Audio-Visual Speech Processing and effects of multisensory asynchronicity

Aron Petau

aron@petau.net

967985

Bachelor's Thesis

at the Institute of Cognitive Science,

Osnabrück University

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 real world, multiple stimulus modalities are present at all times and asynchronies and inconsistencies are frequent. The brain integrates and synchronizes these modalities to create the world as we know it. I review the literature on multimodal integration and present the current scientific status. In the present study, I seek to identify possible problems related to learning and speech processing in general when presented with temporal audiovisual delays in stimuli. I also examine application-specific properties such as the temporal delay between passively and actively transmitted auditory signals in smart hearing protection device (SHPD). I present the design, methodology, and results of an online psycholinguistic study conducted with German-speaking students. Participants are presented with a set of uttered German sentences with the speaker and his lips visible on the screen. Participants perform an identification task where they have to choose which noun was modified by a target adjective. The audiovisual delay is modified and a simulated passively transmitted attenuated audio signal is introduced, while reaction times and response accuracy are measured. I discuss possible limitations and use cases with a focus on individuals with autism spectrum disorder that could benefit from increased specificity in filtering noise with a tradeoff for increased audiovisual latency. I aim to establish a relationship between audiovisual delays and speech recognition capability while trying to identify a balanced delay making complex filtering possible from an engineering perspective while ensuring that the additional harm to speech processing is minimal.

Keywords: multisensory integration, smart hearing protection devices, audiovisual asynchrony, auditory sensitivity in ASD

Contents

List of Figures

List of Tables

Acknowledgements

I'm extremely grateful to M.Sc. Juliane Schwab (Institute of Cognitive Science, University of Osnabrück), for being an excellent and involved mentor and supervisor, giving regular constructive advice and helping me navigate through the ocean of statistical analysis.

I would also like to extend my deepest gratitude to 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, literally lending her ears for this project and filling large gaps in my knowledge of audiology, as well as providing large parts of the automated echo simulation code.

Also the whole research team at the NSERC-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS) for providing very useful code simulating the echo effect and the possibility to record samples on the ARP 3.1.

I would like to thank Rosemann and Thiel (2018) for providing the video stimuli used in this study.

1 Introduction

Perceiving the world is a cooperative effort of the senses. Our senses are constantly providing large amounts of data that together form our perception of the world. Speech perception, like most sensory information processing, is inherently multisensory. In a face-to-face conversation, the listener will be exposed to auditory information of the uttered speech as well as some form of visual information, such as bodily cues, gestures, facial expressions, and articulatory movements. Speech information stemming from visual and auditory modalities complement each other (Kavanagh et al., 1972) and it follows that speech perception cannot be fully understood by exclusively looking at one isolated sensory modality[1]. The perceptual incorporation of multisensory streams of information into a single percept[2] is researched under the term multisensory integration. There are significant individual differences in perceptual integration and even prior to this, perception is shaped by external factors, such as the properties of the stimuli themselves. Some of the individual differences are structured; atypical perceptual processing is often observed in individuals with Autism Spectrum Disorder (ASD) as defined in American Psychiatric Association (2013) and there is compelling evidence that their multisensory integration and by extension their audiovisual integration work differently compared to neurotypical individuals[3] (NT) (Noel et al., 2017; Turi et al., 2016).

Auditory sensitivities Another common aspect in individuals with ASD, relevant for the experiment at hand, is increased auditory sensitivity and discomfort, which often leads to avoidant behavior displayed in various forms. Common symptoms of ASD as stated by the American Psychiatric Association (2013) include an "adverse response to specific sounds", which can vary in intensity and scope for individuals, but being reported by parents in more than 40% of records of individuals with ASD (Stiegler and Davis, 2010). This selective hyperreactivity to sound is often paired with socially stigmatized

[1] 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.

[2] specifying the content of a perception; information obtained though our senses

[3] This describes neurologically typical individuals, used as a term to describe non-autistic individuals, in the stricter definition individuals without any learning- or neurodevelopmental disorder.

behaviour and does not seem to be explained by physiological perceptual differences when compared to perception in neurotypical (NT) (Stiegler and Davis, 2010). The consequences are inhibitory to societal life for many individuals and require external help for sound protection, such as earmuffs or earplugs, alleviating the experienced discomfort as reported by Neave-DiToro et al. (2021) and Ikuta et al. (2016). Such passive hearing protection devices (HPDs) work for reducing the impact of sound in noisy environments, but if the signal is powerful enough, it will penetrate the HPD in attenuated form (Samelli et al., 2018).

Also, there are concerns about the prolonged use of passive HPDs to combat acoustic hyperreactivity as this may lead to an increase in long-term avoidant behaviour (Stiegler and Davis, 2010). Furthermore, passive HPDs lack *selective* sound attenuation, rendering them highly impractical for use cases requiring such. Ideally we would be able to attenuate unwanted noise exclusively and still adequately hear relevant auditory signals. Among other usages, individuals suffering from selective auditory hyperreactivity in ASD in particular would benefit from increased selectivity in attenuation.

Smart hearing protection devices One potential answer to increased sound sensitivity are smart hearing protection device (SHPD)s. These do not rely solely on passive attenuation as HPDs. While the attenuation works in a similar fashion to that of HPDs, an additional digital signal processor (DSP) with external microphones and in-ear speakers is able to transmit some selected signals. Advanced techniques can be applied in SHPDs, such as voice activity detection (Lezzoum et al., 2014), which is used to discriminate speech signals from noise. The result can then be used to transmit useful signals without attenuation while still being effective as a noise isolator and attenuating other signals. This selective attenuation promises to provide a viable solution for selective sound sensitivity issues in ASD. An SHPD could be programmed to attenuate individual distressing sounds and environmental noise while still transmitting speech (Lezzoum et al., 2014). Some of these advanced techniques, however, rely on algorithms that can introduce potentially significant transmission delays, resulting in persistent AV asynchronies (Lezzoum et al., 2016).

Audiovisual latency A possible limitation of SHPDs is that the selective attenuation applied might introduce adverse effects in speech processing through increased audiovisual latencies[4]. Traditional passive hearing protection device (earplugs, for example) function by presenting a physical barrier for sound, not introducing a relevant additional delay between the visual and the auditory signal[5]. Selectively attenuating an acoustic signal, however, comes with the trade-off that it requires active processing, which introduces additional latency to the perception of that signal by the wearer of an SHPD (Lezzoum et al., 2016). A higher audiovisual (AV) delay, when compared to passive HPDs, posits largely unknown risks that can potentially interfere with the speech processing of the wearer. The effects of large AV delays on speech processing are studied under the temporal window of integration (TWIN) paradigm (van Wassenhove et al., 2007; yu Zhou et al., 2020), but there is a lack of research, especially when looking at small, sub-perceptible delays. SHPD usage is typically taken to mean that the input is digitally processed in-situ, in "real-time". Real-time here is a flexible term that depends on the intended application and is defined as an acceptability threshold enabling a continuous output signal within time constraints (Kuo et al., 2013). For applications including audiovisual stimuli, such as online meetings, Younkin and Corriveau (2008) posit that this threshold should ideally stay below 45ms asynchrony (video leading) when visual information is present to prevent lip-syncing issues. Additionally, measuring perception thresholds is subject to individual perceptive and hearing differences (Hay-McCutcheon et al., 2009; Rosemann and Thiel, 2018) and other external factors like room acoustics (Haas, 1972). Results also depend on self-reported data, a problem stated in Eg et al. (2015).

The echo-effect Another effect specific to the usage of SHPDs is the multiple asynchronous presence of the auditory stimulus: an echo. Through the same limiting factors in noise isolation as passive HPDs, sufficiently loud auditory signals will be only partially

[4]A small audiovisual asynchrony is virtually always present due to light and sound signals travelling at different speeds and scales with the distance of sound source and target. This asynchronous arrival of stimuli is insignificant for typical distances a face-to-face conversation.

[5]For earplug-type passive HPDs, this is typically in the order of microseconds (μs) , and substantially smaller than the delay introduced by the actively transmitted signal, which is in the millisecond (ms) order (Lezzoum et al., 2016)

attenuated by the SHPD[6]. The initial auditory stimulus will still be heard, in an attenuated and not significantly delayed form. After a delay, the transmitted signal will be perceived, potentially overlapping with the original. When that happens, the wearer of an SHPD will hear a specific echo, meaning that, unlike a conventional echo, the original, but passively attenuated signal will be perceived first.

In such a situation, multiple relations can be observed: The asynchrony between the visual stimulus and the transmitted auditory signal, as well as the asynchrony between the two conflicting auditory stimuli, attenuated by the isolating capacity of the SHPD. As a complicating factor, the temporal delay is only one of many relational aspects. The difference in volume between two signals (Lezzoum et al., 2016) or the causal coherence of auditory and visual information are further examples of factors that potentially impact speech processing (Li et al., 2021).

Effects of SHPD and research focus In the specific case of wearing an SHPD to prevent exposure to potentially distressing sounds, an individual with ASD displaying hyperreactivity would benefit most from wearing the SHPD for a sustained amount of time, across all kinds of social activities, typically also requiring speech processing. Therefore, knowledge about the effects of sustained usage of an SHPD, potentially coming with several mentioned side-effects is necessary. A fragile balance has to be found between the temporal necessities of introducing delay in audio processing to increase quality and the potential speech and learning deficits a more delayed auditory signal could bring. The short- and long-term consequences of a small pervasive AV delay that is likely below the detection threshold in speech-related situations are unclear and motivate the present experiment.

To gain insights on audiovisual speech delays, we conducted an online experiment that aims to serve as a repeatable and extendable protocol to investigate the effects of audiovisual asynchrony as well as asynchrony detection capacity in participants using a referent identification task and a simultaneity judgment task, indirectly measuring the

[6]A typical noise reduction rating (NRR) for HPDs given by manufacturers is 25dB, but the real attenuation of signals is highly dependent on the individual fit and type of HPD, as well as acoustic properties of the signal and individual hearing capacities. Using the personal attenuation rating (PAR), the average attenuation is rather in the range of 16-20dB (Samelli et al., 2018).

efficacy of speech processing using reaction time (RT) and accuracy. Simulating the use of an SHPD for a non-hearing impaired NH neurotypical population, we investigate how much time is permissible to process the input signal before detrimental effects occur in speech processing and if they occur at all. The paradigm employed enables us to test for the effects of perceptible and subperceptible delays. To simulate effects specific to wearing an SHPD while engaging in speech processing, we introduce conditions where the original auditory signal is present in an attenuated form, prior to the actively transmitted signal, representing the echo introduced above. This condition is compared against conditions with a simple visual-leading delay and a condition with no modifications made.

In section 2, I proceed to present prior work on audiovisual delay perception and multisensory integration, specifically examining results regarding speech perception and changes under asynchrony. I present the experimental paradigms used to test the integration processes and delay perception and evaluate them. The experiment presented in section 3 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. After presenting our findings in subsection 3.2, I discuss them, examine possible shortcomings and pitfalls, and relate the paradigm and its results to other literature in section 4.

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 (Crosse et al., 2015; Sumby and Pollack, 1954), such that seeing the speaker's lips especially helps in making sense of what is being talked about (Calvert et al., 1997). Speech processing can be successful even with considerable temporal asynchrony between modalities and noise in the stimuli, while still extracting coherent speech information. (Ross et al., 2007) However, this leads to some interesting scientific questions regarding the multisensory integration of AV speech and temporal asynchrony detection. I will review the literature concerning these questions with a focus on audiovisual speech processing. For this, I am introducing 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 asynchrony between modalities and discuss more scenarios in speech perception, specifically, temporal asynchrony occurring when multiple auditory signals conflict and create the perception of an echo.

With that in mind, I will have a look at research on individuals with ASD and explain what we know about the differences when compared to neurotypical individuals [7] concerning multisensory integration. Then I explore why individuals with Autism Spectrum Disorder (ASD) and hearing impaired (HI) individuals can provide special insights into these topics.

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 on audio-visual speech processing and how that relates to multisensory integration as a whole. This is laying the basis of our experiment, such that after the review we can settle on an experimental design and formulate an informed hypothesis.

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 Meredith and Stein (1986), who initially recorded single-cell neurons in several animals, finding that some neurons respond differently to specific sensory inputs. They termed the neurons that react to input in multiple modalities "multisensory", proving that multisensory convergence is a common and essential concept in sensory processing. In their later book, Stein and Meredith (1993) put 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 mutually exclusive sensory information from all modalities has to be integrated by the nervous system to form the coherent picture of the environment we are used

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

to. From the unitary[8] perception of the world it follows that at some point during the processing of any isolated sensory input it has to be incorporated into that very unitary image of the environment. There has to exist some mechanism linking the isolated percepts to each other to create sensory compositions. This is hard to explain with the assumption that there are independent, non-interfering processing pathways for each sensory modality. We have to concede that sensory modalities at the very least have to interact and can likely affect another. This research into information from various sensory modalities perceptually binding is grouped under the umbrella term of multisensory integration and has been studied for more than a century by now. A notable early example is Stratton (1896), who experimented with vision-distorting glasses, finding that he was quickly able to adapt to the sensory discrepancy between inverted vision and haptic feedback of his environment. This suggests that multisensory integration has to occur, in contrast to isolated modality-specific processing. It follows that some mechanism, processing the sensory information, can acquire additional resources from other modalities and thereby adapt to more advanced tasks, even successfully dealing with partially incomplete inputs. For us, being interested in speech perception, the most relevant multisensory interaction is that between auditory and visual information.

In the following, I introduce common paradigms used to study the integration process and important results reported regarding the perception of temporal asynchrony. I will also examine special sensory circumstances, such as in ASD and age-related hearing loss, and look at current research on SHPDs.

2.1.1 Speech and visual lip movement

A strong demonstration of multisensory 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 inferring auditory speech signals from visual stimuli. The study, being conducted on NH participants with functional magnetic resonance imaging (fMRI), showed that access to only visual lip-reading information was enough to specifically activate

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

areas that are known to be involved in auditory language processing. 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 movements, 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 from the participants is specific to language-related processing and speech processing is routinely done utilizing compounded sensory input from the auditive as well as visual modalities.

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 present in the stimuli. 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 auditory input 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. They found an increased accuracy in the comprehension of auditorily presented words with visual articulatory input present by up to three times when compared to the audio-only condition. The team observed that the performance-difference between AV and audio-only condition was highest with a medium SNR (-12). They take this to mean that the human perceptive system might be highly attuned to only partly corrupted inputs, corresponding with a common real-world scenario, with all kinds of (slightly) 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 electroencephalography (EEG). They extend the findings by Ross et al. (2007) by examining continuous speech versus single syllables, providing a more naturalistic framework. The team reports an increase in speech understanding performance for the audiovisual compared to audio-only condition, 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.

2.1.2 Speech and gestures

Another well-established field of research is audio-gestural synchronization. The idea that listeners constantly incorporate information about hand gestures into their processing of speech fits well within the framework of multisensory integration. Specifically for speech and gestures, synchronizing effects between gestures and speech have been demonstrated in a recent replication of a classic study by McNeill (1992) on gestural synchronicity by Pouw and Dixon (2019), who used motion tracking to observe participants' hand gestures while they were either exposed to a 150ms delayed auditory feedback (DAF)[9] or heard themselves normally. Looking at speech production performance, they found that the benefits of audio-gestural synchronization were biggest when the adverse DAF was present. They suggest that in noisy and other environments counterproductive to speech transmission, a stronger binding by synchrony of gestures and speech follows and propose that the observed neural synchronization especially functions to maintain the stability of speech rhythm under noisy, adverse conditions. In another study by Biau et al. (2015) it has been put forward using EEG that rhythmical hand gestures, congruent with speech stimuli, so-called beat gestures, have a significant tuning effect on the low-frequency oscillatory bands in the brain, where theta activity synchronized with the rhythmic gesture, effectively aiding in predicting the onset of the next word. This would be one possible explanation as to how multisensory perceptual binding is realized on a neural level.

2.1.3 The McGurk effect

Extensively studied and well-known in the context of multisensory integration is a classical illusion dubbed the McGurk effect after the first team to note its existence (McGurk and

[9]DAF occurs when a speaker hears her own voice in a (slightly) delayed manner, which has been shown to induce stress, see Badian et al. (1979) and negatively impacts speech production performance. Usually, this occurs when the speaker is wearing hearing aids or a smart hearing protection device, but a karaoke-microphone connected to a speaker with some latency is another easy example where DAF could occur.

MacDonald, 1976). To produce the effect, they took a video of a speaker uttering a syllable of the structure consonant-vowel and replaced the consonant phoneme in the auditory stream of the video clip with a different phoneme. The replacement and the original form an auditory pair [10], 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 video containing "ba" and the auditory information from the "ga" recording, 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, 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 their experiment and were able to show the illusion in the independent dimension in a speeded classification task. Their paradigm is based on the idea that if two dimensions of a stimulus can be attended to independently, meaning the dimensions are perceptually independent, then irrelevant variations along one of the stimulus dimensions should not affect the RT in a discrimination task regarding the other dimension. A classical example is color and shape; in the study by Soto-Faraco et al. (2004) audiovisual recordings of "nonwords" were used, combining syllables to nonexistent words, creating a McGurk pair in the second syllable that was not targeted. Participants were asked to identify the first syllable that was independent of the illusion. A McGurk illusion across the unattended dimension did affect RT, effectively showing that multisensory integration happens automatically and we are unable to just disregard one modality in perceptual processing.

