Factors influencing Audiovisual Speech Integration

Liesbeth Gijbels

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2024

Reading Committee:
Adrian KC Lee, Chair
Kaylah Lalonde
Yi Shen
Mark T. Wallace

Program Authorized to Offer Degree:
Speech and Hearing Sciences

©Copyright 2024

Liesbeth Gijbels

University of Washington

Abstract

Factors influencing Audiovisual Speech Integration

Liesbeth Gijbels

Chair of the Supervisory Committee:
Adrian KC Lee
Speech and Hearing Sciences

Audiovisual (AV) integration, identified as a pivotal factor in comprehending speech in noisy environments, is a complex phenomenon. Understanding speech perception, even within a single modality, presents various nuances due to language specificity. When consolidating information from multiple modalities, it is imperative to understand how the listener processes the speech signals in each modality, and how this information is successfully integrated to benefit our speech understanding. Prelinguistic integration mechanisms, such as synchronous temporal information from both modalities, have a significant role in identifying AV events. Yet, linguistic integration mechanisms, like phoneme-viseme connections of the AV speech signal or individual linguistic knowledge, significantly influence speech intelligibility. This dissertation contains two sections. First, we outline three remote

AV speech perception tasks across developmental stages, and in developmental disorders such as developmental dyslexia (ages 4-15; n = 261). Second, we present a series of four remote AV psychophysical tasks in adults, ages 21-40 (n = 46), to elucidate the role of prelinguistic and linguistic features pertinent to AV speech integration. For the developmental work we find that weighting assigned to the auditory modality in the AV speech signal serves as a better explanation for individual variability across development than age itself. Moreover, atypical weighting of auditory modality explains differences between children with and without developmental dyslexia on a group level. In adults, our findings suggest that how well temporal asynchrony between the auditory and visual signal is tolerated depends not only on the linguistic complexity of the stimulus, but also on the individual. Prelinguistic information like temporal synchrony perception has an important role in AV speech perception as it endows a 3 dB increase in perceived loudness perception of the target speaker, but this increment interacts with linguistic complexity and temporal asynchrony. Together, these results offer novel insights into different factors influencing AV speech integration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible with the constant support of a number of people.

First and foremost, I extend my heartfelt gratitude to my advisor, Adrian KC Lee. Not only did he play an instrumental role in shaping these projects, but he also shaped my entire PhD journey into a truly enjoyable experience. His unwavering support, both personally and professionally, exceeded my expectations. Our shared enthusiasm, drive to progress, and emphasis on maintaining a healthy work-life balance created a stress-free environment, leaving room for me to grow personally and as a scientist.

Second, I would like to thank Jason Yeatman and Patricia Kuhl for their pivotal roles in guiding my entry into these doctoral studies. Moreover, during my PhD journey I received constant encouragement, both scientifically and financially. Working with both of you on diverse side projects over the past few years felt as a nice change of scenery to my day to day PhD experience.

A big thanks to the Kuhl lab members, specifically Julia Mizrahi and Bo Woo, who became close friends throughout the years. And to my 'only' lab mate, Drew J. Here's to more volleyball games in the future!

I also want to thank my committee members for their consistent support in the pursuit of this degree: Kaylah Lalonde, Yi Shen, Mark Wallace and Scott Murray.

None of these projects would have been possible without the willingness to share scientific efforts. I specifically want to thank Rachael Holt, Kaylah Lalonde, Yi Shen, Kristin Van Engen, Antoine Shahin, Sarah Al-Salim and Dawna Lewis for sharing their stimuli recordings with me.

Assessing and scoring tasks is laborious and intensive work. Therefore, a special thanks to Piper Doering, who made key contributions to the data collection of the developmental studies.

Furthermore, I want to thank other current and previous PhD students of the Speech and Hearing Department at UW. A specific thanks to Patrick Donnelly, Elle O'Brien, William Gray, DJ Audet, Elise Lebovidge, Erica Peterson, Amanda Ciana, Dale Summers and Bertan Kursun for all the great talks.

As I spent six months of this 3.5 year PhD trajectory at Meta, I would also like to thank Melinda Anderson, Christi Miller, and Khia Johnson to help me grow into a better scientist.

And a specific shout out to my support team in Belgium:

Hannah, wat ben ik dankbaar voor een vriendin als jij. Wat een unieke ervaring, om ondanks onze atypische leeftijd, toch samen - van op een hele afstand - dit traject parallel te kunnen doorlopen, en dan nog eens 'samen' mama te kunnen worden!

Christa en Philipe, jullie wil ik toch ook specifiek bedanken voor alle steun en alle interesse in dit traject en al mijn werk.

Mama en papa, ik ben wie ik ben door jullie. Jullie oneindige steun en trots, van ver of dichtbij, maken elke dag een beetje beter, en maken elk obstakel een beetje makkelijker om te trotseren. Ik ben ervan overtuigd dat ik dit doctoraat op deze manier tot een succesvol einde heb kunnen brengen door alles wat jullie mij hebben meegegeven door de jaren heen.

En de laatste dank is gereserveerd voor Hans, en Vic. Hans, in 2018 eindigde je het dankwoord van jouw doctoraat met het statement "Dat wij samen nog veel zouden winnen!". En ik denk dat je toen nog geen idee had van wat wij nog allemaal gingen trotseren samen in de komende vijf jaar. Dank je voor wie je bent, om mij te helpen streven naar meer en beter, en om steeds samen vooruit te willen. En vooral, bedankt om mij zo een prachtige zoon te geven tijdens dit heel doctoraatsproces.

Chapter 1

Introduction

1.1. General introduction

A dog barking on a porch, a movie playing in a theater, having a conversation in a busy restaurant, all of these distinct events have something in common. They send out information in a variety of forms, like light and sound, triggering our sensory systems (i.e., vision and audition) to convert this energy into meaningful signals in our nervous system, and coordinate these distinct sensory systems to collaboratively understand coherent, multisensory events in our everyday lives (Lee & Wallace, 2019). One of the most prevalent and relevant social experiences for humans – engaging in face-to-face conversations – is inherently multimodal. For this reason, this work will focus on audiovisual (AV) speech perception, and henceforth the discussion of multisensory processing in this dissertation pertains to both audition and vision, specifically referring to AV perception, unless otherwise noted.

In the context of AV speech perception, the visual cues from the speaker's face and articulators play a crucial role in enhancing our comprehension of the incoming auditory speech signal. This significance is underscored by the fact that more than half of conversational speech events (61.8%) occur in background noise (Walden et al., 2004), highlighting the importance of visual cues in conversations. Multisensory perception is central to both speech perception and production (Venezia et al., 2016). Multisensory integration can alter perception (McGurk & MacDonald, 1976), and often benefits our performance (e.g., Binnie et al., 1974; Erber 969; Macleod & Summerfield, 1987; Sumby & Pollack, 1954), leading to improved accuracy (Tye-Murray et al., 2007a), processing speed (Reisberg et al., 1987), speech comprehension (Tye-Murray et al., 2008), and listening effort (Sommers & Phelps, 2016).

