an attempt to minimize distracting environmental noises beyond our control. Furthermore, we informed people not to use wireless headphones to avoid additional delays.

3.1.2 Materials

Media files Video and corresponding audio files are used with friendly permission by the original creators of the OLAKS Corpus Rosemann and Thiel (2018). These are full HD recordings of a male German native speaker uttering German sentences centered on

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 for the main task, half of which follow an SVO structure, the other half OVS, where both the subject and the object are modified, respectively. 70 sentences were used as main trials, 10 more were used as filler trials. Two example sentences for SVO and OVS structures taken from the corpus are

(1) *Den alten Pfarrer grüßt der kluge Pilot.*
The-ACC old priest greets the-NOM clever pilot.
'The clever pilot greets the old priest.'

(2) *Der stille Postbote grüßt den dicken Frisör.*
The-NOM silent mailman greets the-ACC fat pilot.
'The silent mailman greets the fat pilot.'

Each sentence contains two nouns, each modified by an adjective and a verb connecting the two. The entities referred to by the nouns are either animals, professions, or mythical creatures, typically appearing in tales targeted at children. They are selected to be readily identifiable, with a clear prototypical image coming to mind. The full list of utilized sentences is available in the appendix.

Audio and video streams 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

To generate the delay conditions, the rounded number of sound samples is added in front of the audio signal. The rounding error is 1 sample, or 1/48000th of a second. The resulting, longer audio sample is then merged with the video stream, where a still frame is added for the last few ms where a sound signal is playing, but the video has played through.

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 x 80 = 400 stimuli. The audio and video streams are combined into a single file before the experiment to minimize av-synchrony issues resulting from different media playback handling in different browsers.

Images 53 of the 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. 14 images for sentences which did not have a direct fit in the MultiPic Database were found via Google Image search and are all licensed free for personal use, totalling in 67 images used in the experiment. All of these are then manipulated using GIMP 2.10.22, centered on a quadratic canvas with a transparent background, all resolutions ranging from 500 to 1200 pixels. The full list of the images used can be found in the appendix.

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 their consent. After consent, pseudonymous data were collected, such as age, gender, vision, and hearing capacity. We requested that participants eliminate any possible interfering distractions such as noise or other people in the same room. 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. They are also instructed to

ensure adequate viewing conditions 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. 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 background is white throughout the entire experiment. After the chance to correct missing requirements, such as getting missing glasses or correctly setting up and enabling headphones, participants are presented with an example stimulus that could be repeated at will to adjust the sound level to a comfortable level comparable to a face-to-face conversation. After indicating that they adjusted their volume accordingly, participants are then redirected to another browser window playing in fullscreen, which contains the entire experiment. After a brief instruction to the task, they were presented with 5 trial runs to get familiar with the nature of the main task. No feedback on correct answers was given. Participants are then reminded to answer as fast as they possibly can, using only the middle and the index finger of their dominant hand, to reduce possible differences between the dominant and nondominant hand reflecting in the RTs.

The experiment consists of two different tasks structured in blocks: the main task, an unspeeded forced choice referent identification task (choice task), and a modified SJ task. Between each main trial block, breaks are inserted and not time-restricted, the participant could choose for how long to take each break.

Target Presentation The presented target words are extracted from the sentences in the corpus. Every adjective was presented once as a target word, every sentence was used two times, presenting a different target. The adjectives were displayed for 2500ms in the center of the screen sized at 10 percent of the screen height.