However, some research suggests that using the McGurk illusion is not a fine-grained enough measure to accurately assess audio-visual integration and may hinder research

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

regarding the automaticity of integration Rosenblum (2019). They make a compelling argument that The McGurk effect should not be conflated with speech integration itself and does not comprise a sufficient indicator of the latter.

2.2 Multisensory signal delay, asynchrony, and temporal window of integration

Based on the framework of multisensory integration introduced in subsection 2.1, a sensible area of research might be the limits of integration. Some research about properly functioning integration was already presented, but what happens in situations where integration fails? In a naturally occurring dialogue this may not be the first thing that comes to mind, but in an increasingly digital world of indirectly transmitted speech, we come to note that the temporal alignment of visual and auditory information matters. Think of the mild annoyance when the subtitles are slightly off, or even gross misunderstandings during an online video conference caused by temporal misalignment.

Upper bound (TWIN) A popular term for expressing perceptual binding is the temporal window of integration (TWIN), which specifies the range of AV asynchrony within which multisensory integration performs optimally. It is a probabilistic concept, predicting the likelihood of whether stimuli across different modalities will be perceptually bound or not. Outside of this window, the likelihood of integration decreases and without perceptual binding the asynchrony becomes noticeable and speech perception might be impacted (Stevenson et al., 2012).

van Wassenhove et al. (2007), utilizing the McGurk effect, performed a simultaneity judgment (SJ) and an identification task where syllable pairs were presented both as visual and auditory stimuli. These pairs had to be identified by the participants. In an SJtask, the participant is presented with an auditory and a visual signal asynchronous to each other and has to decide whether those stimuli occurred simultaneously or successive. Usually, the subjective onset asynchrony (SOA) is varied for the presentations. When the percentage of the "synchronous" responses is plotted, a Gaussian curve emerges. The

peak of that curve is denoted as the point of subjective simultaneity (PSS) (Vroomen and Keetels, 2010). Looking at this Gaussian-shaped curve, the TWIN can be approximated by the standard deviation (SD) of the curve, representing a window of strong perceptual binding, such that two stimulus components are treated as the same event. They found that AV stimulus pairs forming a McGurk pair were more often judged synchronous than non-fusing AV pairs. In this scenario, the optimally performing temporal window of integration is estimated to be around 200ms wide, ranging from -30ms (auditory leading) to 170ms (visual leading). With their findings they conclude that AV integration can usually compensate AV asynchrony well within this TWIN, making AV bi-modal integration relatively resilient against temporal asynchrony. With this, they tried to recreate the original findings by Sumby and Pollack (1954), who investigated audiovisual integration and the potential of one modality to enhance the other.

A more recent audiovisual delay study Li et al. (2021) noted that in a standard audiovisual SJ task with stepped delays from -400 to 400ms delay, roughly 50% of the participants incorrectly judged the 200ms delayed stimulus to be synchronous. Even in the 400ms condition, around 10% of the 27 participants still judged the stimulus as being synchronous. In a second experiment looking at audiovisual causality, Li et al. (2021) report for a 400ms AV delay around 15% "synchronous" responses across conditions involving speech with high causality and 25% for speech in their low causality condition. This demonstrates that the temporal corrective capacity of some underlying sensory integration mechanism is surprisingly strong. The authors compared a high causal relationship[11] between auditory and visual stimulus component condition against a low causality condition. AV synchrony perception was impacted more 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, less "synchronous" responses for the asynchronous conditions were recorded, suggesting some form of causal inference can help in asynchrony compensation. Another important finding related is that the temporal

[11]This means that the visual stimulus component shows a plausible source of the sound. One example from the study is the display of a pen being clicked paired with the click sound.

order of the modal information seems to matter. Several TWIN studies suggest that the speech-specific audiovisual TWIN is asymmetric, being larger for visual-leading stimuli over auditory-leading stimuli (van Wassenhove et al., 2007; Maier et al., 2011; Stevenson et al., 2012). Further research suggesting that tolerance for visual-leading asynchrony is bigger can be found in Maier et al. (2011), who conducted a study where they compared different component asynchrony in audiovisual stimuli. Investigating the difference in TWIN for speech stimuli using audiovisual SJ and temporal order judgment (TOJ) [12] tasks, they found that this holds for speech perception specifically. 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. These findings match the asymmetry of TWIN found in van Wassenhove et al. (2007). 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[13]. Maier et al. (2011) 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. Furthermore, they report a more narrow and asymmetric TWIN for unmodified speech stimuli in contrast to distorted or rotated speech argue for the presence of a highly specified recognition system for speech that is not purely dependent on causal correlations but also features some specialized statistical recognition for natural language. The results suggest that humans rely on a learned relationship between visual temporal cues and auditory information, especially when processing speech. One explanation given for the observed asymmetry is that hearing the sound before seeing the source is quite an unnatural situation because light travels faster than sound, usually arriving earlier at the individual[14] (Stevenson et al., 2012). The size of the TWIN for some non-speech stimuli is smaller, with Petrini et al. (2009) measuring a 112ms window

[12]a TOJ task is similar to an SJ task with the difference that the participant now has to report which stimulus component was perceived first. Usually, there is no option to declare them as synchronous. If we plotted the percentage of "stimulus component A first", an S-shaped logistic function would emerge. Here, the SOA would be denoted as the point with 50% "A first" responses.

[13]rotated speech refers to audio signals that are spectrally rotated in the frequency domain, preserving the temporal features of the signal, yet rendering it unintelligible.

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

in an audiovisual SJ task with drumming sounds.

TWIN looks at the breaking point of perceptual binding in an SJ task and is used to approximate the asynchrony between two stimulus components where perceptual binding fails and we start to perceive the components as separate events; The just noticeable difference (JND) denotes the smallest AV asynchrony where we consistently respond correctly in a TOJ task. The JND measures participants accuracy in responding "video first" or "audio first" by computing the difference of the audio-leading SOA at 25% of "synchronous" responses and the 75% video-leading point divided by two. For an SJ task, it is easier defined as the point where the participant correctly responds "not synchronous" 75% of the time (Vroomen and Keetels, 2010). While it makes sense to assume that the TWIN and JND are not quite distinct and should show strong codependence, empirically this is often not the case, possibly due to different cognitive biases employed when solving an SJ versus a TOJ task. (Zampini et al., 2003).

The JND does not seem to be fixed, it can vary depending on the specific needs of processing in a specific environment (Eg et al., 2015).

2.2.1 Minimally noticeable asynchrony

Regarding perceived synchrony and TWIN, it would be of interest whether we can quantify just how small a temporal asynchrony can be noticed and whether there is a detection threshold. This is the research question under the concept of the just noticeable difference (JND). This ability is highly dependent on the type of auditory signal used. People are generally very capable of detecting temporal delays in their own voices. Agnew and Thornton (2000), using 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. Both teams demonstrate that auditory lag with DAF applied is already perceived as annoying to the speaker at around 20-30ms and speech production performance decreases. Within the same study, Stone and Moore (2002), looked at the permissible delays in hearing aids for hearing impaired participants and, also utilizing DAF, identified that no disturbance is noticed under 30ms for regular speech.

For the purpose of identifying a threshold of a minimally perceivable AV asynchrony threshold, studies looking at DAF cannot be applied at face value here, because the detection threshold for own voice recordings consistently seems a lot lower than for external voices. Studies using DAF rely on internally produced speech, where subjects are likely to notice inconsistencies and asynchrony faster due to lifelong exposure to their own voice. Additionally, there are potentially large differences in the processing of internally produced speech and external speech signals, such that a conclusion for general speech perception is limited in predictive power.

Returning to studies on AV speech perception, Vatakis and Spence (2006a) looked at the sensitivity of normally hearing (NH) participants towards audiovisual asynchrony for speech and nonspeech stimuli (musical stimuli in this case) and found that the JND for speech is lower than for other tested stimuli and found the detectable threshold on average to be around 100ms for speech stimuli 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) report average asynchrony detection thresholds of around 200ms when using unfiltered continuous speech.

This would suggest that the JND highly depends on the type of situation of the perceiving individual and the length and complexity of the auditory stimulus. Eg et al. (2015) report in their review that for continuous signals the JND would be much higher than for a short alarm signal, stressing that the audiovisual JND is highly dependent on the context and content of the conveyed information. To verify the impact of the type of stimulus on AV synchrony perception, they conducted an experiment consisting of an SJ task using audiovisual stimuli from the domains of speech perception (news coverage), physical action (a chess game) and music (playing drums). They found a significant change in the size and shape of the TWIN, and, specifically looking at the JND, they found that to be smaller for speech-related stimuli than either of the other conditions.

Effects of echo Turning towards the effects of an echo present in the auditory signal and examining the possible impact on asynchrony perception, Lezzoum et al. (2016) measure a smaller asynchrony detection threshold for simple non-speech stimuli, a bell signal with

delayed echo was detectable by 20% of the participants at 8ms. Zakis et al. (2012) estimate experts even to be able to detect an AV delay in music already at 3-5ms. 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), unattenuated external voice and 20dB attenuated external voice. The tolerance for external voices is the condition that most directly reflects the chosen conditions in our experiment and this reflects general speech perception in the real world more accurately. They found slightly elevated annoyance ratings in the unattenuated condition for the NH participants, which was absent in the attenuated condition. Attenuated stimuli resulted in the first notable increase in annoyance between 20 and 30ms. By and large, HI participants were more tolerant towards auditory delays, with the authors suggesting that experience in using hearing aids likely enlarges the delay tolerance in participants.

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% of the participants at 16 - 22ms delay. In their setup, participants were able to tune the temporal asynchrony between auditory and visual stimuli 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 that was 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: a stronger attenuation results in

a higher threshold for perceivable AV delay. 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.

Subperceptible Asynchrony 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 for the notion that rapid temporal recalibration is determined by the prior physical AV asynchrony and not by the temporal judgments given by participants. They examined the shifting of the point of subjective simultaneity (PSS)[15] in AVSJ trials, who had so-called "adaptor" TOJ trials inserted. Both types of trials had varying asynchrony, audio-leading as well as video-leading, and the measured shift in PSS was shown to correlate more with the physical temporal order of the stimuli components (either visual-leading or auditory-leading) than with the participants' reported synchrony judgments. If this holds, the sub-perceptual temporal lag would indeed influence our speech perception without a detectable change in the percept decisions taken in a test setting. Further, they were able to demonstrate that the shift in PSS was present after just one asynchronous trial, hence the name *rapid* recalibration. The team concludes that the recalibration is "likely mediated by early sensory processes, which operate without reference to one's conscious appraisal of prior temporal events." (Van der Burg et al., 2018). All this was shown using very simple pip tones and visual flashes, such that a generalization to more complex speech stimuli is not possible. Whether the findings hold for AV speech would be an avenue for further research.

2.3 Special circumstances in speech processing

2.3.1 Hearing impairment and effect of Age

Looking at age-related hearing loss, Rosemann and Thiel (2018) brought forward fMRI data to suggest that with increased hearing loss, the AV integration gets stronger. This was shown through an audiovisual McGurk illusion, where a more pronounced multisensory

[15]The PSS is the point with the highest likelihood for a "synchronous" response in an SJ task. Usually, it is computed as the mean of the "synchronous" responses.

fusion effect is indicative of stronger audiovisual integration. This would suggest that there is likely no linear relationship between hearing capacity and integration and it supports other claims discussed earlier that integration shows a window of maximal effectiveness under moderately adverse conditions (Ross et al., 2007; Crosse et al., 2015), as which we could count mild hearing loss. Du et al. (2016) suggest that increased reliance on visual speech information during integration seems to be a common and effective way for older adults to compensate for impaired auditory speech perception. Petrini et al. (2009) also report that there is a clear tendency for NH-participants to be less tolerant towards temporal delay than HI-participants. Even more, 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 delays. For us, this means that designing the experiment with NH people in mind will later apply to HI subjects as well, since a perceptive threshold found for NH individuals is likely smaller than for a HI individual.

2.3.2 Autism spectrum disorder and possible differences towards neurotypicals

Autism Spectrum Disorder often presents itself in social interaction and communication deficits and often goes along with atypical processing of sensory information (American Psychiatric Association, 2013). Multiple studies have established findings regarding common features within the sensory processing in individuals with ASD. One rather well-researched processing difference lies in recalibration speed, or potentially even the overall capacity for re-calibration. As indicated by 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 normally hearing individuals. This rapid re-calibration is very diminished in NH individuals with ASD, one consequence being a lower susceptibility to the McGurk effect. Slower re-calibration also results in a reduced ability to rapidly adjust

to adverse speech perception situations (Beker et al., 2018; Turi et al., 2016). Following Smith and Bennetto (2007); Beker et al. (2018), this provides a possible explanation why individuals with ASD typically start to speak later and under-perform in language production, with language deficits being a recognized symptom of autism (American Psychiatric Association, 2013). In Brandwein et al. (2013), this is discussed and extended to more general, basic non-speech and non-social stimuli, suggesting this to be a consistent effect present even 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[16] is likely the underlying cause for many traits typically associated with ASD, such as impairments in social and communicative skills.

Beker et al. (2018) report a delayed development of multisensory integration in individuals with ASD, with especially adverse consequences for AV speech integration, likely leading to impairments in speech processing and general communication after development. They suggest early sensory training during development and propose the possibility of preventing the delayed development of MSI and thereby mitigating the hallmark symptoms of impaired communication skills American Psychiatric Association (2013) in individuals with ASD. Especially when decreased sound tolerance is present in individuals with ASD, subjects with ASD show elevated physiological response to sounds and rate subjective discomfort higher, but there is no evidence that they habituate to sounds slower or that

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

they have lower auditory detection thresholds (Kuiper et al., 2019).

The comparison between demographic groups, such as individuals with ASD, is outside the scope of the initial experiment, but it is targeted for future research.

2.4 Asynchrony chosen in the experiment

Regarding our study examining temporal asynchrony, an essential question to ask is what AV asynchrony and which modality combinations have been looked at in the literature, and whether we can make some prior claims about typical audiovisual delay detection thresholds or whether the literature converges on a range for the size of TWIN. As an overall goal, we want to simulate the scenario of transmitting sound via an SHPD and as such look at phenomena likely occurring here. The two main points we are interested in are the effects of a larger AV asynchrony when compared to the use of HPDs and the selective attenuation of the signal through the DSP, leading to the presence of multiple potentially perceivable signals. Since we attenuated the echo in our conditions where an auditory echo is present, we should expect a similar effect for perceptive thresholds asLezzoum et al. (2014), who used the same overall technique for sound attenuation, although effect translation is not given since the attenuation function we used is specific to a certain device and fit. For better comparability and approximation of a real usage scenario, the auditory lags in the simple delay and the passively transmitted echo condition should be of the same size. In our simple setup we chose three conditions: a 0-condition, to get a benchmark result, a condition with a small AV asynchrony, and a condition with a large and obvious asynchrony, where we expect to obtain clear results which is enabling us to verify our general hypothesis. We seek to show that speech processing is indeed positively dependent on synchronicity within our specific setup. The two asynchrony conditions either have an additional attenuated echo included or not, yielding five conditions in total that we wish to compare. We expect performance to suffer more when the simulated echo is present compared to conditions without an echo. To ensure we can demonstrate a loss of perceptual binding in our results, our choice of value for the large asynchrony condition should be larger than the optimal performing TWIN. Even for delays larger than 200ms

(the size of the TWIN found by van Wassenhove et al. (2007)) we can still expect some impact on audiovisual speech understanding happening although likely the perceptual binding is less present. An asynchrony not big enough would result in a higher percentage of "synchronous" responses by participants in the SJ task. We can assume this through the loosely Gaussian-shaped response patterns in typical SJ tasks as those found by Maier et al. (2011). Upon reviewing the literature with this specific question in mind, we decided on 400ms for the larger AV asynchrony value. This is estimated to be distinctly noticeable, with an unambiguous impact on speech reception performance.

Slightly more complicated is the choice of the smaller value, since ideally, we want this value to be below conscious perceptibility of the participant. We do not know yet whether this will affect speech performance. A review of intersensory synchrony (Vroomen and Keetels, 2010) concluded that temporal lags among different modalities below 20ms are usually unnoticed, and they put forward that this is due to a strong natural tendency to reduce errors and adaptive temporal recalibration. However, this is a general claim about any intersensory integration, it might not specifically hold for language-specific stimuli. Although Vroomen and Keetels (2010) report findings identifying speech as a special scenario in AV perception (van Wassenhove et al., 2007), and an influence of stimulus complexity (Vatakis and Spence, 2006a) they caution prior claims: "that stimulus factors, such as rise time, need to be controlled more carefully before any sensible comparison can be made across audio–visual speech, complex stimuli, and simple combinations of flashes and beeps." (Vroomen and Keetels, 2010) Therefore, the situation with speech stimuli at small asynchrony is not clear cut and warrants further research. Additionally, the previously presented passively attenuated echo present with an SHPD possibly has an impact on speech perception not studied in prior work to our knowledge and could complicate a possible answer on the value of the smaller AV asynchrony.

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 simple AV delay and the condition with additional passively transmitted echo. Taking the literature on detectable thresholds into account, using

complex speech stimuli, we are confident that a vast majority of our participants will be able to detect neither the "simple" AV asynchrony nor the additional simulated attenuated echo. Should we still find any speech performance impact in these conditions, this should be a good indication for multisensory mechanisms acting below the thresholds of conscious perceptibility involved in speech perception. This would ultimately call into question the utility of highly complex filters additionally introducing larger asynchrony in SHPD, at least in environments where speech understanding is critical.