However, integration of audiovisual (AV) speech cues reflects substantial individual differences, which cannot be entirely accounted for by the information conveyed solely through the speech signal or the perceptual abilities of the individual (Grant et al., 1998). These observed individual differences illustrate (1) changes in response to experience with auditory and visual sensory processing across the lifespan (Baum & Stevenson, 2017; Lalonde & Werner, 2021; Tye-Murray et al., 2016), and (2) individual variability in AV speech integration within a phase of life (Dey & Sommers, 2015; Grant et al., 1998; Nath et al., 2011). Integration of AV speech plays an important role in language development for infants and children (e.g., Bahrik & Lickliter, 2012; Dodd & Burnham, 1988; Kuhl & Meltzoff, 1982; Lewkowicz & Kraebel, 2004; Patterson & Werker, 1999), in second language learning (Erdener, 2016 for review), in understanding speech in adverse listening environments (e.g., Bradlow & Alexander, 2007; Ma et al., 2009; Van Engen et al., 2014), or when perceptual systems are impaired (e.g., Dias et al., 2021; Setti et al., 2013). Moreover, AV integration has been proposed as a contributing factor in defining the etiology of developmental disorders like autism spectrum disorder (e.g., Foxe et al., 2015; Iarocci et al., 2010; Megnin et al., 2012), developmental language disorders (e.g., Meronen et al., 2013; Norrix et al., 2007), or developmental dyslexia (e.g., de Gelder & Vroomen, 1998; Hayes et al., 2003; Ramirez & Mann., 2005).

Since 1954 (Sumby & Pollack, 1954), an extensive body of literature has studied AV speech perception. Various models have been formulated to elucidate AV integration. Each model centers on specific aspects, such as speech stimulus characteristics (Tye-Murray et al., 2007b), individual traits (Grant et al., 1998), optimal integration (Ma et al., 2009), and the weighting of auditory and visual input based on stimulus characteristics (Massaro, 1987) or individual perceptual performance (Schwartz, 2010). In contrast to most auditory speech-perception models (i.e., Liberman et al., 1967; Studdert-Kennedy, 1976) that take basic components (e.g., bottom-up signal related cue-extraction, integration of featural

information, and top-down linguistic processes) into account, AV models run into the additional challenge of incorporating and evaluating visual cues both within and across modalities (Grant & Bernstein, 2019; Massaro, 1987). To better reconcile with communicative abilities in real-world scenarios or other related behavioral performances, AV speech perception models need to consider prelinguistic and linguistic characteristics of the signal, its context, and the individual. First, at the level of the speech signal, both the clarity of the input signal (i.e., prelinguistic) and the linguistic information provided by the speech signal (i.e., linguistic) affect AV speech integration. Second, large individual variability in integration of AV speech signals point to the need for inclusion of linguistic knowledge of the individual as well as basic perceptual abilities.

In summary, AV speech perception is complex, and it is important to use consistent language to describe different aspects of this process. Furthermore, in the AV literature, there are terminologies that have not been clearly delineated. The next section focuses on establishing a framework that relates AV speech processing to AV perception in general by first clarifying a few terminologies that have been used routinely in the literature (Section 1.2). Subsequently, AV speech perception abilities will be discussed from infancy to older adults, considering both prelinguistic and linguistic perspectives (Section 1.3 - 1.4).

1.2. A framework for AV speech integration

The first section (1.2.1) focuses on how information can be combined in a multisensory context. We will first focus on the concept of AV integration. Next, we will introduce different aspects of speech perception and focus on how interpretation of AV speech perception could be described from a prelinguistic and linguistic viewpoint (Section 1.2.2). We then explain the process in which we decide whether sensory information should be combined in the first place (i.e., Causal Inference, see Section 1.2.3) and how this view from an ideal Bayesian observer

perspective can serve as a framework for AV integration of speech accounting for prelinguistic and linguistic components.

1.2.1. General mechanisms of AV integration

Cars passing, dogs barking, construction workers banging their hammers, rain falling, children jumping in puddles – all these distinct events can happen in one scene, right in front of you. These events undoubtedly contain sounds and visuals that uniquely go together; intuitively, it is not a car barking or raindrops banging. Yet, the auditory and visual information of all these events in this specific scene arrive separately at our auditory (i.e., cochlea) and visual (i.e., retina) systems as continuous and overlapping input. Each of the sensory systems therefore needs to appropriately group information into meaningful streams. In audition, this process is called auditory scene analysis (Bregman, 1994), whereas in vision, this is often referred to as image segmentation (Pasupathy, 2015). This grouping of information leads to object formation, and these objects can be formed because they share particular properties that group together different sensory content coming from a common source (Bizley & Cohen, 2013). The characteristics of an object are often low-level features (e.g., auditory: pitch, temporal onset/offset, temporal envelope, timbre; visual: color, radiance, luminance), and the coherent modulation of these features makes the object stand out in the background stream (Griffiths & Warren, 2004; Shamma et al., 2011).

For example, from our scenario where multiple dogs are barking, the auditory system could use interaural time difference (ITD) and interaural level difference (ILD) cues to determine that there are multiple dogs (i.e., segregate objects), and the perceived location of these dogs could then be computed through grouping together their component ITD / ILD values. The visual system can similarly tell us that there are multiple dogs and where they are located. However, when our scene becomes more complex – for example, three dogs are barking while fighting over a toy – the

use of auditory ITD and ILD cues might not suffice to distinguish how many dogs are present, and as they move quickly, where they are located. Allowing both the auditory and visual system to process and combine incoming information will enhance our localization performance. Specifically, by combining auditory with visual information, these single multisensory objects enable us to track each dog separately. Following the features of one object (e.g., temporal envelope of one's bark) simultaneously aids in tracking the location of the same dog.

These AV events are thus *"perceptual constructs which occur when a constellation of stimulus features are bound within the brain"* (Bizley et al., 2016), and are often tracked and perceived so seamlessly that it is easy to forget that the sensory information reaches to different parts of the cortex (via the cochlea and retina), originating from separate physiological coding processes. Nonetheless, at some point this information is integrated, and further optimized by contextual information. When, where, and how exactly this integration process occurs is a more complex question.

It is plausible that both modalities are processed individually first, culminating in combined multimodal output (i.e., late-stage integration – Altieri, 2010; Grant et al., 1998). Such late-stage integration could occur from consecutive processing of both information streams. However, the idea of serial processing has never gained much interest as it seems an unlikely and inefficient way of AV perception (Altieri & Townsend, 2011). The visual stimulus is rarely completed before the start of the auditory stimulus and delaying processing of the auditory stream would be cognitively demanding and inefficient. It is more likely that the auditory and visual streams are being processed parallel in time and a separate decision of each modality leads to the identification of the AV speech signal. This does not necessarily mean that they have to be processed exactly simultaneously, but there is room for overlap (Altieri & Townsend, 2011). In this late-stage integration hypothesis, AV integration takes place only after the completion of unimodal processing (Grant et al., 1998) and thus this integration process constitutes a distinct cognitive stage. This concept

accommodates (1) individual variability in the AV integration process, not explained by unimodal processing, and (2) individual weighting of auditory and visual information based on the informativeness of each modality (Peelle & Sommers, 2015). The posterior Superior Temporal Sulcus most likely plays a major role in this high-level late-stage integration process (Möttonen et al., 2004) as it receives input from both modalities (Calvert et al., 2000; Stevenson & James, 2009). In addition these late-stage integration processes allow the use of context and cognition in decision making. If and how the context is interpreted along with the sensory data depends on prior knowledge, expectations, and conceptual understanding of each individual.