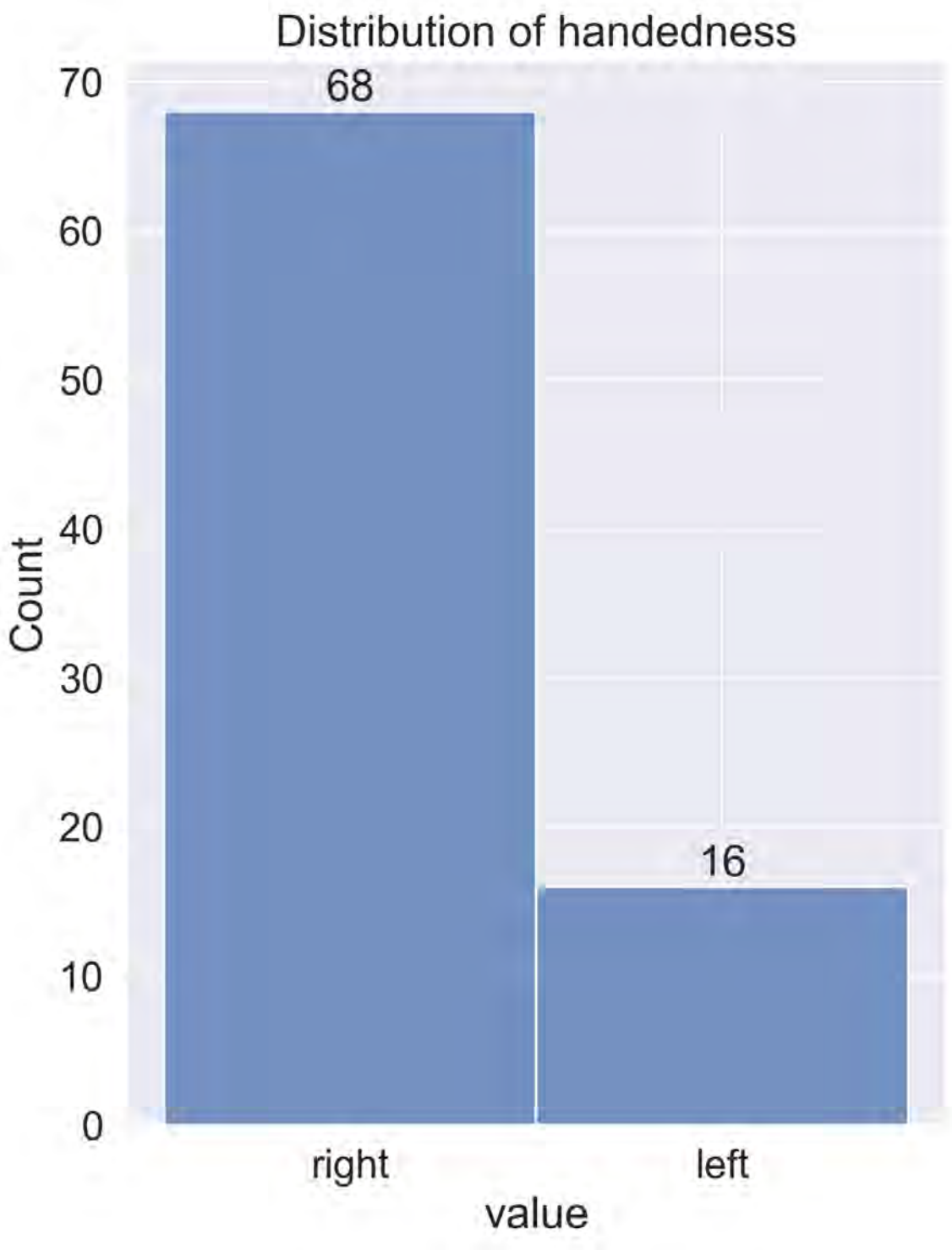


Figure 3.3: Handedness of participants

(3.4.1) Example of image files for a OVS sentence: Pfarrer, Pilot

(3.4.2) Example of image files for a SVO sentence: Postbote, Frisör

Figure 3.4: Example images, corresponding to 1 and 2

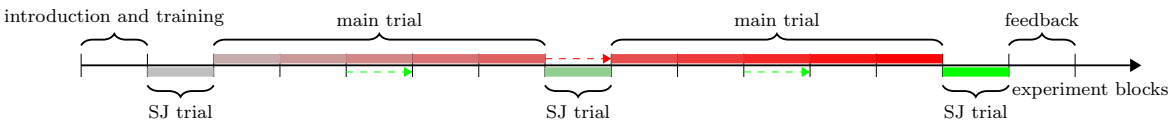


Figure 3.5: Temporal order of entire experiment

Figure 3.6: Example of target presentation

Main Task TODO better word for the main task

The task was a referent identification task, in which the participant has to choose which noun is modified by the target adjective. The main task consists of 10 blocks with 16 trials each, for a total of 160 unique trials. Each trial is divided into target presentation and stimulus presentation and records the response starting with the onset of the video stimulus.

Target presentation The target, which was either one of the two adjectives present in the sentence, is flashed for 2500ms.

Stimulus Presentation Then, in the stimulus phase, the video clip stimulus is shown alongside with two images corresponding to the left or right answer option indicated by the position of the images and helping arrows. To let the participant have a look at the images and ensure proper identification, the images are first shown alone. Then, after 1500ms, a fixation cross is presented at the center of where the audiovisual stimulus will appear. After 1000ms of presenting the images and the fixation cross, the movie clip is presented and keypresses are recorded. An upward arrow is presented alongside to remind the participant to press the upper arrow when no image fits.