2.5 Conclusion and motivation for the experiment

We saw that multisensory integration is a powerful mechanism in speech processing and is capable of bridging substantial temporal AV asynchrony. Multisensory integration increases resilience to adverse factors, such as background noise and distortion effects and provides a scaffolding for speech processing. Environmental cues, such as rhythm, bodily cues, and lip movement are integrated into speech perception. SJ and TOJ paradigms used to measure the degree and success of integration were introduced, but there are still open questions as to how speech perception performance interacts with small audiovisual asynchrony. The PSS and JND provide indicators about the extent of perceptual binding, which then reflects in the size of the TWIN, but there are concerns that a minimal perceived difference is not the same as a minimal difference impacting speech perception.

The noticeable AV asynchrony threshold for a smart hearing protection device is not easily captured in a generalized single number. We saw that it is highly specific to the situation of the speaker and hearer, and interpersonal differences in perceptive processing, the type and fit of the device, and the audio properties of the signal. This value is also highly task-dependent, as we have seen in highly differing results in TOJ and SJ paradigms. Furthermore, perceptive synchrony thresholds often refer to variables collected in paradigms relying on self-report. However, even when a specific subjective AV asynchrony detection threshold is found, there is no reason for it to strictly coincide with a threshold for detrimental effects in speech perception performance, as that could speculatively degrade already within subperceptible asynchrony. There is a knowledge gap

regarding the impact of this subperceptible AV asynchrony.

Studies employing various strategies to define a temporal window of integration, although very helpful in sizing the overall scope of integrative capacity, fail to answer questions about speech processing efficacy under realistic, small, persistent AV asynchrony and other slightly adverse conditions such as background noise.

Specific echo configurations created through the cooccurrence of a passively attenuated audio signal and an actively transmitted signal are likely when wearing an SHPD and their effect and interaction with speech processing are critically missing in research.

With this in mind, we created a browser-based online study to help investigate the effects of wearing an SHPD on speech processing. We will reproduce effects seen in previous SJ paradigms to verify a typical TWIN in our setup and measure the accuracy and speed of speech perception and investigate the presence of adverse effects on language processing with small asynchronies.

Technical limitations The present study was built with PsychoPy 3 (Peirce et al., 2019) for which it has to be acknowledged that our study being browser-based has technical limitations being discussed Bridges et al. (2020). 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. For PsychoPy run via pavlovia.org, executed within Chrome browser on a Windows 10 operating system [17], we can expect an average RT variance of 0.39ms and variance of audiovisual synchronicity of 3.01ms. These values would be slightly higher for Edge users[18] and Firefox users[19]. The mean constant latency for RT measurements can be expected to be around 30 to 60ms, depending on the soft- and hardware combination, in our model case of using Windows 10 with Chrome being on average at 43.95ms. The average AV asynchrony varies more wildly, from -10.21 (meaning auditory leading) to 190.45ms using Ubuntu and Firefox. Again, our prevalent model of Windows 10 and Chrome has an average lag of 65.32ms. Concurring with the

[17]This is the combination we estimated to be most prevalent among participants. This was confirmed during data collection, see Figure A.5 and Figure A.4.

[18]mean RT variance between 1.74 and 2.03ms, depending on architecture, and mean AV synchronicity arond 3.69-5.6ms. For more tested combinations, please refer to Bridges et al. (2020).

[19]mean RT variance 1.96ms and mean AV synchronicity around 3.9ms

authors, we disregard these lag values, since a constant error, although yielding false absolute values, will not affect the relative comparison between conditions. Nevertheless, the large variance (in comparison to the size of our 10ms asynchrony) between different systems is a limitation possibly introducing unforeseen effects in any results for perceptive AV thresholds. Further, differences in internet speed can be disregarded since all resources are loaded locally during the experiment. Regarding hardware, the experiment makes no use of the computer mouse, eliminating some possible errors stemming 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. The overall recommendation there is to refrain from using absolute response times and instead rely on control conditions verified within the actual setup each participant conducted the experiment on. Added up, this leaves us in the best-case scenario with ±3.4ms of variance. Taking the sum of possible variances together, this results in us expecting the smallest meaningful results even under optimal conditions at an audiovisual asynchrony of at least 10ms. Should a large percentage of participants use a less reliable combination, detecting such small AV asynchrony will be challenging. To be able to investigate smaller asynchrony, we would require a more direct control only possible in a lab setting currently not available to us.

3 Experiment

We are interested in sub-perceptible delays and the specific echo that occurs when the auditory input signal is incompletely attenuated by the SHPD such that the wearer perceives both the attenuated original signal and the actively transmitted delayed output signal.

In this experimental setup, we want to establish a valid indirect measure of speech comprehension when presented with audiovisual delay. We choose 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 to respond to the task. We also record the response accuracy to be able to detect any secondary effects, as lower accuracy could indicate a

deterioration in language processing and comprehension.

This experiment enables us to compare different audiovisual delay conditions without reliance on subjective (conscious) feedback of whether the asynchrony 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 on the referent choice task, we can now gain insights on very small audiovisual delays and attenuated echoes without the need for the participant to actively perceive and report a delay, effectively eliminating the lower boundary of testing present in JND paradigms.

Assumptions

- There is a universal underlying mechanism of multisensory 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.

For an overview of the analyzed dependent variables refer to Table A.1.

3.1 Method

3.1.1 Participants

The experiment recruited a total of 72 participants via the university's internal mailing list targeting cognitive science and psychology students. From those we later excluded 22 participants from the analysis. All 50 remaining participants participants were native German-speaking adults with normal or corrected to normal vision and normal hearing.

From those 50 analyzed participants 26 were female, 20 male with a mean age of 24.14, ranging from 19 to 56. The participant data was not collected on the remaining 4 participants due to technical issues. 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 deletion 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 instead of wireless headphones in an attempt to minimize distracting environmental noises beyond our control and to exclude additional latencies introduced by wireless sound transmission. Eight participants were left-handed.

3.1.2 Materials

Media files Video and corresponding audio files were taken from the Oldenburg linguistically and audiologically controlled sentences (OLACS) Corpus created by Uslar et al. (2013) and used with permission by Rosemann and Thiel (2018). These are full HD recordings of a male German native speaker uttering German sentences centered on his lips. They contained a neutral black background and were recorded at 25 fps, all videos having a bitrate higher than 100mbps. They are extensively controlled for speech reception threshold and plausibility within adult native German speakers with full hearing capacity. This ensures that generally, subject and object should be semantically equally likely to be the actor in the sentence. Of the full 160 sentences, we selectively use 80 for the main task, half of which follow a subject-verb-object (SVO) structure, the other half an object-verb-subject (OVS) structure, where both the subject and the object are modified by an adjective, 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. The entities referred to by the nouns are either animals, professions, or mythical creatures, typically appearing in tales targeted at children. They were selected to be readily identifiable, with a clear prototypical image coming to mind. The full list of utilized sentences is available in the appendix.

To create the five different conditions from the stimuli, audio and video streams were separated, the audio stream was then modified using Matlab (MATLAB, 2020), adding the necessary delay and transforming and adding the simulation of the passively transmitted echo. The attenuation function was computed from pink noise recorded with a SHPD prototype called the Auditory Research Platform (ARP 3.1) developed within the NSERC-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS, Montréal, QC, Canada). The sampling rate of the original audio is 48000 samples per second. To generate the simple delay conditions, the rounded number of zeroes was added in front of the audio signal, resulting in added silence of specified length at the beginning of the audio signal. The resulting longer audio sample was then merged with the video stream, where the last still frame was inserted again for the last few milliseconds where a sound signal was playing, but the video already terminated. This was done before the experiment to minimize AV synchrony issues resulting from different media playback handling in different browsers. The rounding error while merging is 1 sample per frame, such that for the average video of 3-4 seconds at 25fps the AV asynchrony can in the worst case reach

2ms[20].

To generate the simulated passively transmitted signal, the algorithm estimates a transfer function between the outer-ear and the inner-ear microphone in order to simulate a passively transmitted signal. With the help of 30 seconds of pink noise recordings with two microphones, one located on the inside of an SHPD, the other one on the outside, the filtering coefficients were computed, which were then applied to all signals in the frequency domain via a fast Fourier transform. These pink noise samples were recorded in an audiometric booth and the original pink noise was played at 85dBA over 4 loudspeakers in the booth, while a researcher involved in the project was wearing the mentioned ARP 3.1 SHPDs with short soft comply foam tips. We thus simulated a signal being passively transmitted and added it to the delayed original signal to create the audio stream where both signals are present.

The video and audio streams were then merged and compressed using FFmpeg (Tomar, 2006) into h.264 mpeg4 format, which is compatible with most modern browsers. Due to browser playback issues during testing, the videos were also resized to 1280x720px resolution. The audio stream was left as is, repackaged into an AAC mp4 format with a sampling rate of 48kHz, 32bits/sample, which corresponds to the original. The original frame rate of 25fps was left as is to leave AV synchrony intact.

Images 53 of the corresponding images were 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 that did not have a direct fit in the MultiPic database were found via Google image search and were all licensed free for personal use, totaling 67 images used in the experiment. All of these were then manipulated using GIMP version 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 and their sources can be found in the appendix.

[20]assumed a 4 second video, which would have 100 frames: the last frame could be 100/48000 off, corresponding to 2.08ms

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

3.1.3 Procedure

Before taking part in the experiment, participants gave informed consent to the procedure and were informed that they could terminate the experiment via the escape key, should they want to withdraw consent. After, pseudonymous data were collected in a questionnaire, 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 asked 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 were also instructed to ensure adequate viewing conditions and subjectively adequate brightness of the screen. The entire experiment was conducted in one browser session requiring internet access, a keyboard, wired headphones, and a display. The experiment was inaccessible from a mobile device and records the participants' operating system, frame rate, resolution, and the browser used. All stimuli of the experiment were downloaded before starting to prevent and mitigate download speed, performance, or playback issues. Then, participants were 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. Participants were then, after indicating that they adjusted their volume accordingly, redirected to another browser window playing in fullscreen, which contained the entire experiment. After a brief instruction to the task, they were presented with 5 training trials to get familiar with the nature of the main task. No feedback on correct answers was given. Participants were

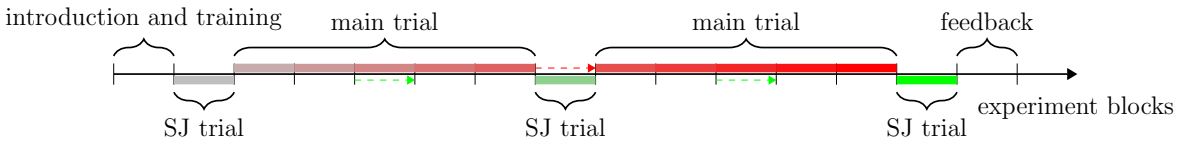
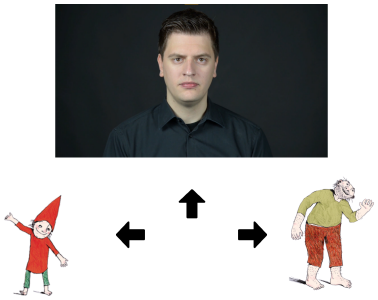


Figure 3.2: Temporal order of the entire experiment

Figure 3.3: Example screens of the referent choice task

(3.3.1) Example of target presentation



(3.3.2) Example stimulus presentation

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 reflected in the RT.

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

Target-noun reference task The presented target words were extracted from the sentences in the corpus. Every adjective was presented once as a target word, and every sentence was used twice, each time presenting a different target. The adjectives were displayed for 2500ms in the center of the screen sized at 10% of the screen height.

The *referent choice task* was a referent identification task, in which the participant had 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. The trials were interleaved with a total of 20 filler trials, generated identical to the actual trials, but the adjective shown is not corresponding to either of the two target images, the video clip and the images are still congruent, just the target is misleading and not modifying any of the referents.

These filler trials are later excluded from analysis, they are merely introduced to prevent inferential task solving. When a filler trial was presented, a randomly chosen adjective present in other target sentences was shown to avoid biasing through novel targets. 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% of the trials being fillers. Each trial was divided into target presentation and stimulus presentation and records the response starting with the onset of the video stimulus.

The target, which was one of the two adjectives present in the sentence, or a random adjective in a filler trial, is flashed in the center of the screen for 2500ms.

After target presentation, the participant was shown two images on the left and right half of the screen as visible in Figure 3.3.2. To let the participant have a look at the images and ensure proper identification, the images were first shown alone. Then, after 1500ms, a fixation cross was presented at the center of where the audiovisual stimulus will appear. After 1000ms of presenting the images and the fixation cross, the movie clip was presented and key presses were recorded. A left and right arrow were used to indicate the two target nouns and their corresponding responses on the keyboard. A third, upward arrow was presented alongside to remind the participant to press the upper arrow when no image fits (a filler trial).

Then, in the stimulus phase, the video clip stimulus was shown alongside two images corresponding to the left or right answer option indicated by the position of the images and helping arrows. The position of the stimulus images on the screen (left or right) was counterbalanced across participants.

The video clip was presented centered in the upper half of the screen, alongside the images in the lower half. The trial ended with a keypress registration of the answer, there was no hard upper response time limit. We used a counterbalanced Latin square design to ensure that participants saw each sentence in only one of the five test conditions. The trial order was pseudorandomized between participants such that trials in immediate sequence never have identical target words, experimental conditions, images, and sentences. In the pauses between each block, the trial progress was presented. Participants could determine

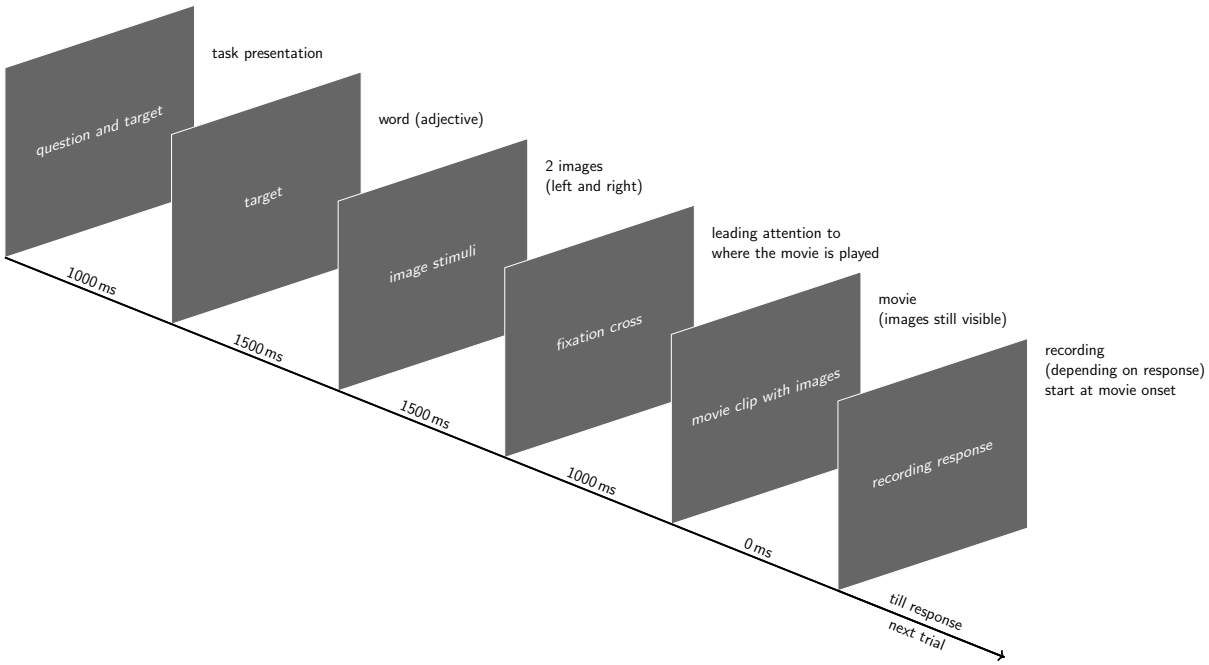


Figure 3.4: Presentation process in referent choice task

on their own for how long to take a break and could continue with a key press.

Adapted SJ task In this customized SJ task we presented participants a video of a German native speaker uttering a sentence of the same structure as in the referent choice task. Participants are shown one video per trial with the audio corresponding to the condition. Subsequently, participants were asked to answer two questions: One whether the auditory signal was perceived to be (1) synchronous with the visual stimulus or not, similar to a traditional SJ task, with a yes/no response recorded. Next, they were also asked whether any (2) auditory distortion, such as multiple overlapping audio signals, was perceived. These questions were asked for the same stimulus in sequence after it has finished playing. For each question, accuracy and RT were measured. The answer was recorded via a keyboard press with separate buttons for yes and no. The mapping of the buttons stays invariant throughout the experiment. The modified SJ task was performed 3 times, distributed once before, once after and once after half of the referent choice trials were completed. Each block consisted of 10 randomly ordered trials, and each of the 5 conditions was repeated twice in each block with a different sentence. The three blocks were Each block consists of the same 10 sentences, also taken from the OLACS set, but not presented in the main task.