In early-stage integration, however, multisensory processing is hypothesized to occur as early as in the primary sensory cortices. Perceptual information from each modality can influence one another (Schroeder et al., 2008), and is combined into a common representation (Altieri, 2010; Calvert et al., 2000). The formed object created by early integration can either be described as modality-specific or as amodal. Modality-specific processing suggests there is a dominant modality that can convert information from other modalities (Massaro, 1987; Woodhouse et al., 2009). For example, looking at a singer's mouth (i.e., visual information) is still represented by an auditory concept, the pitch of the singer. Amodal processing, however, suggests that there is a non-modality-specific percept that combines information perceived by both modalities but is blind to the modalities itself (Rosenblum, 2005). An example of this would be that AV speech containing both information from the visual modality (i.e., visemes) and the auditory modality (i.e., phonemes) are amodally stored in the brain as phonological concepts. Whether early-stage integration is modality-specific or amodal is another topic of discussion, but we postulate that although information of the AV event is transmitted via two different paths, for early-stage integration, information from both sensory systems influence each other from the start. This idea is corroborated by neurophysiological measures showing

that primary sensory cortices are sensitive to other modalities. For example, visual signals activate the auditory cortex (Calvert et al., 1997; Pekkola et al., 2005) and auditory signals activate the visual cortex (Schepers et al., 2015). Furthermore, AV signals increase activity in the primary and secondary auditory cortex, relative to auditory-only signals (Okada et al., 2013). The unimodal streams potentially prepare the 'other' primary cortex for incoming information; more specifically, unimodal cues put the primary cortex in a state of high excitability for when input arrives (Calvert, 2001).

Yet, early- and late-stage integration are not mutually exclusive concepts. Peelle and Sommers (2015) classify these early- and late-stage integration effects as prediction and constraint. Early-stage integration mechanisms shape the processing of future sensory inputs (i.e., prediction) and late-stage integration mechanisms aid combining AV information by imposing restrictions on perceptual interpretations (i.e., constraint). Audiovisual information affects our perception in both of these ways as (1) visual features increase the precision of the listeners' predictions about the upcoming auditory signal (Gagnepain et al., 2012), and (2) at a later stage of integration (a) both modalities can be weighed based on their reliability (Körding et al. 2007, Ma et al., 2009) and (b) prior knowledge and experiences limit the number of possibilities (Körding et al. 2007; Sohoglu et al., 2012). Therefore, AV integration is most likely a multistage process, a combination of early-stage and late-stage integration. From this multistage perspective, the information processing of both modalities in each primary sensory cortex could be explained as multisensory facilitation (Schroeder et al., 2008). When presented with visual cues, synchronized firings of groups of neurons in the primary auditory cortex occur. The neural oscillations in the auditory cortex created by the visual cues shift the phase in a way that when the auditory input arrives the phase will be in a high state of excitability. This process of phase-resetting, in turn, leads to amplification of properties of the auditory signal that results in overall AV enhancement of the incoming signal. The Superior Temporal Sulcus receives converging excitatory multisensory input that interacts. The integrative output can then be normalized by surrounding neuron

responses (referred to as divisive normalization), and therefore can transcend the level of single neuron responses. Divisive normalization also accounts for adaptive changes based on cue reliability (Morgan et al., 2008). For example, if there is a second stimulus that differs on a certain dimension (e.g., timing, location, semantics), it will suppress the excitatory response of the first stimulus and therefore influence the normalization signal of the surrounding neurons. The occurrence of phase-resetting and divisive normalization in this process are likely to complement each other (van Atteveldt et al., 2014). Divisive normalization analyzes the content, where phase-resetting sets the context. They overlap in outcome, yet are relevant in different brain regions (primary cortex vs Superior Temporal Sulcus), at different temporal scales, and operation modes (van Atteveldt et al., 2014).

Now we can tie the earlier described concept of an AV object to the multistage processing in AV integration. To unpack this further, we first go back to our scenario of three dogs fighting over a toy. At a specific time (and place), we would like to selectively attend to one dog (over other dogs, and the rest of the environment). This selective attention therefore is specifically linked to the target AV object (i.e., the target dog), and is generally referred to as object-based attention (Bizley et al., 2016). Selectively attending our target dog will lead to enhancement of features, specific to that dog. For instance, by using auditory ITD and ILD information in combination with visual information of the moving dog, object-based attention will allow the pitch of this specific dog's bark to stand out.

The coherence between the features of different modalities are crucial in this process of selective attention to promote AV object formation, and this has been tested in the laboratory. For example, when the amplitude of a disk is modulated in a temporally coherent way to the intensity of the target tone, but not the masker tone, other features of this target tone (e.g. pitch) – that are orthogonal to the temporally matched AV features – are also enhanced, leading to more accurately completing pitch perception tasks for the target tone but not for the masker. However, when the changes in disk amplitude temporally match the masker tone, not only is the pitch of

the target tone not enhanced, performance on the pitch task is actually worse than when there was no alignment between the auditory and visual information, showing that coherence within an AV object is crucial to AV integration (Maddox et al., 2015).

The process of binding coherent AV information into an AV object has been postulated as an early-stage integration mechanism, as enhancement of all features of this AV object can occur without attentional modulation (i.e., in anesthetized animals; Atilgan et al., 2018), and without needing judgment of the listener on the enhanced features (Bizley et al., 2006). Yet, not all AV integration is a result of object formation. AV integration can be defined as *"any process in which information across sensory modalities is combined to make a perceptual judgment"* (Lee et al., 2019). Let us clarify this with an example. Think about watching a movie with subtitles. The coherent auditory and visual information of a conversation between two actors allows us to benefit from both auditory and articulatory information of the specific actor that is speaking in a busy setting. This allows us to selectively attend two distinct AV objects (i.e, the two actors in the conversation). The visual information, provided by the articulators, can improve the understanding of the auditory speech signal, yet might not be sufficient. When we then also read the subtitles we can improve the intelligibility of the scene. However, these subtitles are not uniquely temporally coherent to one of both actors. So subtitles do not result in binding of the AV input signal, yet the visual information provided by the subtitles can still be used to our benefit. This is then an example of late-stage AV integration.

1.2.2. Audiovisual speech perception

Coherent information provided by incoming AV speech signals thus aids in identifying specific AV events. It is established that both temporal and spatial information of the AV signal play a role in guiding the listener when and where to attend (e.g., temporal: Lalonde & Werner 2021; Munhall & Vatikiotis-Bateson, 2004;

spatial: Möttönen et al., 2011; Rosenblum, 2008), however, whether this type of information is sufficient to bind AV speech signals is less clear.

In contrast to non-speech stimuli (e.g., Maddox et al., 2015), temporal synchrony might not be sufficient to identify AV events for speech. As AV temporal coherence of speech can be defined as "*the low-level physical relationship between the timecourses of mouth movements and sound production*" (Cappelloni et al., 2023), we automatically acknowledge the role of phonetic content (i.e., how the sound matches the movement of the articulators) in this process (Lalonde & McCreery, 2020). To address the contributions of temporal and phonetic content in AV integration of speech, Fiscella and colleagues (2022) measured performance on an orthogonal feature (i.e., fundamental frequency modulation) in AV speech stimuli that varied in perceptual temporal and phonetic coherence (i.e., by varying rotation of the face in the video) and showed that temporal coherence, but not phonetic coherence resulted in improvement of the pitch perception task. Interestingly, Cappelloni and colleagues (2023) tried to further define features important for binding of AV speech signals and found that temporal coherence had no effect on the detection of orthogonal pitch events when the identity of the talker was varied. In their study, the identity of the talker seemed to be the only contributing factor for binding of AV speech signals. They suggested that these fundamentally different outcomes were a result of the complexity of the AV scene that needed to be segregated. When segregation of the scene was easier (i.e., defined by talker identity), synchronous temporal information did not add information and may therefore have had a smaller role in AV object formation.