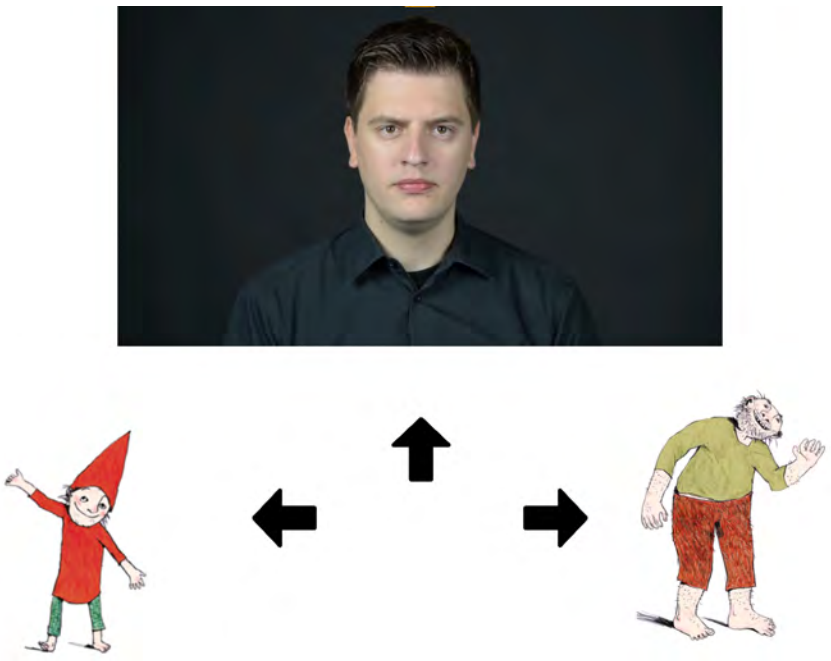


Figure 3.7: The stimulus presentation in the main task

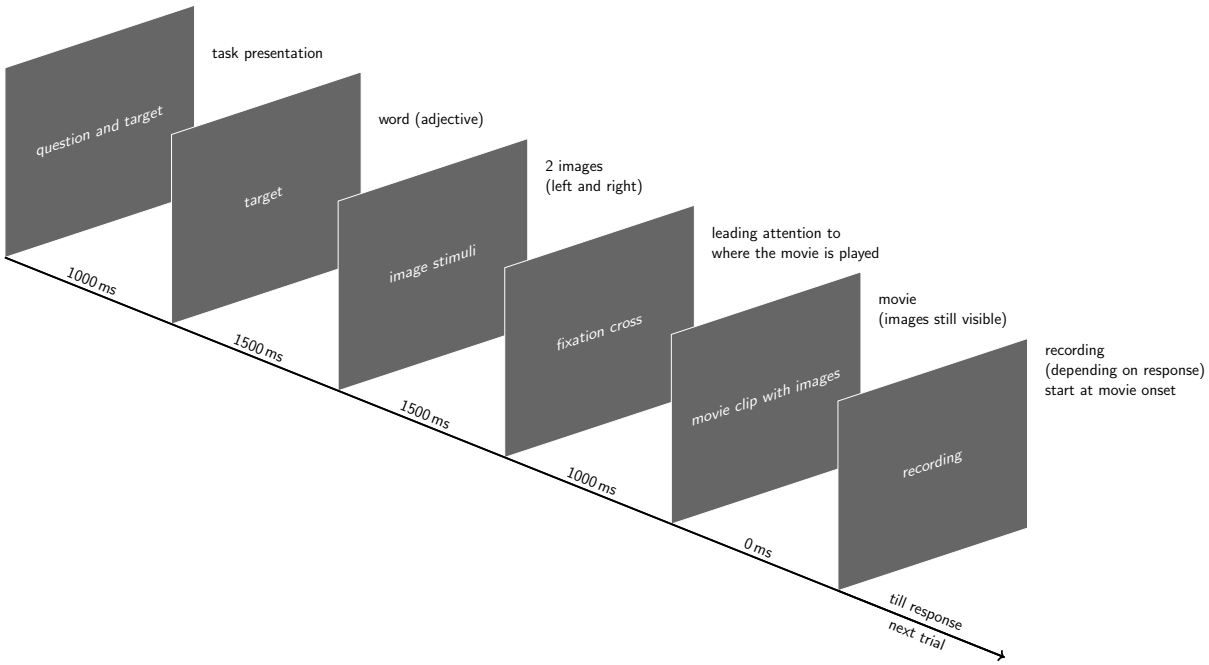


Figure 3.8: Temporal order of presentation in main task

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. Valid keypresses are either the left arrow key or the right one, mapped to the corresponding location of the image stimuli. The task is organized in a pseudorandom order counterbalanced in 5 groups between participants. The randomization applied ensures that trials in immediate sequence never have identical target words, experimental conditions, images, and sentences. Each stimulus sentence is used twice: once with the target being the first noun and the modifying adjective, once with the target being the last pair. To prevent inferential problem solving, we introduced filler trials where the target is misleading and not modifying any of the referents. When a filler trial was presented, a randomly chosen adjective from the trial corpus was displayed. The adjective consequently did not appear in the video clip, resulting in the correct answer to be the upper arrow key. Out of the 160 trials 20 were filler trials, resulting in 12,5 percent of the trials to be fillers. In the time between each block, the trial progress was presented and participants could determine on their own for how long to take a break and could continue with a keypress.

SJ task The participant, after watching one video per trial, is then asked whether the auditory signal was perceived to be synchronous with the visual stimulus. Next, also whether any auditory distortion, such multiple overlapping audio signals, were perceived. These questions are asked for the same stimulus in sequence after it has finished playing. For each question, accuracy 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. The modified SJ task is performed 3 times with 10 randomly ordered trials, and each of the 5 conditions was repeated twice in each block with a different sentence. The blocks are distributed before, after half of the main trials completed, and after completion of the main task. Each block consists of the same 10 sentences, reshuffled, also taken from the OLAKS set, but not presented in the main task.