3.1.4 Hypothesis

Table 3.1: Conditions

	Control Condition	Condition 1	Condition 2	Condition 3	Condition 4
descriptor	*control*	*simple delay, small*	*simple delay, large*	*small delay, simulated echo*	*large delay, simulated echo*
AV delay	0ms	10ms	400ms	10ms	400ms
Attenuated echo	no	no	no	yes	yes

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 RT, meaning that we expect longer reaction times and less response accuracy with worsening perception of the speech stimuli. Concretely, measured 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. We expect the additional echo to be imperceptible in the small delay conditions, such that between echo present versus simple delay we should detect no difference in speech perception. With a larger 400ms delay, for a large percentage of participants the echo should have a noticeable effect on speech perception, reflecting in lower accuracy and longer RT. 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 asynchrony, 400ms passively transmitted echo) is to verify that linguistic processing is indeed dependent on perceived synchrony (indicated loosely by

the TWIN) and noninterference, such that speech perception performance decreases, the further outside of the TWIN the asynchrony is. For interfering factors such as the echo, we expect it to impact performance only when crossing an analogous asynchrony threshold, whose exact position is not generalizable due to external factors. Due to the previously presented detection thresholds and results from the literature, we expect largely correct identification in the secondary SJ task for the 400ms delay conditions, but not for the presumably impercepted 10ms asynchrony conditions.

3.2 Results

Participant exclusion In total 72 subjects participated in the experiment. We excluded participants based on response accuracy in the referent choice task. As an exclusion threshold, we took an overall accuracy (target and filler trials together) below 80% for individual participants. This led to the exclusion of two participants. Frame rate has a significant effect on audiovisual asynchrony detection, as reported in Vatakis and Spence (2006b). To maintain comparable results, we excluded all participants with an average frame rate lower than 25, since the original video clips are encoded at 25 fps, which led to the exclusion of 5 more participants. Due to an imbalance in the 5 inter-participant groups, we had to exclude 15 more participants randomly to fulfill the assumption of counterbalanced sets. This resulted in a total of N = 50 included participants fulfilling all criteria (see Figure A.3).

3.3 Statistical analysis

We cleaned up the data using python, the statistical analysis was performed in R (R Core Team, 2021), employing linear mixed effects models for the RT analysis and binary logistic regression models for the analysis of accuracy in the referent choice task and the adapted SJ task. We used the R package lme4 to conduct the analyses (Bates et al., 2014). We started with a maximally random effects structure and when the model did not converge we simplified the random effects structure in a stepwise procedure.

3.3.1 Referent choice task

RT analysis To remove outliers, we eliminated all reaction time measurements more than 2.5 standard deviations from the individual mean. We also accounted for the respective position of the target, deleting points more than 2.5 standard deviations away from the mean of the respective target position. Since one assumption of our employed statistical model is normally distributed data, we used the boxcox algorithm (Box and Cox, 1964) to identify an appropriate transformation to achieve normal distribution, then we transformed the RT results with a square root function to approximate a normal distribution. To test our hypotheses on RT in the referent identification task, we treated the experimental condition, the position of the target adjective in the sentence (treatment coded as 0 and 1), and the trial number (z-standardized and treated as numeric predictor) as fixed effects. We also included random by-participant and by-item intercepts.

We fit two models to test our hypotheses: In the first model, we used treatment coding for the experimental conditions with the control condition as reference. This allows us to address a difference between the conditions such that we can see overall effects in reference to the control condition and observe the effect of simple delay (no delay, 10ms delay, 400ms delay) on accuracy and RT. In the second model, we directly assessed whether there was a difference between the 10ms and 400ms delay conditions with and without echo, respectively. For that, we added two sum-coded comparisons to the model, one comparing passively transmitted echo vs simple delay (where "echo present" was coded as -0.5, no echo (simple delay) as +0.5) within the small 10ms delay conditions, so that we could directly see whether the presence of a simulated echo produces a significant difference in RT or not. The other comparison was analyzing the same for the large asynchrony conditions, overall letting us infer differences in the perception of the simulated echo over different asynchronies. Further, we included an interaction between the position of the target adjective and the experimental condition.[21]

For an overview of the overall RT for each participant, see Figure A.6. As visible in

[21] The resulting formula tested was

$$RT \sim condition * position + target.position + trial.z.scaled + (1|participant) + (1|sentence) \quad (3.1)$$

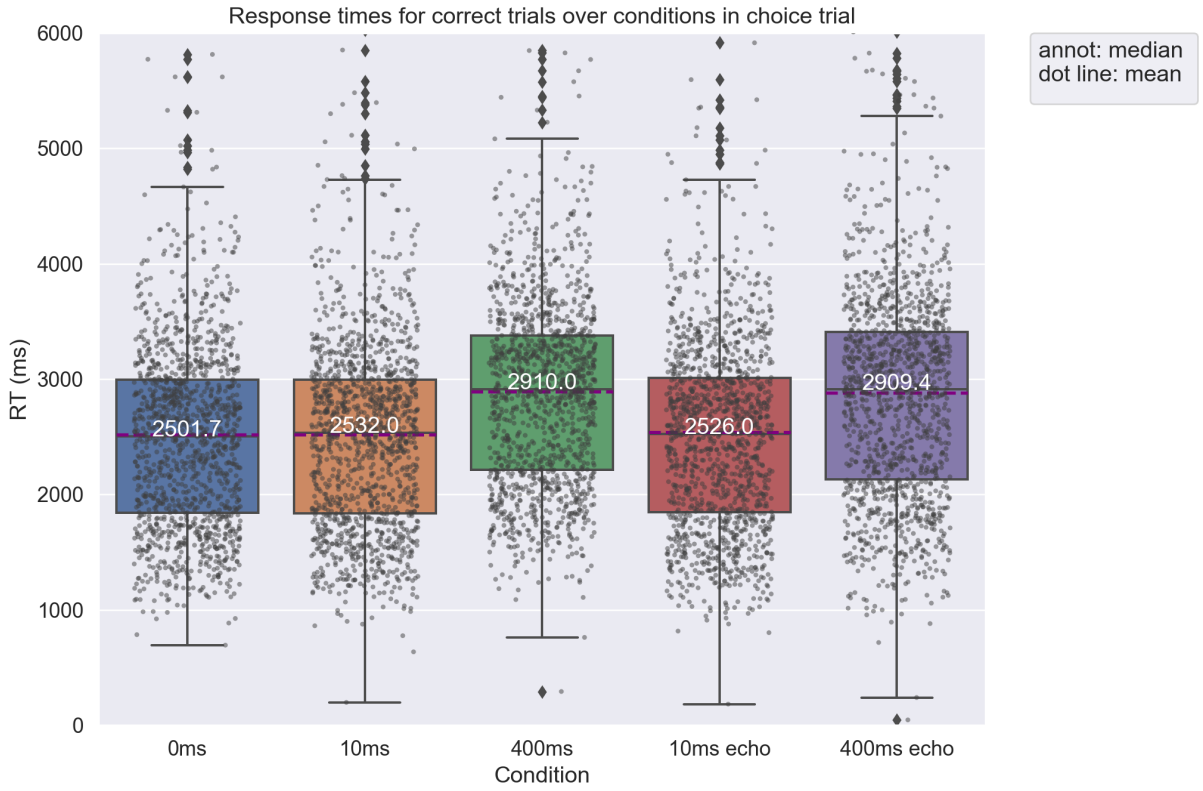


Figure 3.5: RT in referent choice task by condition

subsection A.4, there was no significant difference in RT between the control condition and the 10ms simple delay condition(β = 0.19, SE = 0.25, z = -0.75, p =0.454). Participants did not take significantly longer to respond in the 10ms delay conditions, regardless of whether an echo was present or not. This was not the case for the 400ms delay conditions, here both differed significantly from the control, the simple delay 400ms condition (β = 4.15, SE = 0.25, z = 16.36, p = < 0.001) as well as the 400ms echo condition (β = 2.86, SE = 0.25, z = 11.26, p = < 0.001). In the 400ms delay conditions, participants were able to respond faster when an echo was present. The position of the targeted noun in the sentence (whether it was mentioned first or not) did have an overall significant effect on RT (β = 10.05, SE = 0.25, z = 39.74, p = < 0.001). In the 400ms simple delay condition, participants responded later when they had to wait for the occurrence of the second noun (β = -0.81, SE = 0.36, z = -2.26, p = 0.024), which is an indicator that they did respond

with contrasts to check for echo, that becomes

$$RT \sim condition + crt10ms * target.pos + crt400ms * target.pos + trial.scal + (1|participant) + (1|sentence) \quad (3.2)$$

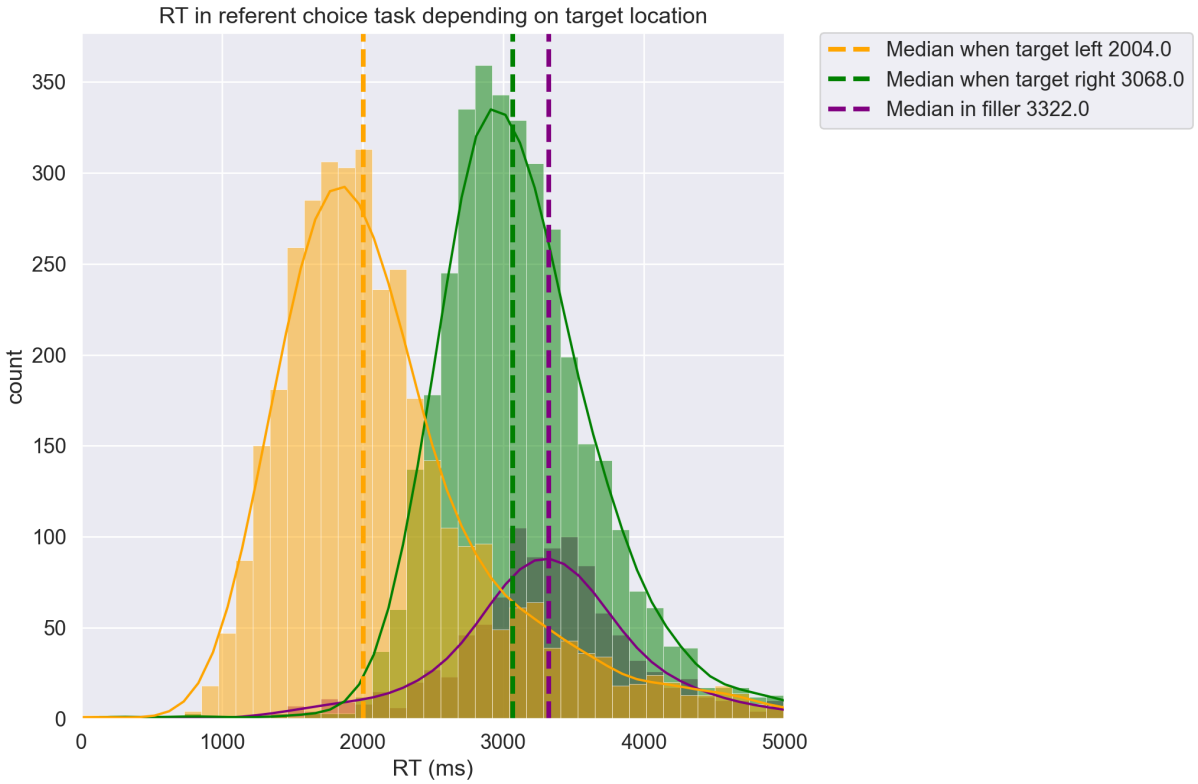


Figure 3.6: RT in referent choice task by location

when they heard the target and not earlier, using inferences. This interaction effect was not visible in both of the 10ms delay conditions. This suggests that for the two 10ms conditions, the effect of the target position was comparable to its effect on the control condition. For the 400ms conditions, participants responded faster if there was an echo than if there wasn't one, but only if the target was at an early position in the sentence.

Now additionally looking at the contrast coded model looking at the effects of the introduced echo, we can observe supporting evidence for the earlier results. Here, a significant difference in the large delay conditions (400ms) is visible, where people responded faster when the echo was present (β = 1.30, SE = 0.27, z = 4.79, p = $<$ 0.001). This effect is not observed in the small delay conditions (10ms). The trial number had a significant effect on RT (β = -1.11, SE = 0.07, z = -16.42, p = $<$ 0.001), which means that participants responded faster towards the end of the experiment.

Accuracy referent choice task To analyze the response accuracy within the same referent choice task, we used the same model structure as for the RT analysis in section 3.3.1,

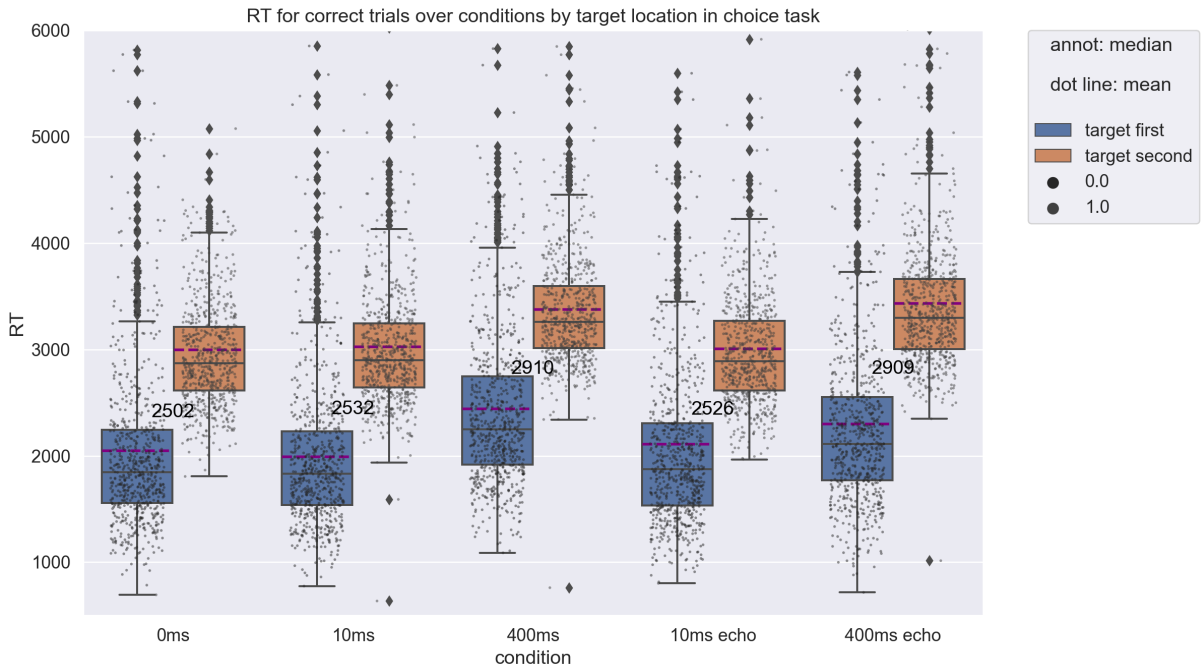


Figure 3.7: RT in referent choice task by location distributed over conditions

except that now the models were fit using a generalized linear mixed-effects model with no interaction of target position with the condition.[22]

All participants had a similar mean accuracy across conditions of almost 90% in the identification task. We found accuracy to be highest in the reference condition and only slightly lower in the other conditions(mean accuracy in control: 91%, other conditions ranging from 88 to 90%). This can be seen in Figure 3.8.

Participants had a marginally significant worse response accuracy in both large 400ms AV asynchrony conditions , for the 400ms AV delay condition without additional echo (β = -0.61, SE = 0.32, z = -1.93, p = 0.054), as well as the condition with echo present (β = -0.56, SE = 0.32, z = -1.77, p = 0.077). However, we found no significant difference in accuracy when comparing the control and the small AV delay conditions. Accuracy increased significantly over trial numbers (β = 0.17, SE = 0.06, z = 2.86, p = 0.004), with

[22]The resulting formula to compare conditions was

$$response.acc \sim condition + target.position + trial.z.scaled + (1|participant) + (1|sentence) \quad (3.3)$$

and for testing the effects of echo:

$$response.acc \sim condition + contrast10ms + contrast400ms + trial.z.scaled + (1|participant) + (1|sentence) \quad (3.4)$$

Coding of p-values in the tables follows the rules: p = 0: '***' p < 0.001: '**' p < 0.01: '*' p < 0.05: '.'

Table 3.2: Fixed effects with contrast for echo

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
(Intercept)	45.77	0.55	82.76	< 0.001	***
contrast10ms	-0.26	0.27	-0.98	0.328	
pos.second	10.12	0.14	74.69	< 0.001	***
contrast400ms	1.30	0.27	4.79	< 0.001	***
trial number (z-scaled)	-1.11	0.07	-16.42	< 0.001	***
contrast10ms:pos.second	0.12	0.38	0.32	0.752	
pos.second:contrast400ms	-1.50	0.38	-3.90	< 0.001	***

more accuracy towards the end of the experiment. The position of the target noun within the sentence showed a significant difference(β = -0.63, SE = 0.31, z = -2.00, p = 0.045), with participants being more accurate in responding when the correct noun was the first instance. There was no significant interaction of target position with any of the conditions.

Again, applying contrasts for echo vs. no echo analogous to the RT analysis, we found no significant difference in response accuracy with or without echo.