Does this mean that rather than having a set of AV features resulting in object formation of AV speech signals, the complexity of the speech scene drives the reliance on perceptual cues like temporal coherence? Whether or not this is true, rather than focusing on the delineation of AV binding in the integration of AV speech, a more practical approach to deal with the complexity of AV speech is

describing AV integration of speech in complex scenes by (1) prelinguistic and (2) linguistic factors. This is not a completely new idea. Peelle and Sommers (2015) and Fiscella and colleagues (2022) suggest that temporal and linguistic correspondences of AV speech differentially contribute to early- and late-stage integration. Peelle and Sommers (2015) more specifically suggested distinguishing between timing of the acoustic signal (i.e., amplitude envelope, influencing attention and perceptual sensitivity) and the content (i.e., place and manner of articulation, constraining lexical selection). Given recent developments in literature using more artificial and synthetic stimuli (e.g., Cappelloni et al., 2023), we felt the need to expand temporal information to prelinguistic information, and specifically delineate the importance of language and all its components in the analysis of speech. This further allows for discussion of the input signal and its context, as well as factoring in the specific experiences of the listener with this input (Vatakis et al., 2008) in describing AV integration in speech.

We define prelinguistic as the perceptual information that is independent of the language of the stimuli and context, or the language and language skills of the talker or listener. Both the auditory and the visual modality provide low-level temporal (i.e., signal onset and offset) and spatial information (i.e., about the location), and in the context of speech, this will inform us when and where to direct attention, and it will help identify the speaker (Bernstein et al., 2004; Fiscella et al., 2022; Peelle & Sommers, 2015). Furthermore, dynamic patterns that are similar in auditory and visual speech help segment a continuous speech stream into distinct meaningful units (Cunillera et al., 2010; Kuhl, 2004) and identify the amplitude envelope of speech (Chandrasekaran et al., 2009). And more recently there even has been growing evidence that articulatory features like tongue-back position, mouth opening, and intra-oral pressure, which are speech specific but not language specific, are also visible in subtle jaw, lip, and cheek movements (Kroos, 2007). Prelinguistic elements describe the temporal and spatial characteristics of the signal (and masker),

but also the perceptual sensitivity of the listener (Peelle & Sommers, 2015). Individual differences in perceptual processing skills create variances across subjects in AV speech integration.

In contrast, linguistic information encompasses every aspect related to language. At the most basic level, both auditory and visual speech signals contain speech-specific cues like phoneme-viseme relationships. Phonemes are minimally distinctive auditory units of speech, whereas visemes are minimally distinctive visual units of speech (Fisher, 1968). Visemes represent the position of the face and mouth when producing a phoneme. A viseme, however, can represent multiple phonemes as they might have identical appearances on the lips (Bear & Harvey, 2017). Manner, place, and voicing of articulation are all represented via the auditory modality, whereas visual speech mainly contains information about the place of articulation (Tye-Murray et al., 2007b). Thus, the position of the talker's articulators improves speech perception linguistically, as this phoneme-viseme relationship helps with selecting the talker, and improves the quality (Tye-Murray et al., 2007a; Eskelund et al., 2011) and the speed (Reisberg et al., 1987) of speech understanding. Each language has a set of visemes that correspond to their specific phonemes, and therefore phoneme-viseme connections are unique to a language (and an individual).

More complex levels of linguistic information that have a role in AV speech integration are (1) the vocabulary used in the speech signal and (2) the complexity of the speech stream (word vs. sentence vs. passages). How much this linguistic information influences speech perception is a result of the individual's lexicon, the knowledge of semantics and syntax, the use of word-level and sentence-level context to compensate for misperceptions and impoverished acoustic and visual information, and memory processes and strategies for lexical access based on partial information in the signal (Grant et al., 1998; Ma et al., 2009).

In the context of multistage processing of AV speech stimuli, prelinguistic information can conceivably be processed at both early and late integration stages, whereas linguistic information is solely processed at a later stage, as it is hard to argue that linguistic information (both auditory and visual) already exists and is processed concurrently in the sensory cortices. As described by Grant and colleagues (1998): *"lexical identity is determined, which can then feed back to auditory and visual processing separately"*. This idea is further corroborated by neurophysiological findings. Atilgan and colleagues (2018) argue that the primary auditory cortex encodes low-level spectral features (i.e., temporal coherence), whereas articulatory features are only present later on at the Superior Temporal Gyrus (Ding et al., 2016). Any feedback connection from the Superior Temporal Gyrus to the primary auditory cortex would then be late-stage integration. This view is further supported by Bayesian Integration in the higher brain areas like the Intraparietal Sulcus (Rohe & Noppeney, 2015 (a+b); 2016).

1.2.3. Causal Inference as an ideal Bayesian observer in the context of prelinguistic and linguistic features of AV integration

1.2.3.1. Prelinguistic features

Prelinguistic AV mechanisms like temporal and spatial information processing of the auditory and visual speech stream are important as they help detect sensory information (Eramudugolla et al., 2011). Temporal information plays a role in predicting the timing of acoustic speech events, and therefore provides information on when to listen (e.g., Bernstein et al., 2004; Grant & Seitz, 2000), and may help target speech to be separated from other competing speech sounds (Devergie et al., 2011; Summerfield, 1992). In our everyday lives our auditory system is often confronted with a complex mixture of sounds. It is a challenge for the auditory system to segregate and group these sounds into their component sources, so we can ignore some while focusing on others. Specifically, the additional dynamic

information provided by visual cues in AV events can influence this process of auditory scene analysis (Atilgan et al., 2018; Maddox et al., 2015).

There is, however, a more fundamental question that needs to be addressed: How do we know what should be integrated, and what should not? The Causal Inference Model is a theoretical framework that can provide us with some behavioral (Körding et al., 2007; Ma et al., 2009) and neurophysiological (Cao et al., 2019) insights.

In an ideal, noise-free world there is a 100% match/mismatch between the two modalities, making it rather simple to decide whether to integrate the input streams. When two events are close in space and time, a single underlying cause is assumed; however, when these two events are distinct in space and time, two independent causes are inferred (Hairston et al., 2003; Wallace et al., 2004). Unfortunately, the real world is not that simple. For example, in AV speech there is no objective temporal synchrony. Visual information precedes auditory information in conversational contexts. At the level of the talker, the timing of the mouth movements precedes the onset of the voice by tens to a few 100 ms (Chandrasekaran et al., 2009; van Wassenhove et al., 2005). At the level of the listener, visual information also reaches the listener sooner than auditory information due to differences in the speed of acoustic versus electromagnetic waves. While we are tolerant of visual-leading asynchrony up to ~ 250ms (regardless of the language used – Munhall et al., 1996; Stevenson et al., 2012), it creates a level of uncertainty. Think about a movie where the audio and video do not match. It takes a second to realize what is going on, and why this looks weird. This is because we expect the audio and video to be in sync, but our system first has to figure out, should we integrate or not? This uncertainty could be estimated by a reliability-weighting principle of judging sensory signals as an optimal Bayesian observer (Alais & Burr, 2004; Ernst & Bülthoff, 2004; Ma et al., 2009 Raposo et al., 2012). In essence, this means that for small disparities there is a higher likelihood of integrating, whereas for greater disparities chances increase that a person does not combine this multimodal information. The model combines cues

from both modalities in a statistically optimal manner and adjusts the cue combination continuously, depending on the degree of belief that the sensory information from different modalities should be combined together (Körding et al., 2007). Thus, this model does not only suggest whether specific information in two modalities should be combined, it also tells us how they should be combined by specifying the weight in the integration process (Ganesh et al., 2018).