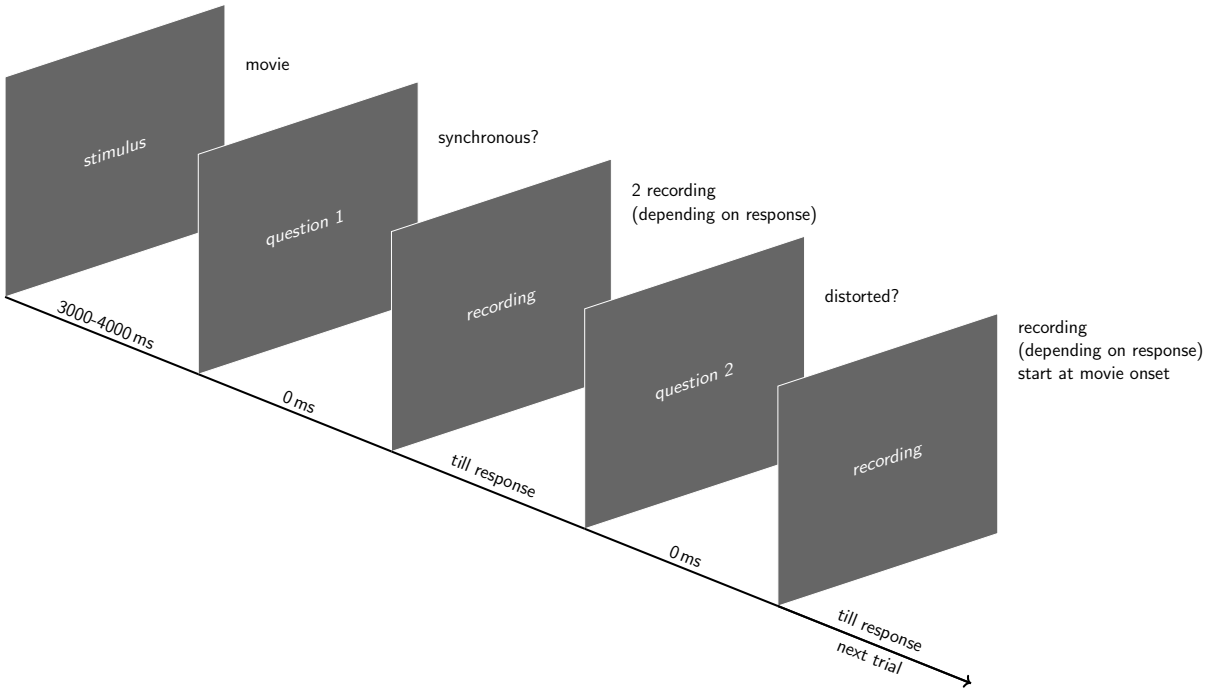


Figure 3.9: Temporal order of presentation in adapted SJ task

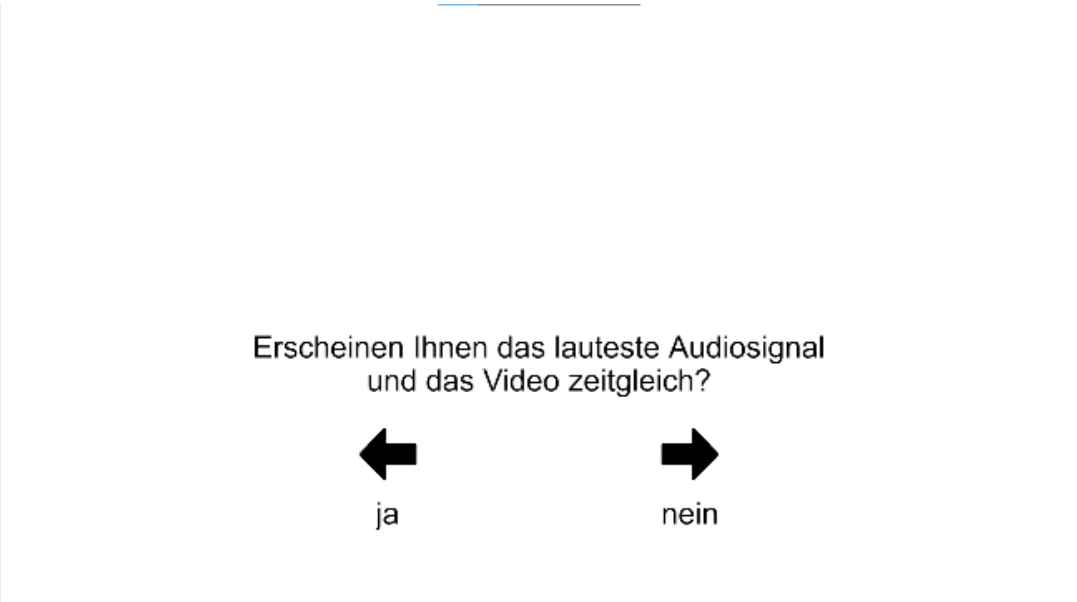


Figure 3.10: Presentation of questions with response indicators

3.1.4 Hypothesis

Table 3.1: Conditions

Modifier	Control Condition	Condition 1	Condition 2	Condition 3	Condition 4
AV delay	0ms	10ms	400ms	-	-
Attenuated echo	0ms	-	-	10ms	400ms