3.3.2 SJ task

We also analyzed the interleaved adapted SJ task and checked for response accuracy for two independent questions. To recall: as described in section 3.1.3, the first one was a classical "synchronous or not?" question, while the second question measured perceptivity of multiple auditory signals (distortion) To test our hypotheses on RT in the simultaneity judgment task, we treated the experimental condition, the position of the target adjective in the sentence (treatment coded as 0 and 1), and the block (either 0, 1 or 2 as numeric predictor) as fixed effects. We also included random by-participant and by-item intercepts. [23] Here, we analyzed the models with respect to response key, giving us the direct "synchronous" or "not synchronous" responses common in other SJ paradigms.

[23]The resulting formula comparing conditions tested was

$$synchrony.response.key \sim condition + block + (1|participant) + (1|sentence.id) \quad (3.5)$$

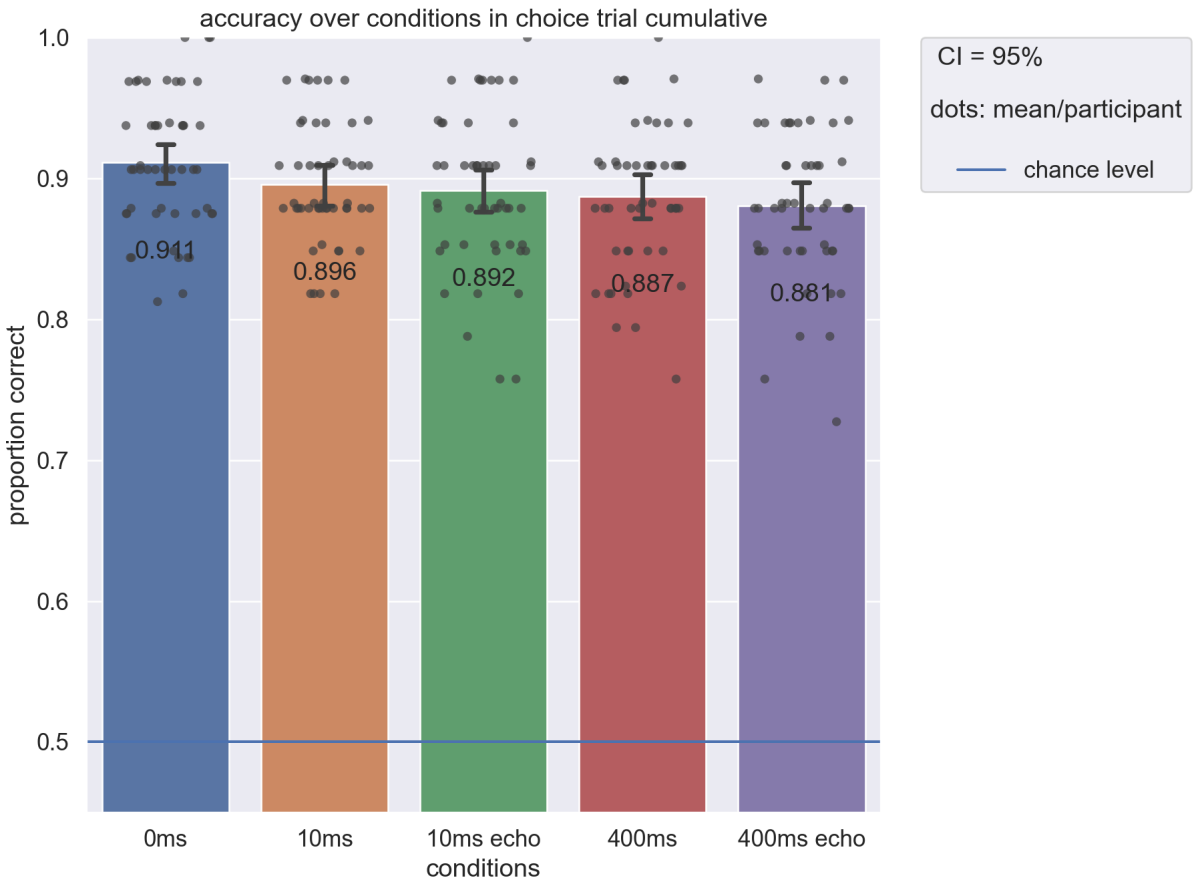


Figure 3.8: Response accuracy by condition, referent choice task

Response key in synchrony question Analogous to the referent choice task, we fit two models here: In the first model, we used treatment coding for the experimental conditions with the control condition as reference. This allows us to address a difference between the conditions such that we can see overall effects in reference to the control condition and observe the effect of simple delay (no delay, 10ms delay, 400ms delay) on accuracy and RT. In the second model, we directly assessed whether there was a difference between the 10ms and 400ms delay conditions with and without echo, respectively. For that, we added two sum-coded comparisons to the model, one comparing passively transmitted echo vs simple delay (where "echo present" was coded as -0.5, no echo (simple delay) as +0.5) within the small 10ms delay conditions, so that we could directly see whether the presence of a simulated echo produces a significant difference in response accuracy or

the formula testing for the echo:

$$synchrony.response.key \sim crt10ms + crt400ms + condition + block + (1|participant) + (1|sentence.id) \quad (3.6)$$

Table 3.3: Difference in accuracy over conditions in referent choice task

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
control condition	4.26	0.29	14.74	< 0.001	***
10ms simple delay	-0.06	0.35	-0.17	0.865	
400ms simple delay	-0.61	0.32	-1.93	0.054	.
10ms delay, echo	-0.35	0.33	-1.06	0.291	
400ms delay, echo	-0.56	0.32	-1.77	0.077	.
pos.second	-0.63	0.31	-2.00	0.045	*
trial number (z-scaled)	0.17	0.06	2.86	0.004	**
10ms simple delay:pos.second	-0.25	0.44	-0.57	0.569	
400ms simple delay:pos.second	0.06	0.40	0.15	0.878	
10ms delay, echo:pos.second	-0.05	0.42	-0.12	0.909	
400ms delay, echo:pos.second	0.25	0.41	0.62	0.534	

not. The other comparison was analyzing the same for the large asynchrony conditions, overall letting us judge SJ performance of the simulated echo over the small and large asynchronies.

As visible in Figure 3.9, most participants correctly identified the control condition as synchronous, and misidentified both conditions with small delay (no echo and echo present). As predicted, participants did not perceive any asynchrony at 10ms delay. More surprising is that for both large AV delay conditions, on average they performed at chance level. A closer inspection showed that some participants consistently fail to recognize the 400ms asynchrony, while others can correctly identify it as asynchronous. This is especially evident when looking at the mean response accuracy of the individual participants in the large delay conditions, showing a much higher variance than the other conditions.

Synchrony perception was not significantly changing over the timespan of the experiment, not providing evidence that the PSS shifted throughout performing the referent choice task as other studies reported in section 2. Both 400ms delay conditions significantly differed from the control condition, the 400ms simple delay condition($\beta = 3.46$, SE $= 0.34$, z $= 10.25$, p $= < 0.001$), as well as the large asynchrony with echo condition($\beta = 3.56$, SE

Table 3.4: Difference in accuracy in referent choice task with and without echo

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
(Intercept)	3.92	0.17	22.44	< 0.001	***
contrast10ms	0.26	0.31	0.83	0.406	
pos.second	-0.61	0.12	-4.94	< 0.001	***
contrast400ms	-0.05	0.31	-0.16	0.875	
trial number (z-scaled)	0.16	0.06	2.71	0.007	**
contrast10ms:pos.second	-0.17	0.39	-0.43	0.668	
pos.second:contrast400ms	-0.21	0.39	-0.54	0.586	

Table 3.5: Response keys in SJ task, synchrony question

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
control condition	-3.65	0.37	-9.83	< 0.001	***
10ms simple delay	0.22	0.40	0.55	0.584	
400ms simple delay	3.46	0.34	10.25	< 0.001	***
10ms delay, echo	0.57	0.38	1.50	0.135	
400ms delay, echo	3.56	0.34	10.64	< 0.001	***
block1	-0.01	0.20	-0.04	0.969	
block2	0.12	0.20	0.59	0.555	

= 0.34, z = 10.64, p = < 0.001). As predicted, the large delay conditions were easier to identify as asynchronous than the small conditions.

After applying the contrasts, we see that there is no evidence that accuracy of synchrony perception was affected by the presence of the echo in either delay condition. Seeing no significant effect here might indicate that participants did correctly regard the presence of an echo (distortion) and asynchrony (SJ question) as independent, as suggested by asking 2 separate questions targeting each phenomenon.

Response key in distortion question On the question whether participants perceived any distortion, like multiple signals within the audio component of the stimuli, a more

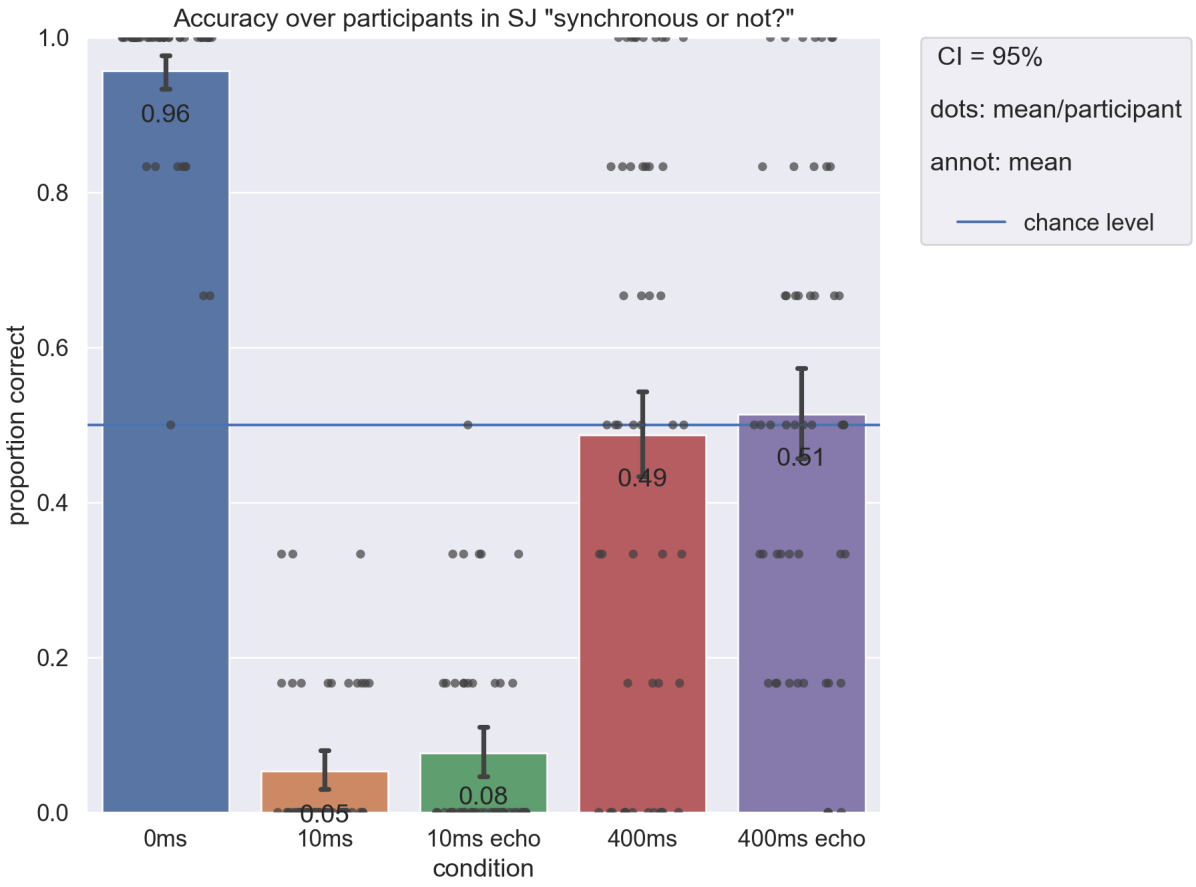


Figure 3.9: Response accuracy by condition for synchrony question

unified response pattern is present. The model applied here was identical to the one for the SJ question in section 3.3.2, just using data from the distortion question.[24] Analyzing the second question asked in the SJ task, whether or not the participants perceived any additional signals or distortions in the auditory stimulus components, we found that, as predicted, participants were almost unanimously not perceiving the additional attenuated echo in the small delay condition, but most were able to spot the presence of an echo in the large delay condition. For an overview, see Figure 3.10

All participants performed better than the chance threshold and correctly identified the presence of a passively transmitted signal with a large delay in the auditory part of

[24]The resulting formula tested was

$$distortion.keys \sim condition + block + (1|participant) + (1|sentence.id) \quad (3.7)$$

with contrasts:

$$distortion.keys \sim crt10ms + crt400ms + condition + block + (1|participant) + (1|sentence.id) \quad (3.8)$$

Table 3.6: Accuracy in SJ task, synchrony question with contrasts

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
(Intercept)	-1.43	0.19	-7.58	< 0.001	***
contrast10ms	0.03	0.23	0.12	0.908	
contrast400ms	-0.11	0.22	-0.50	0.616	
block1	0.004	0.16	0.03	0.980	
block2	0.10	0.16	0.60	0.550	

the stimulus. In contrast, almost all participants failed to respond to the presence of the passively transmitted additional signal when it was paired with a small AV delay.

Table 3.7: Responses in SJ task, "distorted or not?"

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
control condition	2.82	0.37	7.55	< 0.001	***
10ms simple delay	0.86	0.52	1.65	0.099	.
400ms simple delay	-0.46	0.46	-1.00	0.317	
10ms delay, echo	0.66	0.56	1.19	0.235	
400ms delay, echo	-5.58	0.41	-13.72	< 0.001	***
block1	0.74	0.31	2.40	0.017	*
block2	1.08	0.37	2.94	0.003	**

The echo in the large AV asynchrony condition was significantly better identified than the control(β = -5.58, SE = 0.41, z = -13.72, p = < 0.001). In contrast, the smaller echo seemed harder to detect and is only marginally significant(β = 0.86, SE = 0.52, z = 1.65, p = 0.099). Again, the performance difference between blocks is significantly increasing towards the end (difference between first and last block: β = 1.08, SE = 0.37, z = 2.94, p = 0.003), providing some indication that performance in the SJ task increased over time, possibly being influenced by performing the referent choice task and providing grounds to argue for perceptual adaptation towards sensory optimization for the task at hand.

In the contrast-coded comparison between passively transmitted echo present versus not present we see no significant impact on distortion recognition in the small asynchrony.

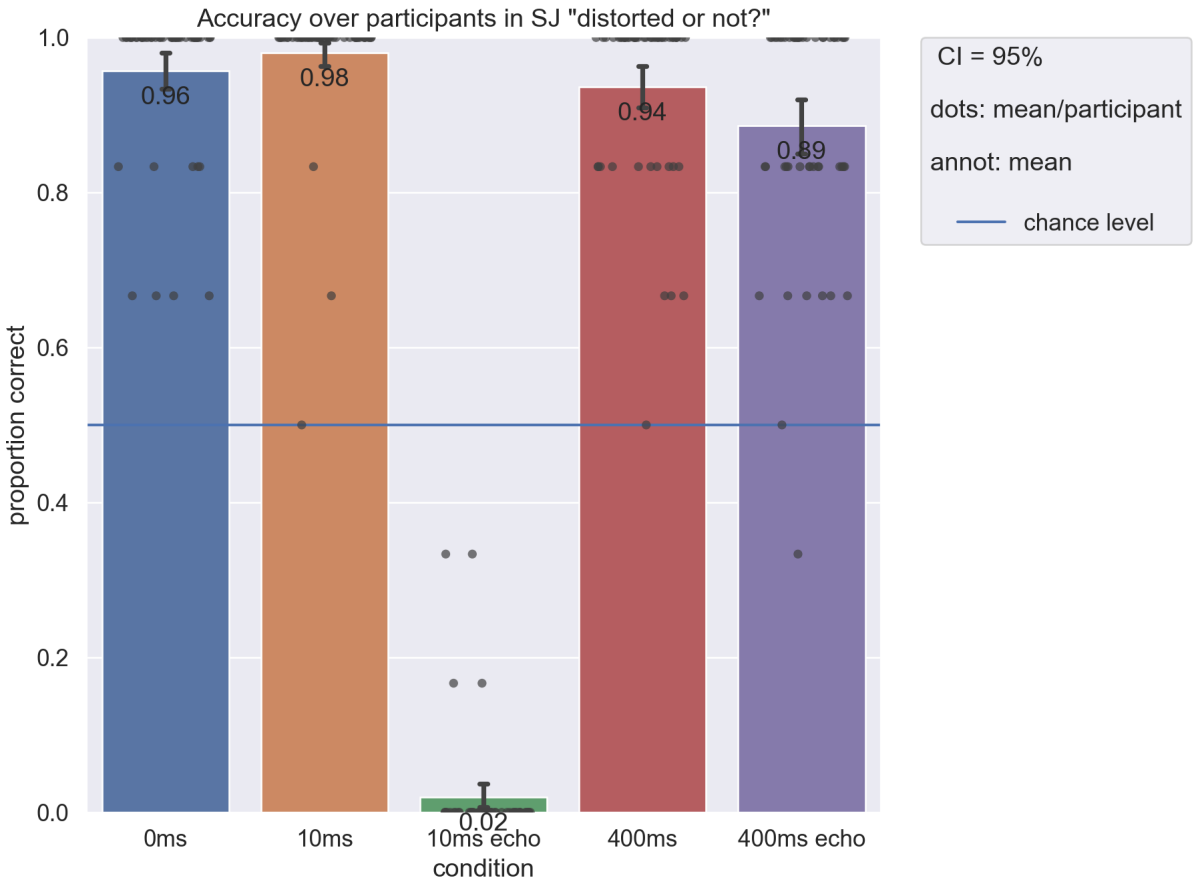


Figure 3.10: Response accuracy by condition for distortion question

Since we analyzed the response key here, this means participants noticed a distortion in the large asynchrony($\beta = 8.53$, SE $= 0.45$, z $= 18.83$, p $= < 0.001$), but failed to pick it up in the small one. This indicates that participants were able to correctly identify the echo in the 400ms AV asynchrony condition, but not in the 10ms one. We have therefore no evidence indicating that an attenuated echo as simulated here with an AV asynchrony of up to 10ms impacts speech recognition accuracy in our setup at all. For the larger asynchrony of 400ms, we can see a statistically significant difference of echo present versus no echo. We can infer that participants, as intended, identified the simulated echo as a distortion, indicating that they did indeed perceive it. For the full table see Table 3.8.