For every incoming multisensory signal the Causal Inference Model evaluates four parameters: (1) the uncertainty of the visual system (i.e., visual likelihood), (2) the uncertainty about the auditory system (i.e., auditory likelihood), (3) the knowledge the observer has about the cues (i.e., interaction prior), and (4) the prior probability that there is a single cause versus two causes (Körding et al., 2007). For real-life speech events this implies that when evaluating the uncertainty of the individual sensory systems, one has to look at the characteristics of the environment (e.g., visual cues are less reliable in the dark, auditory cues are less reliable in noisy backgrounds) and the individual observers (e.g., vision or hearing difficulties). Even in an ideal scenario, intrinsic physiological noise exists throughout the nervous system (Faisal et al., 2008), and thus no sensory stimulus reveals the true sensory value but rather induces a distribution of probabilities for the sensory percept.

In order to understand how priors play a role in this causal inference model, let us examine the ventriloquist effect (Pick et al., 1969). The ventriloquist effect describes an illusionary phenomenon in which you perceive speech produced by the artist as coming from the mouth of a puppet. Our prior knowledge biases us in such a way that we believe that speech sounds should go together with congruently moving articulators. The ventriloquist's mouth is hardly moving, but the puppet's mouth is (and often in a more exaggerated way to capture your attention). Therefore, we would naturally assume the sound comes from the moving visual source, the puppet. Furthermore, and beyond AV illusions like the ventriloquist effect, each feature has its own priors. Audition and vision are not weighed equally in

localization or temporal decisions. More specifically, vision is weighed more heavily for localization, whereas audition is weighed more heavily in temporal decisions. Accumulated experience gives confidence that visual cues are more reliable for spatial localization. These priors are in agreement with the observers' uncertainty about the world's causal structure as accounted for by Bayesian Causal Inference (Körding et al., 2007; Meijer et al., 2019). For example, an individual's unimodal perception is on average biased towards the center when localizing visual stimuli and to the periphery when localizing auditory stimuli. When integration of the two modalities occurs, the visual bias becomes dominant (i.e., the prior), but will still be reduced compared to the unimodal visual localization (i.e., the likelihood). Multimodal perception, therefore, improves precision (over unimodal performance) and accuracy of the perpetual estimates (Odegaard et al., 2015).

Interestingly, the prior can be shifted. Odegaard et al. (2017) showed that the tendency to integrate AV stimuli changed after repeated exposure to spatially discrepant (but temporally coherent) stimuli. Repeated simultaneous exposure to auditory and visually spatially discrepant stimuli led to an increased tendency to integrate future spatially discrepant stimuli. This might feel counterintuitive, but good temporal precision of the auditory and visual stimuli possibly provided evidence of a common cause. This evidence was weighed more heavily than the less reliable spatial discrepancy and resulted in an interpretation of a common cause for the latter presented spatially disparate stimuli. It led to a correction of the earlier assumed prior by increasing the integration tendency. When stimuli had no strong indication to be perceived as a common cause (e.g., temporally asynchronous; Nahorna et al., 2015 or phonetically incongruent; Nahorna et al., 2012), the opposite scenario was observed. Repeated exposure of incongruent stimuli led to a decrease in the tendency to integrate. Moreover, repeated exposure to AV stimuli with fixed spatial discrepancy, led to modification of auditory spatial perception (i.e., likelihood functions), rather than adjustment of spatial prior expectations (Frissen et

al., 2005). Thus, when the spatial relationship is fixed (as in Frissen et al., 2005) the system treated the more precise signal as the teaching signal and recalibrated the less precise representations by shifting them to reduce discrepancy. In contrast, when the relationship is variable (as in Odegaard et al., 2017), the system did not try to correct either sensory representation but it corrected the model of the world by relaxing its criteria for the perception of an object.

1.2.3.2. Linguistic features

How is the Bayesian Causal Inference Model interpreted for AV speech? Specifically, are there factors specific to language that influence the outcome of the model? AV integration of speech is mainly beneficial in disrupted or complex scenes (Bernstein et al., 2000; 2004; Campbell, 1998; Sumby & Pollack, 1954). So, the model has to (1) predict performance for speech in noise, and (2) interpet linguistic characteristics like word density (Luce & Pisoni, 1998; Auer, 2002) and vocabulary size (Miller et al., 1951). Ma and colleagues (2009) offer a comprehensive response to these questions. First, they showed that evidence for inverse effectiveness (the largest multisensory enhancement is expected when a unisensory stimulus is weakest – Meredith & Stein, 1986), was only present when limited vocabulary was assumed – which is artificial in comparison to conversational environments. When referring to vocabulary size, Ma and colleagues (2009) are not discussing the quantity of unique word stimuli; rather, they are addressing the total number of words that the listener may take into consideration. They reported that AV integration for more realistic scenarios (large vocabulary sets), align more with a Bayesian inference model where a benefit from added visual information is maximal at intermediate noise levels.

From a linguistic perspective, speech recognition can be viewed as a process where perceived phonetic information is compared to a mental lexicon, which is specific to the language experience of the listener. From a Bayesian inference viewpoint; "*word recognition is a process in which vocabulary words are prototypes defined by a conjunction of*

phonetic features, with each phoneme in a word contributing to a set of features" (Ma et al., 2009). A word can be viewed as a specific 'point' within that space, and various acoustic variations of the same word are dispersed in the immediate vicinity of an average prototype (i.e., complementary-systems approach to abstract and episodic speech perception; Goldinger, 2007). When the word is presented in noise, uncertainty manifests about this set of features. The feature space can be viewed as a topographic space with defined neighborhood relationships (Luce & Pisoni, 1998), and words that are closer are more likely to be confused. Additionally, some words have more neighbors than others, and these words in high-density regions are harder to recognize (Auer 2002; Luce & Pisoni, 1998). This can make the feature space high dimensional. The high-dimensional feature space described in this model is exactly what might set integration of AV speech apart from other more simple AV stimuli (~low-dimensional feature space). Luckily, not all phonetic representations represent words in a lexicon (MacEachern 2000) and high-frequency words are easier to recognize (Luce & Pisoni, 1998), reducing the dimensions for listeners who master a language. And exactly the combination of these aspects point to the language-specific prior probability in a Bayesian Inference model.

Based on the clarification of some terminology regarding AV integration (1.2.1.), the proposal to discuss integration of AV speech from a prelinguistic and linguistic perspective (1.2.2.), within the framework of the Causal Inference Model as a Bayesian observer (1.2.3.), we aim to summarize an extensive body of behavioral literature regarding AV integration of speech signals, starting with adults as a reference point, followed by research involving infants, children, and older adults.