Table 3.2: 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

Table 3.3: Independent Variables (IV)

Symbol	Variable	Values
lat	latency of audio to visual stimulus	0ms latency (no latency)
		10ms latency,
		400ms latency,
ech	attenuated inverse echo	0ms echo (no echo),
		10ms echo,
		400ms echo

TODO restructure table, group by condition

As the primary effect of interest, we expect that when presented with greater temporal dis-alignment of sensory inputs, 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 increasingly adverse of the speech stimuli. Concretely, RTs will be the shortest in the control condition, 0ms latency, 0ms echo, and the RTs will be higher for latency and echo conditions,

respectively. With more adverse stimuli for processing by either temporal misalignment 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 manipulated adverse conditions in comparison to the unmodified base condition.

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 a large proportion of incorrect responses in the SJ task. The intent of the large modification (400ms delay, 400ms echo) is to verify that linguistic processing is indeed positively dependent on synchrony and noninterference. Therefore, we expect largely correct identification in the secondary task.

3.2 Results

3.2.1 Choice Task

3.2.2 SJ Task

3.2.3 Statistical Analysis

RStudio Team (2020)

Descriptive Statistics

4 Discussion

4.1 My Results in other current research

4.2 Later studies with ASD

4.3 Conclusion

Does the hypothesis hold?