In short, we find that participants are in the present study unable to detect a 10ms audiovisual asynchrony. Also the additionally inserted passively attenuated echo remains undetected with such a small delay. We find no impact on RT and accuracy and with the present paradigm there is no grounds to argue that there is an impact on speech reception

Table 3.8: Responses in SJ task, "distorted or not?", impact of echo

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
(Intercept)	2.30	0.19	12.14	< 0.001	***
contrast10ms	< 0.01	0.33	0.00	1.000	
contrast400ms	8.53	0.45	18.83	< 0.001	***
block1	0.49	0.25	1.97	0.049	*
block2	0.66	0.25	2.59	0.010	**

performance at such delays. For a larger delay of 400ms, there is evidence that participants can notice such a delay, although a surprisingly high amount of participants seemingly was unable to consistently recognize it. We also find a significant difference in the large asynchrony conditions (400ms) between echo and simple delay condition, but not at 10ms delay.

4 Discussion

In the presented experiment, we used a referent identification task interleaved with an adapted SJ task demonstrating a method to measure the effective speed of speech perception via reaction time, potentially detecting adverse effects of sub-perceptible AV asynchrony. As shown through the SJ task, the small asynchrony of 10ms and also the simulated echo when paired with 10ms asynchrony were consistently below perceptive thresholds. We were unable to find any (adverse) effects on speech processing for this small 10ms range, neither in accuracy nor in the processing time of speech (operationalized through RT). This provides no evidence against 10ms asynchrony being an acceptable AV asynchrony used in SHPDs. With the data collected, we are unable to draw conclusions on whether an even higher asynchrony than 10ms might be acceptable to use as a permissible threshold as well, and some limitations apply. These will have to be addressed in further studies to validate the findings and the procedure. In the small 10ms asynchrony conditions of the referent choice as well as the SJ task, we found the null hypothesis to be true, meaning there is still the possibility that there are still adverse effects on speech understanding

present that we failed to identify or measure.

When looking at the larger 400ms delay in the referent choice task on the other hand, the AV delay was significant for the impact of the echo on reaction times of the participants, in contrast to the smaller delay. This validates the simulated echo as a relevant factor to be studied further. We were effectively able to demonstrate that at large enough AV delays, asynchronous duplicated auditory input has a consistently measurable effect on the speed of speech processing. Supposing that our simulation of the auditory stimuli was close enough to the real effects occurring when wearing an SHPD, we can infer that such impact is likely to be found in real-world usage too.

We also found rather high response accuracy variance[25] in the large 400ms asynchrony conditions in the SJ task. This would be an important finding if found to be valid, suggesting that there might not be an optimal one-size-fits-all solution. Instead, the success of the SHPD would not only depend on fit and individual attenuation but also individual size and shape of TWIN and suggest that permissible asynchrony thresholds are very individual indeed and might warrant diverse approaches tailored to the individual.

Specifically regarding usage of SHPDs, half of the participants noticing the asynchrony shows that 400ms is, at least for a significant part of the NH users, not a permissible delay and at such large delays other factors like the echo become distinctly noticeable. Since individuals with ASD typically have a slower adapting PSS as discussed in subsubsection 2.3.2, and therefore asynchrony presents a larger hindrance in speech understanding for them[26], it is likely that a potential working permissible threshold in noise attenuation for individuals with elevated sound sensitivity will have to be much smaller than 400ms. This also provides a first argument that further research into speech perception specifically in individuals with autism wearing SHPDs is warranted, since a systematic difference between NT and ASD population in permissible asynchrony thresholds cannot be excluded. Such SHPDs, for which the ARP 3.1 introduced in subsubsection 3.1.2 is one prototype, are contenders for alleviating and potentially solving communication deficits individuals with ASD are experiencing.

[25] as visible in Figure 3.9

[26] see Beker et al. (2018) and Turi et al. (2016) as reviewed in subsubsection 2.3.2

By and large, the designed task which measures speech perception by way of response speed and accuracy on a referent choice task was significant for both AV delay and echo on the 400ms conditions. This shows that it was a valid measure in principle and can measure the impact of phenomena specific to wearing an SHPD. In particular, we successfully measured the impact of multiple overlaid auditory signals produced by the still audible passively transmitted signal and the consequences of increased audio latencies introduced through real-time selective noise filtering. The success of the paradigm presented then shows that additionally introduced latencies, when small enough, and auditive distortions, such as the simulated echo, present no insurmountable obstacle to speech understanding with an SHPD.

The prediction that a large asynchrony of 400ms is readily perceivable by a majority of the participants does not fully hold in the present setup, and we will explore possible reasons for that and other potential issues in section 4.

In the following, I will discuss some limitations that apply to the experiment as conducted. I will examine potential technical issues regarding the upper bound of the TWIN in section 4, limits in detecting a potential sub-perceptible permissible asynchrony threshold discussed in section 4.

Upper bound issues Contrary to our expectations, an unexpectedly high percentage of participants (51% in the 400ms simple delay condition, 49% in 400ms echo condition, see Figure 3.9) was not able to detect the 400ms asynchrony, which is not in line with prior studies. A failure in normally hearing neurotypical population to detect asynchrony in the 400ms range in AV speech stimuli is not unheard of, but other studies present a much lower percentage of "synchronous" responses in unmodified SJ tasks. (Li et al., 2021) report "synchronous" responses closer to 15-25% at 400ms delay, as presented in section 2.2. Noel et al. (2017) report a proportion of "synchronous" responses closer to 40%(see section 2.2). As also mentioned in subsection 2.2, van Wassenhove et al. (2007) did only measure asynchrony as large as 267ms, but already at this smaller asynchrony they report a percentage of "synchronous" responses closer to 20%. While all of these values are lower than what we found, their large variance can be an indicator that potentially

not only procedural errors are responsible, but also the individual experimental conditions and differences in the used stimuli provide possible explanations.

Interindividual variance A look at our individual participants' data suggests that there are high individual differences with some participants consistently being able to correctly identify 400ms asynchrony and others consistently misidentifying the condition as synchronous. This would be in line with the findings by Ipser et al. (2017), who report stable individual differences in AV asynchrony perception through a McGurk effect. They suggest to not attribute this high variance between participants to measurement errors or biases and argue for a personal base asynchrony present in every perceiver, warranting finding individual asynchrony thresholds and correcting for them artificially. There are (at least) two other potential reasons for the high inter-participant variation in perceptibility of the large 400ms asynchrony: (1) a larger TWIN due to the used sentence material, and (2) limitations in the timing due to the online study format. For the former, the usage of whole sentences as stimuli in our setup would let us argue for a larger TWIN (Eg et al., 2015; Lezzoum et al., 2016), being translated into a stronger perceptual binding, and therefore resulting in the worse performance in the SJ task that we observed. We have also argued in section 2.2 that perceptual properties of the stimuli and the auditory environment, as well as background noise, can impact the shape and size of the TWIN, such that the strength of the perceptual binding varies with the stimuli used. In our experiment, the low frame rate and association biases added through non-random selection of the subset of sentences used, as well as the unknown perceptual fit of pictures and sentences are all factors that could impact the shape of the TWIN and therefore change RT and response accuracy. For the latter, the failure of the present experiment to reproduce these values could indicate problems stemming from the realization as an online study. Although all participants were informed to wear wired headphones and we assume that they did, the overall temporal accuracy of sound reproduction on participant devices can vary (Bridges et al., 2020). While wired headphones share more similarities in terms of reproduction speed than for example wireless headphones, where the latency variance is bigger, as we already established for hearing protection in section 1, the perception

of sounds transmitted depends on the individual fit (Lezzoum et al., 2016) and with headphones also on subjectively manipulated audio properties like audio gain (Stone et al., 2008) or presence of background noise (Lezzoum et al., 2016). All these factors play a role in shaping the TWIN and hence are deciding factors of whether an auditive and a visual component are perceptually bound or not, even at high asynchronies of 400ms. In our case, this could mean that since there is no unified environment between the participants, performance differences stemming from preexisting interindividual differences are additionally enlarged by interindividual differences in the experimental environment.

To mitigate the issue of an unexpectedly large proportion of participants still perceptually binding at 400ms asynchrony, we recommend using even larger values for the large asynchrony condition, paired with tested hardware in a lab-based replication to corroborate our findings. As a possible remedy, with more resources, the same experiment could be repeated with more variation in asynchrony, either introduced as additional conditions to maintain comparability or replacing the asynchrony values in the present experiment, to report more fine-grained results. More claims could also be made if the overall frame rate would allow for a more accurate relationship between visual stimulus perception and response time. To achieve this, a set of stimuli with a higher recording frame rate could be chosen. Experimental hardware should then be chosen accordingly, fulfilling the requirements to ensure a true representation of the intended stimuli. To mitigate other issues stemming from the uncontrolled stimuli and presentation parameters, conducting a norming study on the available stimuli or exclusively utilizing another set of prenormed audiovisual stimuli would be a good avenue of future research.

The present study would benefit from individualized prior calculation of the PSS, to account and correct for differences in synchrony perception. The difficulties to estimate the TWIN via the McGurk effect mentioned in subsubsection 2.1.3, are exemplary to show that more research is needed to robustly determine individual TWINs and use them to better account for perceptive differences of the participant as well as linking it to the properties of the stimuli.

Lower bound issues Turning to the smaller asynchronies measured to detect possible effects on speech perception, as stated, we did not find such effects, but the validity and extent of these findings are limited. One limitation of the online design is that it remains unclear whether the small AV asynchrony could technically be realized by the individual hardware combinations.

We do have information about the mean frame rate discussed earlier in section 3, but to make substantiated claims about which frames were dropped during actual playback and whether this drop had any impact on the measured effects, we would need a lab setting with defined and repeatable hardware. The actual mean frame rate used of 25fps means that all analyzed participants had a new image appearing every 40ms. While improving slightly through the exclusion of low average frame rates, this situation is not ideal for testing effects in the 10ms range, and it would be worthwhile to repeat the experiment with higher frame rates in a lab setting to determine its effects on speech processing. Eg et al. (2015) recommend maintaining a maximal asynchrony of less than 100ms to stay imperceptible to digitally stream AV content. It is advisable to be able to check in a future experiment whether this actually holds with the methods used in the present experiment.

Setup issues We lack information about the graphics rendering unit, the input devices, and the monitors used by the participants, whose model and implementation is a major factor in the speed of displaying each frame on the screen (Ivkovic et al., 2015; Bridges et al., 2020). Further, hardware compatibility and settings, like the mode the monitor is in or whether the browser supports WebGL, all impact the lag present in the intended and the real frame display (Bridges et al., 2020). Although the differences in display speed and other lags produced by hardware are in the low milliseconds' range(Ivkovic et al. (2015) found local lag ranging from 23 to 243ms), considering that we are trying to observe the effects of a 10ms AV delay, it would be worthwhile to repeat the experiment in a setting that can precisely account for the additional unintended delays in frame presentation. Moreover, as discussed in Bridges et al. (2020), measuring the actual delay in audio playback is even harder to control and would require measuring the physical sound with a microphone to ensure proper consistent playback. This is not a real option in any

remote online study. Here again, a reasonable avoidance of these issues is the repeated execution of the experiment under known external conditions in a lab setting. In principle, the experiment can be performed without changes offline, as the experimental software allows for that. There, additional hardware such as an external microphone could be used to verify the timing accuracy of auditory stimuli, as well as a consistent set of hardware, where visual display consistency could be checked using photodiodes (Bridges et al., 2020).

Methodological issues One possible methodological issue in the experimental setup is the validity of the chosen forced-choice referent identification task as an indication of speech processing. The intention here was that participants have to employ natural language understanding. While it is evident that participants have to comprehend the target adjective to attach it to one of the pictorial representations of the nouns, we cannot exclude the possibility that participants solved the task without direct reference to the sentence presented. One example would be through association (linking the adjective *green* to the *frog*) without taking the stimulus sentence into account.

Furthermore, there could be inaccuracies introduced through memory issues; Even though we display the target adjective for 2500ms, which is plenty of time to read, recognize, and memorize the target, feedback by participants indicates that it proved challenging to remember the targeted adjective throughout all 160 trial repetitions, resulting in some random answers in the referent choice task. Based on the relatively high overall accuracy, we do not expect this to be problematic in the present experiment.

Moreover, possibly problematic is the validity of the image representation of the target nouns. We took care to select prototypical images for all nouns referred to, but we cannot exclude speed differences in recognition and attribution to the correct noun. As stated earlier, all pictures taken from Duñabeitia et al. (2018) are tested for correct labeling by German-speaking participants, but this is not the case for images that were supplemented from other sources to create complete stimuli sets. These images were also not further analyzed for discernibility and perceptual properties like variance in luminance and contrast distribution. Regarding the image stimuli, the perceptual similarity between them is not tested for the configuration they are used in here together with the recorded videos of the

sentences (Rosemann and Thiel, 2018) from the OLACS corpus(Uslar et al., 2013) .

Possibly problematic is also the relative sparsity of data points for the SJ task, where we decided to only include 30 trials per participant purely for time considerations and taking care not to overly fatigue the participants. Looking at the individual subject level, from a statistical point of view, more data would give us more reliable results, especially when considering that the small asynchrony in the 10ms conditions is subject to various external variances. Here, the low sample size per participant prevents us from drawing conclusions and more rigorous research is needed.

The presented paradigms studying asynchrony in speech processing are reliant on operating variables and direct participant judgments, and speech processing as a specialized and highly adaptive mechanism is not easily studied directly. We hope to partially mitigate this problem by taking RT as the operating variable instead of synchronous/asynchronous responses. Rapid recalibration, represented through a shift in PSS, is essential to speech processing in real-world environments and promises to be an indicator of perceptual binding via predicting the size of the TWIN and vice versa.

Regarding an extension to subjects with ASD and potential communicative impairments previously discussed in subsubsection 2.3.2, it is problematic that the present experiment relies on the reading faculties of the participants. The presentation of the target adjectives could be realized as auditory cues or some other form of presentation could be constructed. In the present form, there is no direct extension of the experiment to non-reading participants possible without modification, limiting future direct group comparisons. We believe that a minor procedure modification and subsequent comparison still present worthwhile research, one could for example add spoken versions of the target adjectives to the already present written versions.

Other studies regarding selective noise attenuation with SHPDs After examining the issues of the present experiment and their potential solutions, we turn towards other examples already examining SHPDs as a potential solution for elevated sound sensitivity to see where our results are situated and can contribute. There are studies already examining technical solutions to increased sound sensitivity: Ikuta et al. (2016) for

example, tested active noise-canceling headphones in children with ASD. They describe a large potential and demonstrate effectiveness for coping with high sensory sensitivity, but were unable to test for long-term usage effects and lament that those noise-canceling headphones are usually designed to not attenuate voices, such that for application in a noisy classroom-type of situation they are useless.

Employing a different, but still technical strategy to solve the issues with noisy classrooms, Rance et al. (2017) tested a system with ear-level remote microphones and classroom amplification systems and found reduced cortisol levels in children with ASD when employing strategic sound amplification. They established a link between stress response levels and functional hearing impairment despite presenting with average auditory capacity, suggesting that reducing the stress elicited through stressful sounds alone could already help with speech perception in ASD.

There are also several commercial products available already targeting elevated sound sensitivity, not only in ASD but also in patients with tinnitus for example, who report similar sound sensitivities as individuals with ASD. One presents a manual version with a volume slider, such that the selectivity of attenuation remains problematic, but the wearer can easily adjust the attenuation level, addressing potential long-term usage effects: link to dbud. Another company uses an algorithmic approach employing an SHPD with selective speech in noise control; *link to nuheara*. The presence of these devices demonstrates that a device capable of attenuating specific sounds is worthwhile as a potential solution to auditory sensitivities and anecdotal testimonies of customers where usage improves quality of life indicate a potential solution coming from these technical approaches.

4.1 What did we show?

Reviewing the limitations coming along with our measured results and keeping in mind the large differences of thresholds measured in prior literature (see subsubsection 2.2.1, the evidence for an acceptable AV asynchrony threshold that provides freedom for many processing algorithms which are demanding in terms of introduced lag, is thin. Nevertheless, in the present experiment, we did produce some expected effects and the findings of the

SJ task are mostly in line with other findings in the literature, although it is reasonable to repeat the experiment with a predictable hardware combination in a lab environment to ensure consistency across participant results. If validated, we have shown that there exists an overlap of subperceptible asynchronies and asynchronies permissible for successful speech processing as introduced in section 2.2.1. We did reproduce some expected effects for the large 400ms delay conditions, as discussed in section 4, showing that with more individualized threshold selection, synchrony perception can also be studied online, potentially providing research scaling options to more participants with less work. We were able to establish a valid protocol to test for sub-perceptual temporal AV asynchrony in speech perception.