1.3. Prelinguistic factors in multisensory integration

1.3.1. Adults

Audiovisual speech is beneficial over auditory-only speech as visual information helps detect speech in a complex scene (Eskelund et al., 2011). The visual features in

AV speech provide temporal markers that correspond to a target auditory speech signal, indicating when to direct attention to a speaker (Bernstein et al., 2004; Fiscella et al., 2022; Peelle & Sommers, 2015). Whether these prelinguistic features contribute to an improvement of speech intelligibility itself is less clear.

The most basic temporal information presented in AV signals is the onset and offset of the auditory and visual signal. The onset information helps us to detect the incoming information (Grant et al., 1998) and get ready for the incoming speech signal as the visual signal slightly proceeds the auditory signal (Chandrasekaran et al., 2009; van Wassenhove et al., 2005). Additionally, the listener can use the temporal amplitude envelope of speech, provided by articulatory information like mouth opening in connected speech. Independent of a language, a larger opening of the mouth corresponds to a louder amplitude (Chandrasekaran et al., 2009). The integration of this AV information enables the listener to attune to the rhythmic aspects of the speech signal and guides the listener about when to anticipate specific acoustic information (Peelle & Sommers, 2015), thereby enhancing the ability to detect speech in complex scenarios (Grant & Seitz, 2000; Kim & Davis 2003). It is argued that exactly this process aids in decoding syllabic and lexical categories (Peelle & Davis, 2012), resulting in an improvement of speech intelligibility. However, one could argue that information regarding syllabic and lexical category decoding only aids the listener when it matches the listener's linguistic knowledge (i.e., language-specific), and therefore the detection benefit is prelinguistic, but the actual intelligibility benefit is linguistic. This idea is corroborated by an experiment of Yuan and colleagues (2020) where they used a visual analogue (sphere) to match the amplitude of the speech signal over time. Although they reported a speech recognition benefit by presenting the visual analogue, it was small in comparison to typically observed AV speech perception benefits, indicating that when the linguistic information was eliminated there was mainly a detection benefit. Another indication for language-independent detection benefits shows from similar sized benefits

between speech and non-speech stimuli (Eramudugolla et al., 2011). Additionally, Kim and Davis (2003) argued that the detection benefit – which was assumed to be prelinguistic, so regardless of the language – was only present for connected speech when the acoustic envelope was highly correlated with the mouth movements. These highly correlated stimuli resulted in an equal AV threshold detection benefit whether you knew the language or not. However, when this correlation was low the outcome was heavily influenced by whether the listener knew the language. Not knowing the language even resulted in worse detection thresholds for these AV stimuli compared to the auditory signal. This suggests that the prelinguistic information provided by stimuli with a low correlation between the acoustic envelope and the mouth movement were not sufficient to integrate these stimuli, regardless of the temporal coherence between the two modalities, and that not knowing the language could have a negative outcome on the perceptual processing of these stimuli.

The intricate connection between prelinguistic and linguistic elements in audiovisual speech signals is also evident in literature employing sine-wave speech (SWS). SWS is a type of synthetic speech where the natural harmonic content of the original signal is replaced by a series of tones or sine waves at the frequencies of the formants in the original speech. Therefore, the temporal and fundamental frequency information is retained, but the spectral details are simplified (Lachs & Pisoni, 2004). If integration of prelinguistic features are sufficient to increase intelligibility, one could argue that listeners should benefit from AV SWS. Interestingly, SWS only supported the perception of the linguistic message when listeners were informed that they could get information out of the signal (Remez et al., 1982; Tuomainen et al., 2005). So, although the speech signal was identical in both scenarios, only when individuals were made aware that it contained 'hidden' phonetic information, they would employ the coherent temporal and fundamental frequency information to create an AV speech perception benefit (Baart et al., 2014). Therefore, similar to the

findings of Kim and Davis (2003), the (language-specific) prior knowledge of the listener needed to be activated to benefit from AV cues in speech.

Nonetheless, prelinguistic cues like temporal coherence play an important role in AV integration. Overlapping, highly-correlated temporal features in both modalities can be sufficient to drive AV integration for speech (Jones & Munhall, 1997) and non-speech (DeLoss & Andersen, 2015; Keetels & Vroomen, 2007) signals in relatively simple scenes. Yet, conversational speech is often not that simple. In more complex scenes (Fleming et al., 2021), or when stimulus localization in space is essential (Stein et al., 1996), spatial coherence can serve as an important secondary cue for AV integration. Fleming and colleagues (2021) showed furthermore that, at least for a closed response set, spatial coherence of the auditory and visual information stream actually improved intelligibility compared to spatial disparity scenes and auditory-only scenarios, indicating that this prelinguistic advantage goes beyond a detection benefit.

Both temporal and spatial information provide statistical indicators to the decision whether or not AV cues originate from a common source and integration is even beneficial (Wallace & Stevenson, 2014). However, these statistical indicators are not fixed. The level of spatial coherence influences the tolerance for AV asynchrony. Zampini and colleagues (2005) showed, for example, that an increase in spatial discrepancy resulted in a decrease in tolerance of temporal AV asynchrony. Furthermore, the importance of AV temporal coherence depends on the larger context. If there are other features more distinct in separating competing streams (i.e., fundamental frequency between a male and female talker), the benefit from temporal coherence can disappear (Cappelloni et al., 2023).

Besides, the sensitivity to coherence of these prelinguistic features in AV integration depends on the listener as an individual (e.g., temporal: Stevenson et al., 2012; spatial: Mihalik & Noppeney, 2020). Whether this is linked directly to prelinguistic

features as temporal synchrony perception (Freeman & Ipser, 2016; Wilbiks & Dyson, 2018), or rather to broader skills and strategy (Van Atteveldt et al., 2014; Wilbiks et al., 2020), it is accepted that individual listeners have a perceptual sensitivity (Peelle & Sommers, 2015), a preference for a certain modality (Seewald et al., 1985), even when they have no impairment in auditory or visual speech perception. Multisensory integration is a weighted average of sensory estimates, and the weight assigned to each modality can thus vary by individual (Schwartz, 2010). Notwithstanding the individual variability, or specific group differences (e.g musicians vs. less musical individuals – Proverbio et al., 2016), most young adults weigh auditory information more heavily than visual information. This is not surprising since the auditory stream provides more information (voicing, manner, and place of articulation) and can lead to 100% intelligibility in a conversation. Visual speech on its own does not provide sufficient information (mainly place of articulation) to have similar intelligibility.

A last prelinguistic component contributing to individual differences of AV integration are differences in cognition influencing our sensory performance. Whether or not cognitive processes should be treated as distinctively different from perceptual processes is a discussion for another time (Michel, 2020; Tacca, 2011). We define both as prelinguistic processes, as they are not unique to a language. Schneider and Pichora-Fuller (2000) argue for an integrated perceptual-cognitive system of AV perception. The perceptual benefit of AV speech perception will lead, according to this theory, to larger availability of resources for higher-order cognitive processes. These can in turn lead to increased behavioral performance for both individual modalities and for the multimodal speech signal (Alsius et al., 2018). Attention (Navarra et al., 2010), expectation (Tuomainen et al., 2005), awareness (Baart et al., 2014; Palmer & Ramsey, 2012), mental imagery (Berger & Ehrsson, 2013), suggestion (Déry et al., 2014), and memory (Frtusova & Philips, 2016) are identified as key cognitive components in this process. So, cognitive processes can

influence our sensory performance (uni- or multimodal), and the benefits perceived by multimodal speech perception can improve our cognitive functioning. For example, for individuals with hearing loss – who have lower working memory performance in the auditory modality – the facilitation caused by the added visual modality can lead to equal AV speech perception performance as their peers with normal hearing thresholds (Grant et al., 2007).