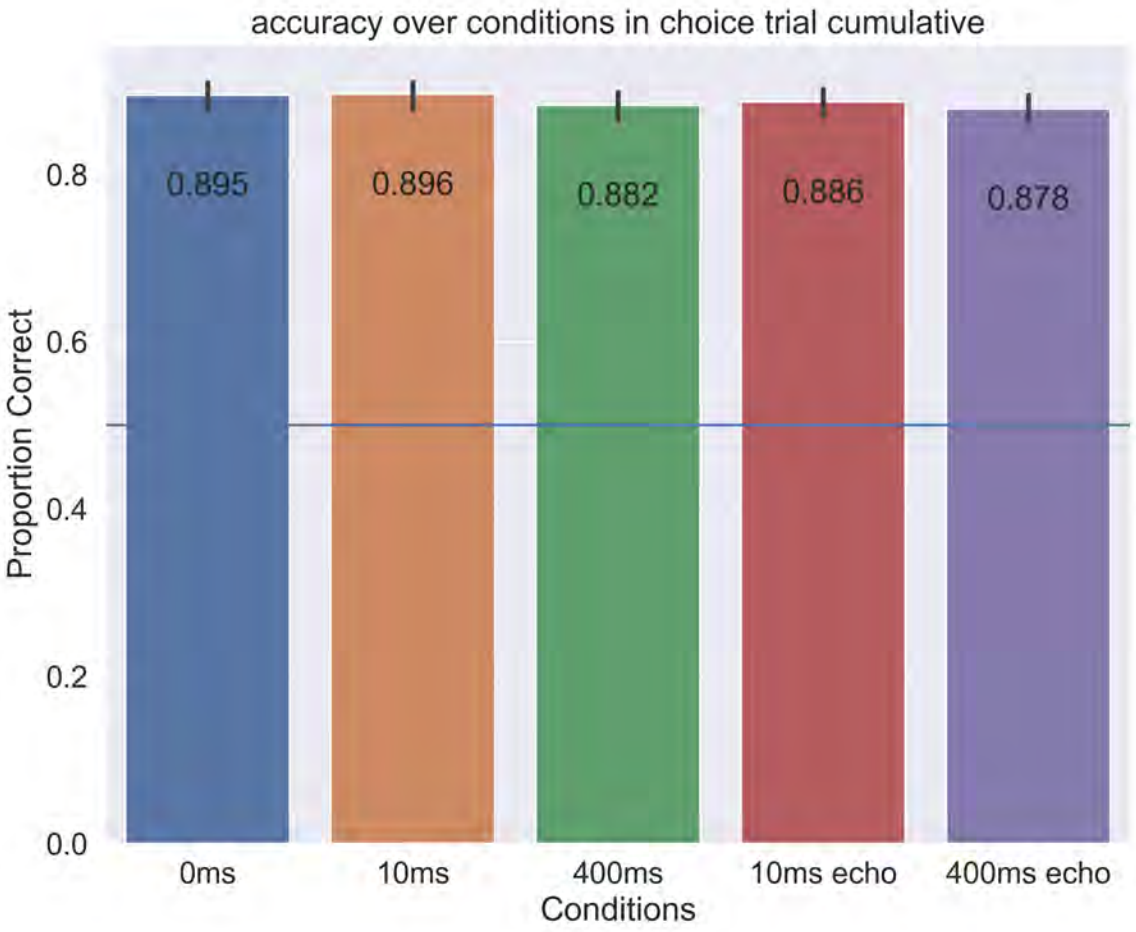


Figure 3.11: Example result table

4.4 Implication

4.5 Suggestions for further research

Issues and open questions

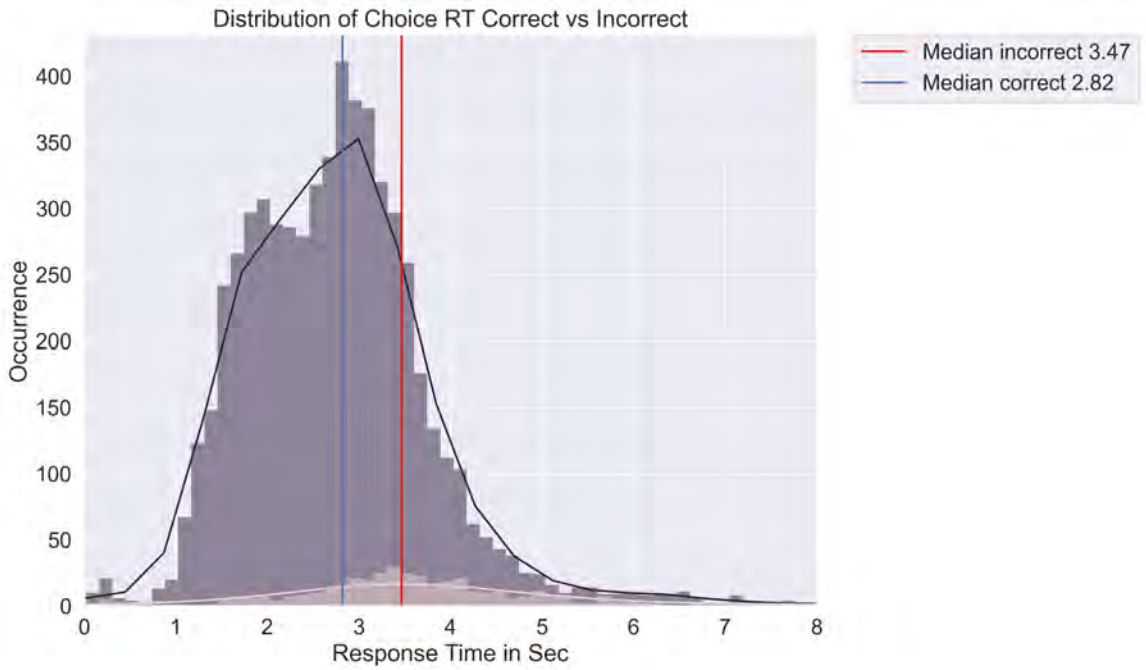


Figure 3.12: Example result table

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.

Eg, R., Griwodz, C., Halvorsen, P., and Behne, D. (2015). Audiovisual robustness: exploring perceptual tolerance to asynchrony and quality distortion. *Multimedia Tools and Applications*, 74(2):345–365.

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.

Grant, K. W., van Wassenhove, V., and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication*, 44(1):43–53. Special Issue on Audio Visual speech processing.

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.

Kavanagh, J. F., Mattingly, I. G., et al. (1972). *Language by ear and by eye: The relationships between speech and reading*, volume 50. Mit Press Cambridge, MA.

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 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.

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.

Samelli, A. G., Gomes, R. F., Chammas, T. V., Silva, B. G., Moreira, R. R., and Fiorini, A. C. (2018). The study of attenuation levels and the comfort of earplugs. *Noise & Health*, 20(94):112–119.

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.

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.

Van der Burg, E., Alais, D., and Cass, J. (2018). Rapid recalibration to audiovisual asynchrony follows the physical—not the perceived—temporal order. *Attention, Perception, & Psychophysics*, 80(8):2060–2068.

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.

Vatakis, A. and Spence, C. (2006). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters*, 393(1):40–44.

Vroomen, J. and Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics*, 72:871–84.

Wang, M., Kong, L., Zhang, C., Wu, X., and Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions. *The Journal of the Acoustical Society of America*, 143(4):EL255–EL259.

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.

Zampini, M., Shore, D. I., and Spence, C. (2003). Multisensory temporal order judgments: the role of hemispheric redundancy. *International Journal of Psychophysiology*, 50(1):165–180. Current findings in multisensory research.