Another novelty is the effective demonstration of the simulated echo effect introduced in section 2.2.1, we were able to observe a significant difference in larger asynchronies, showing that effects specifically pertaining to the situation of wearing an SHPD are observable and deliver novel knowledge on phenomena influencing speech perception under asynchrony.

4.2 Outlook

Overall, results from the present study mirror general results and expectations from the literature, we can clearly see high individual differences in speech perception, and, if repeatable in a lab setting, likely presenting also in large differences in TWIN and strength of perceptual binding.

Since sensory hypersensitivity is associated with elevated adverse responses to certain sounds (see section 4.2), research into the algorithms handling the selection part of selective attenuation would benefit individuals suffering from elevated sensitivity to only very certain sounds. Also, it would take argumentative force away from criticism targeting over-attenuation with HPDs and its adverse long-term behavioral effects.

As open areas of research also remain other demographics. A repetition of the experiment with still-developing children and other age groups promises insights into the temporal unfolding of asynchrony detection in speech. Concerning the intended application in SHPD for individuals with ASD, a repetition and comparison with the target population

would be of special interest. As mentioned in section 4.2, SHPDs are only one avenue of research into ameliorating negative social life experiences of individuals suffering from decreased sound tolerance. Therefore, therapeutic effects of a combination of measures such as using SHPD in conjunction with cognitive-behavioral therapies should be a target of future research.

Should our findings be found reliable, this would have implications beyond the concrete usage of selective sound attenuation in ASD. The high response accuracy variance in the adapted SJ task in larger 400ms asynchronies, if repeatable, would be an argument in favor of individually adapted selective attenuation algorithms. This adaptivity is further complicated by the need to be adaptive to the individual and different environmental situations and usage scenarios as seen in section 2. Further research also testing environmental dependence of speech processing could further illuminate this relationship and also more lifelike sentences with varying complexity and syntactic structure could be used to further approximate actual usage in the wild.

ASD As discussed in subsubsection 2.3.2, atypical multisensory integration is a major reason for individuals with ASD to take longer to develop linguistic skills (Beker et al., 2018) during childhood, in some severe cases even remaining completely nonlinguistic, with prominent features of ASD being difficulties in speech and social interaction (American Psychiatric Association, 2013). There is some evidence that in ASD, a slower shift of the PSS and therefore a less adaptive TWIN is responsible for part of the deficits in language processing, (see Stevenson et al. (2014)) and Stevenson et al. (2018) provides experimental evidence that there is a link between atypical multisensory processing in ASD and impaired communication skills.

Regarding the decreased sound tolerance that is a prominent symptom in ASD, as discussed in section 4.2, Pfeiffer et al. (2019) provide support that noise-attenuating headphones, both in-ear and over-ear, are effective to reduce sympathetic responses of the nervous system to sound in individuals with decreased sound tolerance. While there is protest against long-term usage of HPD[27], SHPDs have the potential to deliver

[27]Jüris et al. (2014) suggest that sustained wear of HPD for individuals with ASD and decreased sound

the reduction in stress and sympathetic response, while avoiding issues that a blanket attenuation of all sounds carries. While behavioral measurements like cognitive behavioral therapy are shown to be effective here, we believe that the potential additional option of a selectively sound attenuating device is worthwhile to support the effectiveness of these and possibly providing more quality of life through reduced discomfort for neurodivergent individuals with decreased sound tolerance.

The present experiment serves as a first step towards research in ASD by providing initial results from a NH neurotypical population such that factors specific to a potential research population with ASD can be easily identified in a comparison. The implementation as an online study may help in future data collection as it can be done from almost anywhere, enabling participants to conduct the experiment in a safe and non-distracting environment. The results further provide initial cautionary evidence that a practical application for the presented purpose is possible. Speculatively, because we are unable to make definite claims about the significance regarding the null hypothesis, there is at least some engineering headroom for additional latencies caused by noise selection algorithms, that is we were unable to detect evidence against this null hypothesis.

5 Conclusion

We conducted an online study examining the impact of potentially increased audiovisual asynchronies present in the wearing of an SHPD, simulated through adding artificial lag to AV speech stimuli and also simulating an attenuated secondary auditory signal with a transfer function using real-world data recorded on an SHPD.

The prediction that participants do not show any significant speed differences or differences in accuracy when presented with a sub-perceptible asynchrony holds for 10ms, with a perceptibility threshold or a threshold for adverse effects on speech understanding yet to be found through further research. The significant impact on RT through the presence of an attenuated echo in the larger 400ms asynchrony and the overall effects of

tolerance is counterproductive, as it may lead to increased anxiety levels, they put forward cognitive behavior therapy instead.

asynchrony behaving largely as predicted further show that the paradigm can detect an impact on the speed of speech processing. The presence of the echo impacting speech processing in large delays (400ms) but not small delays (10ms) is demonstrating that the paradigm works as intended and detects effect differences tied to specific asynchrony ranges. Carefully repeated with controlling the limitations of our setup in a setup with less variance, the procedure could gain valuable insights on real-world problems creators and wearers of such an SHPD would face when applied with respect to increased sensory sensitivity in ASD.

More research is needed to issue concrete recommendations on permissible delay thresholds for algorithms, but the current results present no evidence against two important issues, namely higher latency and the resulting multiple asynchronous signals, likely not presenting insurmountable barriers towards a usage targeting speech understanding. The findings on the echo also serve to show that research concentrating on effects of asynchrony alone is not sufficient when considering possible effects of SHPDs, phenomena unique to the application at hand are worth considering.

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.

American Psychiatric Association (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.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4.

Beker, S., Foxe, J. J., and Molholm, S. (2018). Ripe for solution: Delayed development of multisensory processing in autism and its remediation. *Neuroscience & Biobehavioral Reviews*, 84:182–192.

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.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

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.

Haas, H. (1972). The influence of a single echo on the audibility of speech. *Journal of the audio engineering society*, 20(2):146–159.

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, twin. *International Journal of Audiology*, 48(6):321–333.

Ikuta, N., Iwanaga, R., Tokunaga, A., Nakane, H., Tanaka, K., and Tanaka, G. (2016). Effectiveness of earmuffs and noise-cancelling headphones for coping with hyper-reactivity to auditory stimuli in children with autism spectrum disorder: A preliminary study. *Hong Kong Journal of Occupational Therapy*, 28(1):24–32. PMID: 30186064.

Ipser, A., Agolli, V., Bajraktari, A., Al-Alawi, F., Djaafara, N., and Freeman, E. D. (2017). Sight and sound persistently out of synch: stable individual differences in audiovisual synchronisation revealed by implicit measures of lip-voice integration. *Scientific Reports*, 7(1):1–12.

Ivkovic, Z., Stavness, I., Gutwin, C., and Sutcliffe, S. (2015). *Quantifying and Mitigating the Negative Effects of Local Latencies on Aiming in 3D Shooter Games*, pages 135–144. Association for Computing Machinery, New York, NY, USA.

Jüris, L., Andersson, G., Larsen, H. C., and Ekselius, L. (2014). Cognitive behaviour therapy for hyperacusis: A randomized controlled trial. *Behaviour Research and Therapy*, 54:30–37.

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.

Kuiper, M. W., Verhoeven, E. W., and Geurts, H. M. (2019). Stop making noise! auditory sensitivity in adults with an autism spectrum disorder diagnosis: Physiological habituation and subjective detection thresholds. *Journal of autism and developmental disorders*, 49(5):2116–2128.

Kuo, S. M., Lee, B. H., and Tian, W. (2013). *Real-time digital signal processing: fundamentals, implementations and applications*. John Wiley & Sons.

Lezzoum, N., Gagnon, G., and Voix, J. (2014). Voice activity detection system for smart earphones. *IEEE Transactions on Consumer Electronics*, 60(4):737–744.

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.

Neave-DiToro, D., Fuse, A., and Bergen, M. (2021). Knowledge and awareness of ear protection devices for sound sensitivity by individuals with autism spectrum disorders. *Language, Speech, and Hearing Services in Schools*, 52(1):409–425.

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.

Pfeiffer, B., Stein Duker, L., Murphy, A., and Shui, C. (2019). Effectiveness of noise-attenuating headphones on physiological responses for children with autism spectrum disorders. *Frontiers in Integrative Neuroscience*, 13:65.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rance, G., Chisari, D., Saunders, K., and Rault, J.-L. (2017). Reducing listening-related stress in school-aged children with autism spectrum disorder. *Journal of autism and developmental disorders*, 47(7):2010–2022.

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.

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.

Smith, E. G. and Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, 48(8):813–821.

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.

Stevenson, R. A., Segers, M., Ncube, B. L., Black, K. R., Bebko, J. M., Ferber, S., and Barense, M. D. (2018). The cascading influence of multisensory processing on speech perception in autism. *Autism*, 22(5):609–624.

Stevenson, R. A., Zemtsov, R. K., and Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions., twin. *Journal of Experimental Psychology: Human Perception and Performance*, 38(6):1517.

Stiegler, L. N. and Davis, R. (2010). Understanding sound sensitivity in individuals with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, 25(2):67–75.

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.

Stone, M. A., Moore, B. C., Meisenbacher, K., and Derleth, R. P. (2008). Tolerable hearing aid delays. v. estimation of limits for open canal fittings. *Ear and Hearing*, 29(4):601–617.

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 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, twin. *Neuropsychologia*, 45(3):598–607. Advances in Multisensory Processes.

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

Vatakis, A. and Spence, C. (2006b). Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neuroscience Letters*, 405(1):132–136.

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

Younkin, A. C. and Corriveau, P. J. (2008). Determining the amount of audio-video synchronization errors perceptible to the average end-user. *IEEE Transactions on Broadcasting*, 54(3):623–627.

yu Zhou, H., Cheung, E. F., and Chan, R. C. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions, twin. *Neuropsychologia*, 140:107396.

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). Audiovisual temporal order judgments. *Experimental brain research*, 152(2):198–210.

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

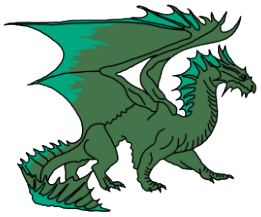
(A.1.6) Bär (bear), Source: Duñabeitia et al. (2018)

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

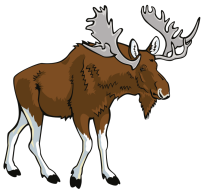
(A.1.8) Clown (clown), Source: Duñabeitia et al. (2018)

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

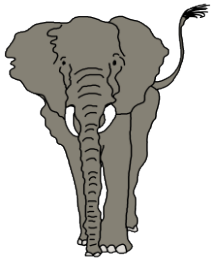
(A.1.10) Dieb (thief), Source: Duñabeitia et al. (2018)



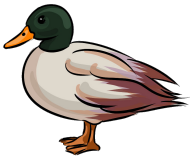
(A.1.11) Drache (dragon), Source: Duñabeitia et al. (2018)



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

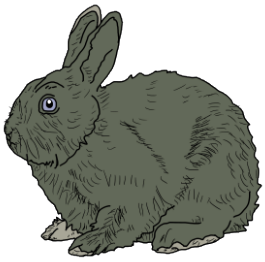
(A.1.15) Esel (donkey), Source: Duñabeitia et al. (2018)

(A.1.16) Frisör (hairdresser), Source: Duñabeitia et al. (2018)

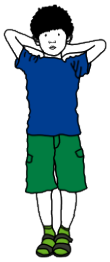
(A.1.17) Frosch (frog), Source: Duñabeitia et al. (2018)

(A.1.18) Gespenst (ghost), Source: Duñabeitia et al. (2018)

(A.1.19) Gärtner (gardener), Source: Duñabeitia et al. (2018)



(A.1.20) Hase (hare), Source: Duñabeitia et al. (2018)



(A.1.21) Junge (boy), Source: Duñabeitia et al. (2018)

(A.1.22) Jäger (hunter), Source: Duñabeitia et al. (2018)

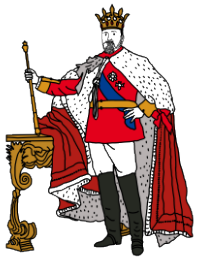
(A.1.23) Kapitän (captain), Source: Duñabeitia et al. (2018)

(A.1.24) Kasper (punch), Source: Duñabeitia et al. (2018)

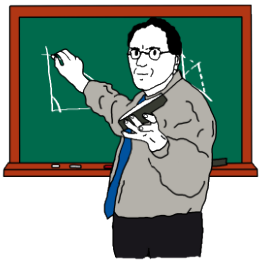
(A.1.25) Koala (koala), Source: Duñabeitia et al. (2018)

(A.1.26) Kobold (cobold), Source: https://de.cleanpng.com/png-85z9z8/download-png.html

(A.1.27) Koch (Chef), Source: https://pngtree.com/so/chef-hat-clipart



(A.1.28) König (King), Source: Duñabeitia et al. (2018)



(A.1.29) Lehrer (Teacher), Source: Duñabeitia et al. (2018)

(A.1.30) Löwe (Lion), Source: Duñabeitia et al. (2018)

(A.1.31) Maler (Painter), Source: Duñabeitia et al. (2018)

(A.1.32) Mann (man), Source: Duñabeitia et al. (2018)

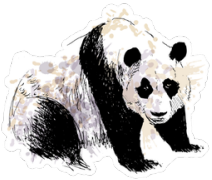
(A.1.33) Matrose (sailor), Source: https://www.nicepng.com/ourpic/u2w7q8a9q8w7w7u2_navy-drawing-sailor-line-art-soldier-navy-drawing/

(A.1.34) Maulwurf (mole), Source: https://www.pngwing.com/de/free-png-tuucp

(A.1.35) Metzger (butcher), Source: Duñabeitia et al. (2018)

(A.1.36) Mönch (monk), Source: https://de.cleanpng.com/png-zsofj1/download-png.html

(A.1.37) Nikolaus (Nikolaus), Source: https://nikolaus-von-myra.de/de/darstellung/galerie/



(A.1.38) Panda (panda), Source: Duñabeitia et al. (2018)

(A.1.39) Papagei (parrot), Source: Duñabeitia et al. (2018)

(A.1.40) Papst (pope), Source: Duñabeitia et al. (2018)

(A.1.41) Pfarrer (pastor), Source: Duñabeitia et al. (2018)

(A.1.42) Pilot (pilot), Source: Duñabeitia et al. (2018)

(A.1.43) Pinguin (Penguin), Source: Duñabeitia et al. (2018)

(A.1.44) Jäger (hunter), Source: Duñabeitia et al. (2018)

(A.1.45) Polizist (Policeman), Source: Duñabeitia et al. (2018)

(A.1.46) Postbote (mailman), Source: Duñabeitia et al. (2018)

(A.1.47) Prinz (prince), Source:https: //www.pngegg.com/de/png-iptbx

(A.1.48) Punker (punk), Source: https://www.vippng.com/preview/hRbTRJR_punk-tribu-urbana-vestimenta-punks/

(A.1.49) Radfahrer (biker), Source: https: //de.pngtree.com/so/die-jungs

(A.1.50) Riese (Giant), Source: https://www.pinterest.de/pin/766386061570400292

(A.1.51) Ritter (Knight), Source: Duñabeitia et al. (2018)

(A.1.52) Roboter (robot), Source: Duñabeitia et al. (2018)

(A.1.53) Räuber (bandit), Source: Duñabeitia et al. (2018)

(A.1.54) Soldat (soldier), Source: Duñabeitia et al. (2018)

(A.1.55) Tiger (tiger), Source: Duñabeitia et al. (2018)

(A.1.56) Tourist (tourist), Source: Duñabeitia et al. (2018)

(A.1.57) Vater (father), Source: https://de.cleanpng.com/png-hu6n8o/download-png.html

(A.1.58) Wikinger (viking), Source:https://de.cleanpng.com/png-en6r2o/

(A.1.59) Zauberer (wizard), Source:Duñabeitia et al. (2018)

(A.1.60) Zwerg (dwarf), Source: Duñabeitia et al. (2018)

A.1.2 Sentences

Here is a full list of all sentences taken from the OLACS corpus and appearing in the experiment(Uslar et al., 2013).

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, examples of all screens shown to the participants in chronological order are listed.

Figure A.2: Screens presented in online experiment

Willkommen zum Experiment!

Sie sollten bereits alle relevanten Fragen beantwortet, verkabelte Kopfhörer eingestellt und ca. 40 Minuten Zeit mitgebracht haben. Am Ende werden wir Ihnen noch ein paar freiwillige Fragen zum Experiment selbst stellen.
Viel Erfolg.

[Fortfahren mit Enter]

(A.2.1) Welcome Screen

In diesem Experiment werden Sie in jedem Durchgang zunächst ein Adjektiv auf dem Bildschirm sehen, welches eine Eigenschaft wiedergibt, (z.B. "tapfer").
Bitte merken Sie sich dann diese Eigenschaft.

[Fortfahren mit Enter]

(A.2.2) Introduction

Im Anschluss erscheinen dann zwei Bilder von Tieren oder Personen auf der linken und rechten Seite des Bildschirms.
Sie haben kurz Zeit, um sich diese Bilder anzuschauen, bevor Ihnen ein Videoclip gezeigt wird.
In diesem sagt ein Mann einen deutschen Satz über die angezeigten Bilder.