Let us focus more on the role of attention for AV integration. In the context of AV integration we have already discussed object-based attention in Section 1.2.1. There is evidence that object-formation of coherent AV features can be formed in the absence of attentional modulation, as it has been observed in anesthetized animals (Atilgan et al., 2018). With these AV objects formed in a complex scene, a particular target object can be attended selectively and as based on the theory of this object-based attention, all crossmodal features in this AV object are also enhanced. When there are multiple AV objects (e.g., in a bar there is AV information of multiple speakers), competition between AV objects is biased towards information that is currently relevant for the listener (Desimone & Duncan, 1995), towards the information they selectively attend (Bizley et al., 2016), and towards information that is more salient (Shinn-Cunningham, 2008; Yantis, 2005). However, dividing our selective attention works to our detriment. For instance, when we are watching a sportscaster's mouth on TV, while simultaneously attending the voices of the conversation of our friends in the couch, our attention strategy is doomed to fail, as the visual stream is coherent with the masker (i.e., the sportscaster's voice) and not the target talker (i.e., the friends' voices). We analyze the selectively attended information stream (i.e., friends' conversation) at a cost as our attention will be divided (Maddox et al., 2015), and the AV object (i.e., the sportscaster on TV) is the more prominent information, overriding top-down attention to the target (Shinn-Cunningham, 2008). Yet, in social settings we seem to be able to understand multiple sources, with multiple AV objects, even when the conversation is chaotic

and unpredictable. The combination of attention switching between objects in a complex scene, the use of our short-term memory, and the use of contextual cues (e.g., linguistic knowledge) is instrumental to fill in missing information (Shinn-Cunningham, 2008).

1.3.2. Infants

One of the first things an infant experiences in life is a parent bringing their face close to the infant while communicating. This immediate interaction exposes the baby to both auditory and visual speech cues. Hence, it is not that surprising that newborns as young as 4 hours (Aldridge et al., 1999) already prefer matching AV stimuli (in time and space), and start recognizing this innate relationship between vision and sound in the first months of life (Dodd, 1979). Infants of four months already look for physical objects after being presented with sound (Spelke, 1976) and selective attention will have a significant role in finding information that is meaningful and coherent in a world filled with relevant and irrelevant information (Bahrick & Lickliter, 2012).

For adults, this process of selective attention is often driven by experience and top-down processes, whereas for infants – who have limited prior knowledge – it is more stimulus-driven (Bahrick & Lickliter, 2012). Infants initially use prelinguistic information like multisensory redundancy (temporal: Spelke et al., 1983; spatial: Walker-Andrews & Lennon, 1985), and combine this later with more complex features of the vocal tract (e.g., intensity, duration, tempo, and rhythm – Spelke, 1979), or gender of the speaker (Patterson & Werker, 2002), to shape early selective attention and, in turn, perception and learning (Bahrick, 2010; Lewkowicz, 2000) in the first months of life. The sensitivity to these redundant features, in combination with increasing sensitivity to the statistical regularities of the environment allows young infants to attend AV events such as a person speaking, a dog barking, a car passing (Bahrick & Lickliter, 2012).

Bahrick and Lickliter (2000, 2002) suggest in their framework of "Intersensory Redundancy Hypothesis" that selective attention, perceptual processing, learning, and memory all follow from an infant's sensitivity to intersensory redundancy. Infants focus on these redundant features at the cost of other unimodal features due to limited attentional resources in early development. Later in development, when attention is educated, less salient, non-redundant features can be processed. The model explains through this concept of 'educating attention' that human (Castellanos et al., 2006) and non-human (Jaime et al., 2010) infants need to be exposed to multisensory redundancy features to translate them to unimodal features. For example, (4-month-old) infants can only detect a change in tempo of a visual representation of a toy hammer tapping (i.e., unisensory) after they have been exposed to a synchronous multisensory AV representation of this toy hammer tapping (Castellanos et al., 2006). When infants become older and more experienced, attention becomes more efficient and flexible, processing speed increases, and perceptual processing improves (Gibson, 1969), so more attentional resources become available to learn from both redundant and non-redundant, and less salient (Bahrick & Lickliter, 2000) and higher-level features (Lewkowicz, 2010).

Next to multisensory redundancy, statistical learning – the sensitivity to probabilistic regularities in sensory input – has been pointed out as an important mechanism in the development of language learning (Romberg & Saffran, 2010) and scene perception (Fiser & Aslin, 2002). This sensitivity to probabilistic regularties helps segment a continuous speech stream into meaningful units and a complex scene into multiple objects. Statistical learning has been demonstrated in both auditory and visual modalities in infants (e.g., Fiser & Aslin, 2002; Kirkham et al., 2002; Saffran et al., 1996). Yet, statistical learning might not be a multisensory (amodal) process in itself. Emberson and colleagues (2020) reported that infants (8-10 months old) preferred novelty in the auditory modality, whereas they preferred familiarity in the visual modality. In the context of statistical learning a familiarity preference reflects

a weaker stage of encoding than a novelty preference (Hunter & Amers, 1988), suggesting that infants have weaker statistical learning in the visual modality and process these modalities at a different speed, in a different phase in life. However, while infants (and adults – Conway & Christiansen, 2009) might mainly use statistical learning from the auditory modality to segment (temporal) speech information, when presented with AV speech, statistical learning mechanisms still track multimodal input and only 'allow' statistical learning in one modality when there is (temporal) redundancy between features of the AV signal (Mitchel & Weiss, 2011).

The importance of redundancy of features in both modalities this early in life is not surprising. Everyday AV events are temporally synchronized and therefore temporally synchronous patterns of AV signals are available from birth on and provide a great basis for AV integration (Gibson, 1966). Additionally, detection of AV temporal synchrony requires minimal experience (Cox et al., 2021). It only requires onset and offset detection and no complex feature processing – as the latter might not be available for infants – is needed (Lewkowicz, 2010). Consequently, newborn infants (< 3 days) already show a preference for congruent AV stimuli (Guellaï et al., 2016). However, this does not mean infants are equally sensitive to temporal coherence as adults. Infants need temporal offsets four to five times larger than adults to detect asynchrony (Lewkowicz, 1996).

Thus, similar to what we see in adults, temporal coherence is a key contributor in integration of AV speech stimuli. Furthermore, work with adults showed that when there are more salient features (i.e., gender differences) present to inform about the AV integration process, temporal coherence had no effect (Cappelloni et al., 2023). Infants at age 4.5 months showed similarly that temporal AV matching performance could be disrupted based on conflicting gender information. This indicates that certain features (like gender) already play a role in the integration decision of the Causal Inference process early on.