A Appendix

A.1 Stimuli

A.1.1 Images

Figure A.1: All image stimuli and their sources

(A.1.1) Arzt (doctor), Source: Duñabeitia et al. (2018)

(A.1.2) Bauer (farmer), Source: https://de.cleanpng.com/png-m0f5mp/download-png.html

(A.1.3) Boxer (boxer), Source: Duñabeitia et al. (2018)

(A.1.4) Bräutigam (spouse), Source: https://de.pngtree.com/so/braut-clipart

(A.1.5) Bäcker (baker), Source: https://www.pngwing.com/de/free-png-zboip/download

(A.1.6) Bär (bear), Source: Duñabeitia et al. (2018)

(A.1.7) Büffel (buffalo), Source: https://de.cleanpng.com/png-eosws1/

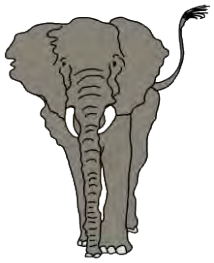
(A.1.8) Clown (clown), Source: Duñabeitia et al. (2018)

(A.1.9) Cowboy (cowboy), Source: Duñabeitia et al. (2018)

(A.1.10) Dieb (thief), Source: Duñabeitia et al. (2018)

(A.1.11) Drache (dragon), Source: Duñabeitia et al. (2018)

(A.1.12) Elch (elk), Source: https://www.pngaaa.com/detail/2563273



(A.1.13) Elefant (elephant) Source: Duñabeitia et al. (2018)

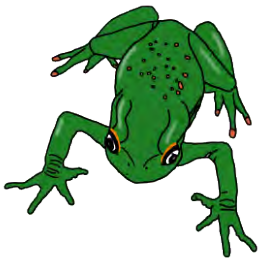
(A.1.14) Engel (angel), Source: Duñabeitia et al. (2018)

(A.1.15) Ente (duck), Source: Duñabeitia et al. (2018)

(A.1.16) Esel (donkey), Source: Duñabeitia et al. (2018)

(A.1.17) Fee (fairy), Source: Duñabeitia et al. (2018)

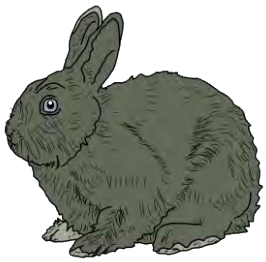
(A.1.18) Fee (fairy), Source: Duñabeitia et al. (2018)



(A.1.19) Frosch (frog), Source: Duñabeitia et al. (2018)

(A.1.20) Gespenst (ghost), Source: Duñabeitia et al. (2018)

(A.1.21) Gärtner (gardener), Source: Duñabeitia et al. (2018)



(A.1.22) Hase (hare), Source: Duñabeitia et al. (2018)

(A.1.23) Hexe (witch), Source: Duñabeitia et al. (2018)

(A.1.24) Hund (dog), Source: Duñabeitia et al. (2018)

(A.1.25) Junge (boy), Source: Duñabeitia et al. (2018)

(A.1.26) Jäger (hunter), Source: Duñabeitia et al. (2018)

(A.1.27) Kapitän (captain), Source: Duñabeitia et al. (2018)

(A.1.28) Kasper (punch), Source: Duñabeitia et al. (2018)

(A.1.29) Kellner (Waiter), Source: Duñabeitia et al. (2018)

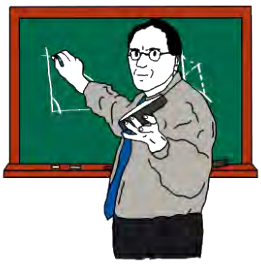
(A.1.30) Koala (koala), Source: Duñabeitia et al. (2018)

(A.1.31) Kobold (cobold), Source: Duñabeitia et al. (2018)
`https://de.cleanpng.com/png-85z9z8/download-png.html`

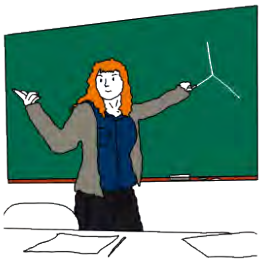
(A.1.32) Koch (Chef), Source: Duñabeitia et al. (2018)
`https://pngtree.com/so/chef-hat-clipart`

(A.1.33) König (King), Source: Duñabeitia et al. (2018)

(A.1.34) Königin (Queen), Source: Duñabeitia et al. (2018)



(A.1.35) Lehrer (Teacher), Source: Duñabeitia et al. (2018)



(A.1.36) Lehrerin (Teacher), Source: Duñabeitia et al. (2018)

(A.1.37) Löwe (Lion), Source: Duñabeitia et al. (2018)

(A.1.38) Maler (Painter), Source: Duñabeitia et al. (2018)

(A.1.39) Mann (man), Source: Duñabeitia et al. (2018)

(A.1.40) Matrose (sailor), Source: Duñabeitia et al. (2018) https://www.nicepng.com/ourpic/u2w7q8a9q8w7w7u2_navy-drawing-sailor-line-

(A.1.41) Maulwurf (mole), Source: Duñabeitia et al. (2018) https://www.pngwing.com/de/search?q=maulwurf+zeichnung

(A.1.42) Metzger (butcher), Source: Duñabeitia et al. (2018)

(A.1.43) Mönch (monk), Source: https://de.cleanpng.com/png-zsofj1/download-png.html

(A.1.44) Nikolaus (Nikolaus), Source: https://nikolaus-von-myra.de/de/darstellung/galerie/