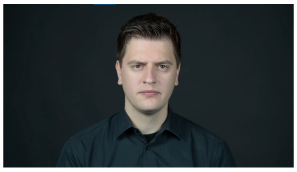
[Weiter mit Enter]

(A.2.3) Explanation of the target

Ihre Aufgabe ist es, dem Satz zu entnehmen, welche der beiden Personen oder Tiere die zuvor gezeigte Eigenschaft (z.B. "tapfer") besitzt, und dann schnellstmöglichst zu antworten.
Nutzen Sie die linke Pfeiltaste für das linke Bild,
und die rechte Pfeiltaste für das Rechte.

[Weiter mit Enter]

(A.2.4) Task Description



Hier sehen Sie ein Beispiel für ein unverändertes Video. Bitte stellen Sie Ihre Computerlautstärke so ein, dass Sie den Satz klar und deutlich verstehen können.

[Wiederholen mit Leertaste]
[Fortfahren mit Entertaste]

(A.2.5) Sound Adjustment Screen

Im folgenden Block sollen Sie beantworten, ob in den Clips, die Ihnen gezeigt werden, Audio und Video synchron sind.

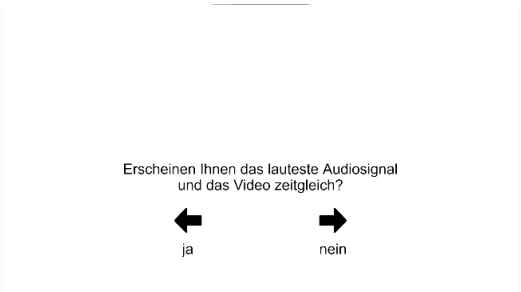
Sie sollen außerdem bewerten, ob Sie eine Audiovererzerrung, wie zum Beispiel ein doppeltes Signal, hören.

Sie können nur JA oder NEIN antworten.

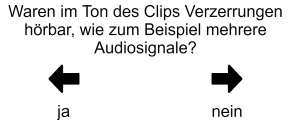
Die linke Pfeiltaste entspricht JA, die rechte Pfeiltaste NEIN.

[Fortfahren mit Enter]

(A.2.6) Introduction SJ Task

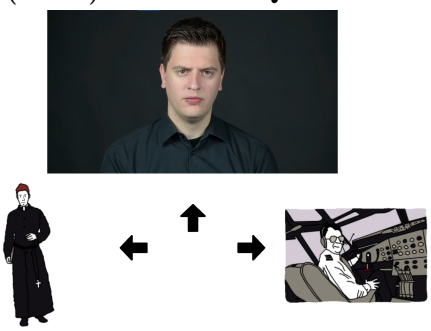


(A.2.7) Synchrony Question



(A.2.8) Distortion Question

(A.2.9) Target Presentation



(A.2.10) Stimulus Presentation

A.3 Extended experiment

Table A.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

A.4 Extended results

Table A.2: Difference in RT over conditions, referent choice task

Measure	Estimate (β)	Std. Error (SE)	z-value (z)	p-value (p)	
control condition	44.05	0.58	76.01	< 0.001	***
10ms simple delay	-0.19	0.25	-0.75	0.454	
400ms simple delay	4.15	0.25	16.36	< 0.001	***
10ms delay, echo	0.08	0.25	0.32	0.751	
400ms delay, echo	2.86	0.25	11.26	< 0.001	***
pos.second	10.05	0.25	39.74	< 0.001	***
trial number (z-scaled)	-1.15	0.06	-20.09	< 0.001	***
10ms simple delay:pos.second	0.28	0.36	0.78	0.437	
400ms simple delay:pos.second	-0.81	0.36	-2.26	0.024	*
10ms delay, echo:pos.second	0.15	0.36	0.41	0.682	
400ms delay, echo:pos.second	0.69	0.36	1.91	0.056	.

A.5 Semantic network of the cited literature

Network of all cited literature visualized in a semantic network

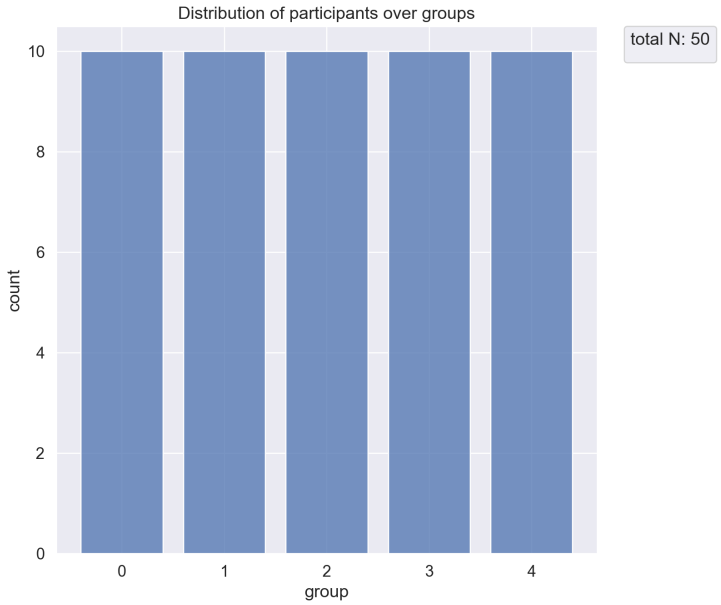


Figure A.3: Distribution over participant groups after exclusion

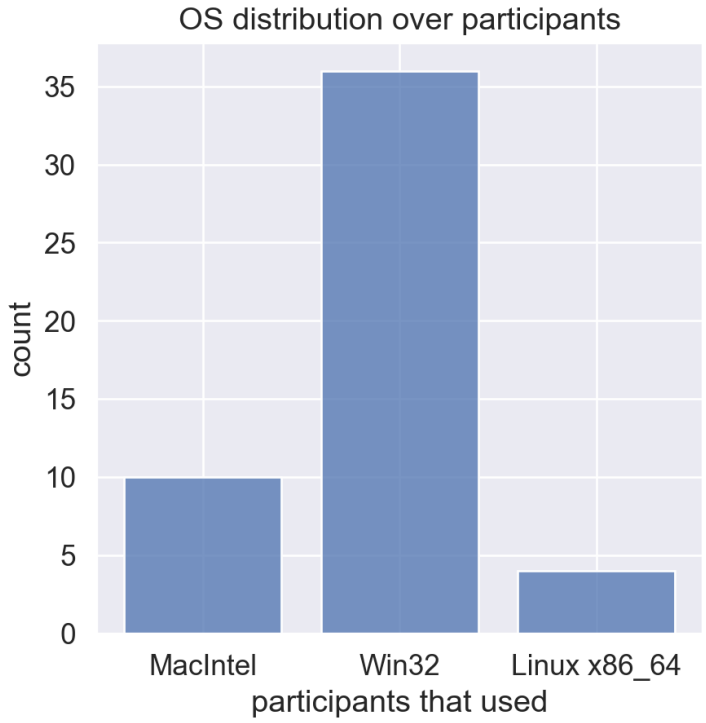


Figure A.4: Operating systems used by participants

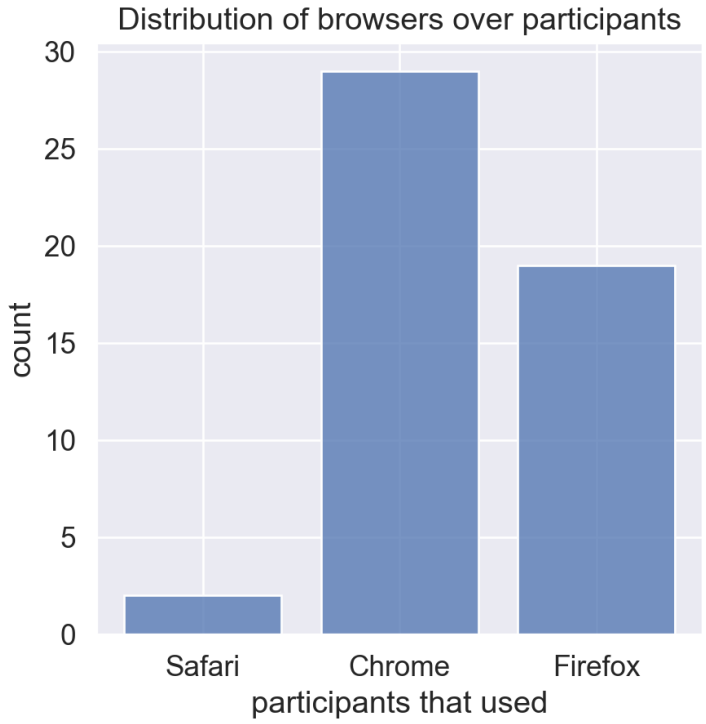


Figure A.5: Browsers used by participants

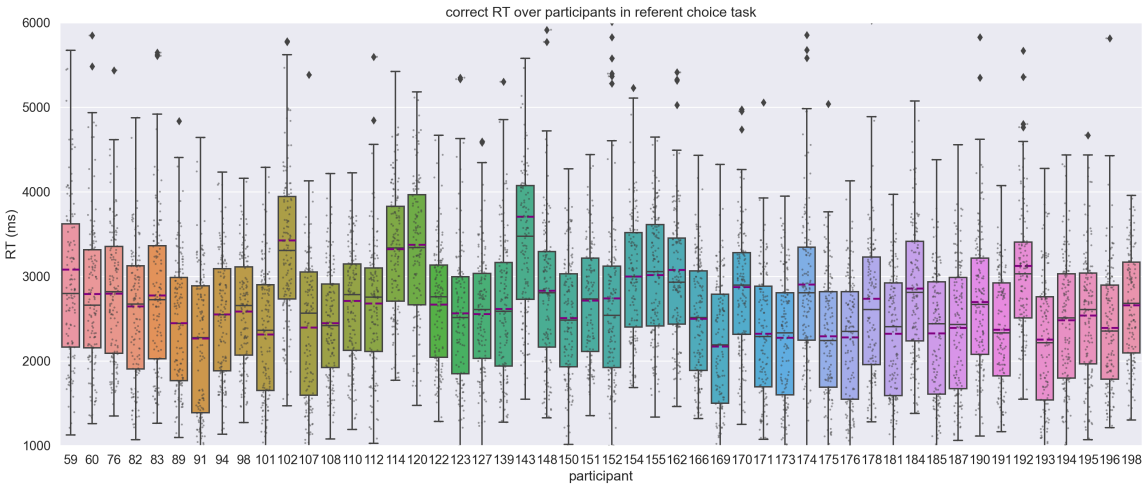


Figure A.6: Overall RT for correct responses in referent choice task

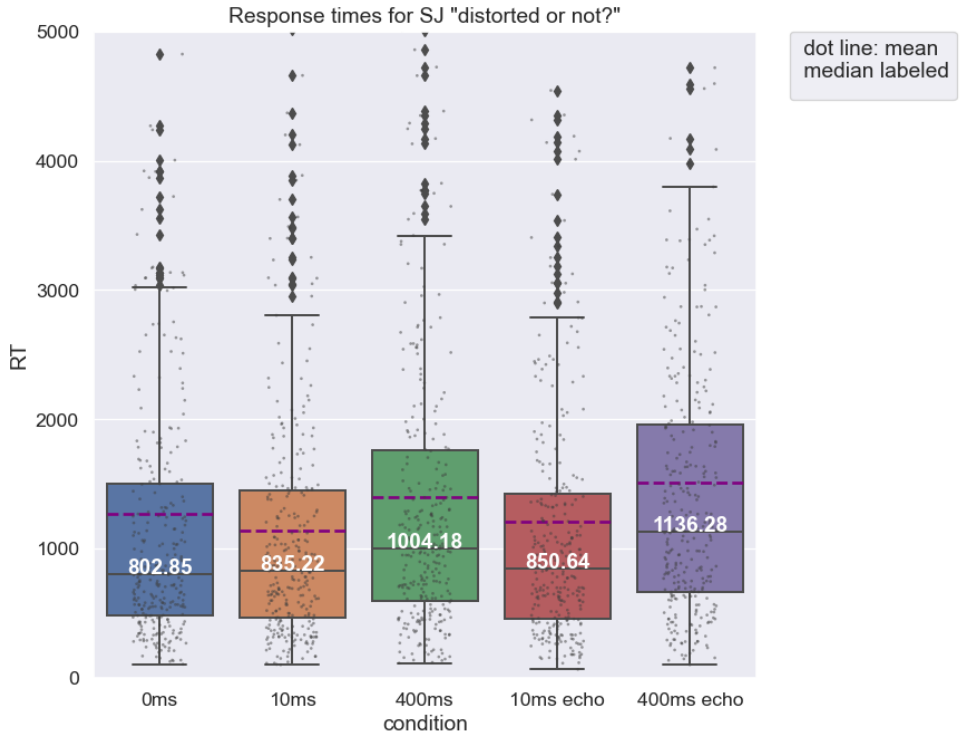


Figure A.7: RT for correct responses in SJ task over conditions for "distorted or not?"

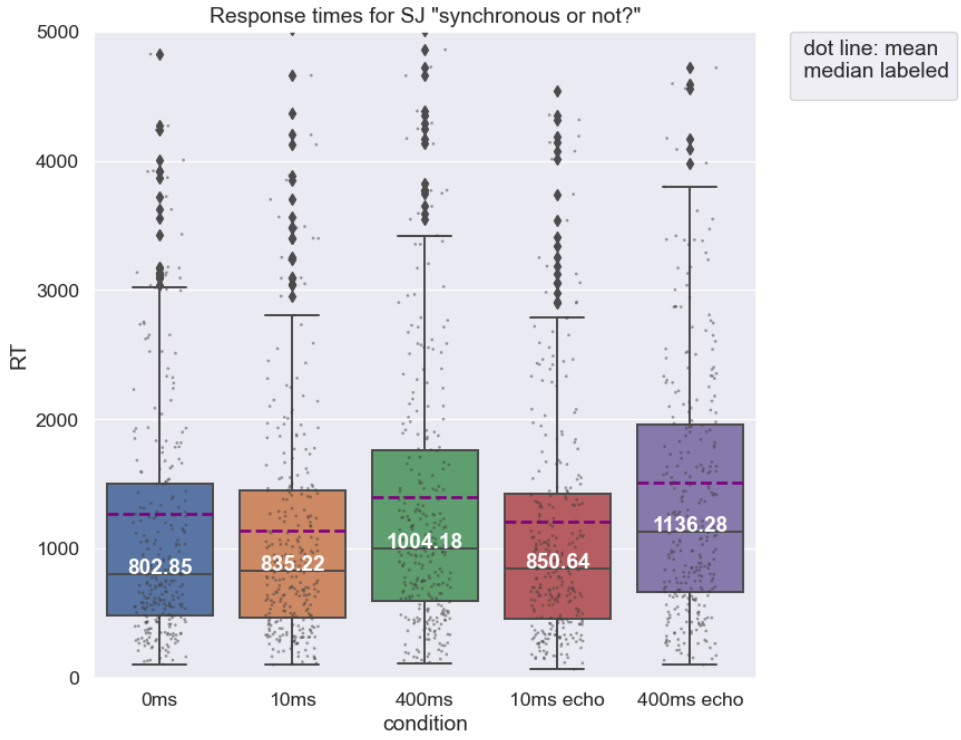


Figure A.8: RT for correct responses in SJ task over conditions for "synchronous or not?"

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, the 24.05.2021

city, date

signature

Acronyms

AAC advanced audio coding. 28

ASD Autism Spectrum Disorder. I, 1, 2, 4, 6, 7, 18–20, 47, 53–58

AV audiovisual. 1–5, 8–24, 27, 28, 33, 38, 41, 44–49, 51, 54, 55, 57

DAF delayed auditory feedback. 9, 14–16

DSP digital signal processor. 2, 20

EEG electroencephalography. 8, 9

FFT fast Fourier transform. 28

fMRI functional magnetic resonance imaging. 7, 17

fps frames per second. 26

HI hearing impaired. 5, 6, 14, 16, 18

HPD hearing protection device. 2–4, 20, 55, 56

JND just noticeable difference. 14, 15, 22, 25

MSI multisensory integration. 19

NH normally hearing. 5, 7, 15, 16, 18, 47, 48, 57

NRR noise reduction rating. 4

NT neurotypical. 1, 2, 5, 6, 47, 48

OLACS Oldenburg linguistically and audiologically controlled sentences. 26, 32, 53

OVS object-verb-subject. 26, 29

PAR personal attenuation rating. 4

PSS point of subjective simultaneity. 12, 17, 22, 41, 47, 50, 53, 56

RT reaction time. 5, 10, 23, 24, 30, 32–37, 39, 40, 45–47, 49, 53, 57, 75

SD standard deviation. 12

SHPD smart hearing protection device. I, 2–5, 7, 9, 20–24, 27, 28, 46–48, 53–58

SJ simultaneity judgment. 11–15, 17–19, 21–23, 30, 32–34, 39, 41–44, 46–49, 53, 55, 56

SNR signal-to-noise ratio. 8

SOA subjective onset asynchrony. 11, 13, 14

SRT speech reception threshold. 26

SVO subject-verb-object. 26

TD typically developed. 6, 18, 19

TOJ temporal order judgment. 13–15, 17, 22

TWIN temporal window of integration. 3, 6, 11–15, 20–23, 34, 47–50, 53, 55, 56