(A.1.45) Panda (panda), Source: Duñabeitia et al. (2018)

(A.1.46) Papagei (parrot), Source: Duñabeitia et al. (2018)

(A.1.47) Papst (pope), Source: Duñabeitia et al. (2018)

(A.1.48) Pfarrer (pastor), Source: Duñabeitia et al. (2018)

(A.1.49) Pilot (pilot), Source: Duñabeitia et al. (2018)

(A.1.50) Pinguin (Penguin), Source: Duñabeitia et al. (2018)

(A.1.51) Jäger (hunter), Source: Duñabeitia et al. (2018)

(A.1.52) Polizist (Police-man), Source: Duñabeitia et al. (2018)

(A.1.53) Postbote (mail-man), Source: Duñabeitia et al. (2018)

(A.1.54) Prinz (prince), Source:https://www.pngegg.com/de/png-iptbx

(A.1.55) Punker (punk), Source: https://www.vippng.com/preview/hRbTRJR_punk-tribu-urbana-vestimenta-peinados/

(A.1.56) Radfahrer (biker), Source: https://de.pngtree.com/so/die-jungs

(A.1.57) Riese (Giant), Source: https://www.pinterest.de/pin/766386061570400292

(A.1.58) Ritter (Knight), Source: Duñabeitia et al. (2018)

(A.1.59) Roboter (robot), Source: Duñabeitia et al. (2018)

(A.1.60) Räuber (bandit), Source: Duñabeitia et al. (2018)

(A.1.61) Soldat (soldier), Source: Duñabeitia et al. (2018)

(A.1.62) Tiger (tiger), Source: Duñabeitia et al. (2018)

(A.1.63) Tourist (tourist), Source: Duñabeitia et al. (2018)

(A.1.64) Vater (father), Source: `https://de.cleanpng.com/png-hu6n8o/download-png.html`

(A.1.65) Wikinger (viking), Source:`https://de.cleanpng.com/png-en6r2o/`

(A.1.66) Zauberer (wizard), Source:Duñabeitia et al. (2018)

(A.1.67) Zwerg (dwarf), Source: Duñabeitia et al. (2018)

A.1.2 Sentences

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

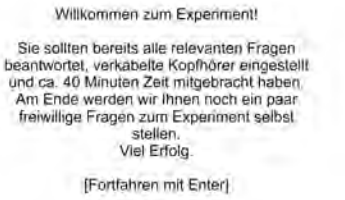
1. Der schlaue Kasper beschattet den faulen Vater.
2. Der blinde Jäger erschießt den braven Soldaten.
3. Der fiese Pirat erschießt den braven Soldaten.
4. Der faule Bäcker ersticht den bösen Koch.
5. Der fiese Koch ersticht den armen Touristen.
6. Der böse Gärtner erwürgt den dreisten Postboten.
7. Der 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.

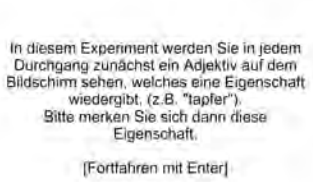
A.2 Experiment screens

Here you will find all screens shown to the participants in chronological order.

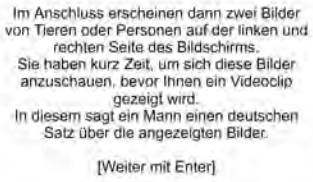
Figure A.2: Screens presented in online experiment



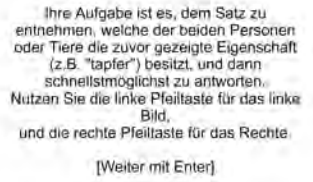
(A.2.1) Welcome Screen



(A.2.2) Introduction

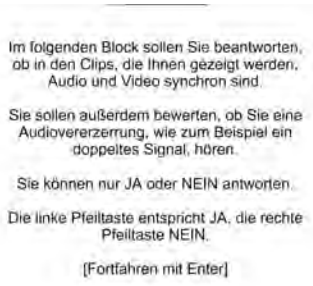


(A.2.3) Introduction

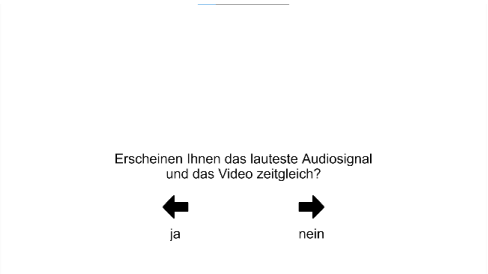


(A.2.4) Introduction

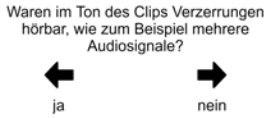
(A.2.5) Sound Adjustment Screen



(A.2.6) Introduction SJ Task



(A.2.7) Synchrony Question



(A.2.8) Distortion Question

(A.2.9) Target Presentation

(A.2.10) Stimulus Presentation

A.3 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.4 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, May 12, 2021

Aron Petau

city, date	signature