Spatial information also has a role in speech perception of infants. Infants orient their head (and eyes) towards new visual and auditory stimuli from birth (Fantz, 1963), and therefore, it is not surprising that infants are already sensitive to the benefits of AV localization. Young infants (8-months-old) can already combine AV localization cues, however, the actual benefit only arises after 10 months of age (Neil et al., 2006). Very young infants (2.5-months-old) also rely less on the AV spatial congruency than older infants, to learn about objects that are temporally bound (Bahrick & Lickliter, 2001).

Thus, infants are already sensitive to multisensory information right after birth. They are even able to use cues like gender and spatial localization in their integration decision, however the integration process itself – to obtain AV speech benefits – is only influenced by these cues when they become more developed.

1.3.3. Children

Children are often exposed to noisy environments e.g., on the playground, in daycare, at home. (Knecht et al., 2002). Similar to infants, they have immature perceptual, cognitive, and linguistic skills (McCreery et al., 2017), resulting in a need for efficient AV integration of speech information.

Perceptual auditory and visual systems show improvement into late childhood (Moore, 2002). This long maturation process of the auditory system represents an increasing ability to recover auditory speech in noise with increasing age (Nittrouer & Boothroyd, 1990; Wightman et al., 2006). For both the auditory and visual system most basic skills are acquired in infancy. It is only more complex perceptual skills that develop until adolescence (e.g., auditory: localization, visual: face selection and visual attention – Dye & Bavelier, 2010). The later start of visual processing in general, and more specifically of these more complex, but important, skills in AV speech perception could explain (1) relatively poor speechreading (visual-only) performance in children (Holt et al., 2011; Massaro et al., 1986), (2) lower

performance on tasks that require face-processing in general (de Gelder et al., 1998), (3) and favoring the auditory modality in AV speech, especially for conflicting information (Hockley & Polka, 1994).

Following the Bayes model for Optimal Integration (Ma et al., 2009), children will, like adults, apply weights to individual perceptual modalities in their decision if and how to integrate AV speech. Children on average weigh auditory information more heavily than visual information (Hockley & Polka, 1994). The auditory stream provides more information about the speech signal and is therefore in general the more reliable input stream. This auditory dominance phenomena – attention is captured by the auditory stimulus when presented with AV stimuli – is even more expressed in children than adults (Hockley & Polka, 1994). Their more limited experience with visual speech, compared to auditory, results in more limited representations of visual phonological knowledge and therefore larger uncertainties in the visual modality (Desjardins et al., 1997; Jerger et al., 2009). The combination of (1) weighing the auditory stream more heavily (Hockley & Polka, 1994), (2) immature auditory performance in noisy environments (Wightman et al., 2006), and (3) predominantly depending on prelinguistic mechanisms, such as temporal cues, for children in integrating AV stimuli (Lalonde & Holt, 2016), might explain why AV integration, and AV speech perception benefits increase from early childhood into adolescence (Lalonde & Holt, 2016; Lalonde & McCreery, 2020; Massaro et al., 1986; McGurk & MacDonald, 1976).

However, these reported age-related differences are much more nuanced. Ross and colleagues (2011) indeed showed an increase in AV speech perception benefits from 5-year-olds to adults, however only in signal-to-noise ratios (SNR) worse than -3dB. Although not impossible, this (SNR < -3dB) is a scenario children (and adults) do not often encounter when listening to speech. At home (SNR = +9 to 15 dB; Benitez-Barrera et al., 2020) or even in primary school classrooms (SNR = +12dB; Sarantopoulos et al., 2004), the SNR is most often positive. Regardless, in everyday

noisy speech environments listeners need to constantly leverage perceptual input with cognitive constraints (i.e., limited working memory and attentional capacity) and linguistic experience. A more immature frontal cortex hinders the adoption of automatic strategies (Thillay et al., 2015), placing greater demands on sustained attention in children. Coupled with their limited processing capacity (Huang-Pollock et al., 2002), this contributes to a broader maturational process in psychophysical testing performance. Undoubtedly, this leads to lower performance on (AV) speech tasks in younger children compared to older children or adults. To reduce these covarying influences, Gijbels and colleagues (for details see Chapter 3) designed a task for children age 4-15-years old, limiting cognitive constraints and linguistic experience to a level such that the AV speech signal should be similarly accessible for children age 4-years-old and up. With cognitive constraints and linguistic experience taken into account, they asked (1) whether developmental differences in AV speech perception benefits would still remain, (2) how perceptual weights were assigned in AV speech perception across this age range, and (3) how the causal inference problem is addressed in childhood development. They showed, based on data from 161 children, that even the youngest children presented reliable AV speech perception benefits, and that these benefits were consistent throughout development when auditory and visual signals matched. The large individual variability between these children was explained by how the child weighed their auditory speech-in-noise performance in AV speech, rather than the quality of the signal itself, or the age of the child. Interestingly, they did show developmental differences between younger and older children when comparing performance between congruent and semantically incongruent stimuli. Older children seemed to be more affected by these incongruent stimuli, which were mixed in between the congruent AV speech trials. Older children might therefore employ prior information more to make decisions about the AV speech stimuli (i.e., to recalibrate) and consequently be more influenced by the incongruency of both input streams. These findings are in line with the suggestion from Rohlf et al. (2020) that multisensory integration

develops prior to crossmodal recalibration, therefore AV speech perception benefits might already be stable in kindergarten, crossmodal recalibration needed for incongruent stimuli is not.

In light of the potential differences in the Causal Inference decision across development, processing of temporal and spatial information needs to be considered. Younger children have a harder time detecting asynchrony of AV stimuli (Lewkowicz & Flom, 2014), resulting in wider temporal binding windows (TBW). The narrowing of the TBW continues, on average, well into adolescence (non-speech: Hillock-Dunn & Wallace, 2012; speech: Stevenson et al., 2018). And as we know that on an individual level, adults with wider TBWs are less susceptible to the McGurk illusion[1] (McGurk & MacDonalds, 1976) – and therefore make less errors of incorrectly integrating incongruent AV speech stimuli – (Stevenson et al., 2012), and younger children – who as a group have wider TBWs – are less susceptible to the McGurk illusion, it is suggested that developmental differences in sensitivity to temporal synchrony perception are important in explaining developmental differences in the Causal Inference process in AV speech stimuli.

For spatial information the developmental differences might be less pronounced. Complex localization cues (in the horizontal and vertical plane) mature around age 6-years-old. In younger children, vertical localization primarily depends on visual cues. However, by the age of six, children begin incorporating auditory information into their decision-making process for localization, aligning with the principles of the Causal Inference Model (Gori et al., 2021).

[1] The McGurk effect is an illusion where the listener is presented with temporally and spatially coherent AV signals, however the phoneme-viseme connection does not match. A visual /ga/ is presented simultaneously with an auditory /ba/, triggering the perception of a third illusionary syllable /da/ or /ða/. The illusion demonstrates that speech perception, even in quiet, is not only an auditory process, and has therefore often been used as a proxy measure for AV integration (Alsius et al., 2018). The effect has been reported at all stages of development (from infancy to older adults), however not all listeners – even within the same age group – show the same degree of susceptibility to the illusion (Alsius et al., 2018; McGurk & MacDonald, 1976; Van Engen et al., 2022).

1.3.4. Older adults

The aging process is associated with a gradual decline in function, commencing in early adulthood. Both auditory and visual perceptual decline are integral to this normal aging process (Fozard & Gordon-Salant, 2001). Subtle perceptual changes, such as reduced speech perception in noise and declining contrast sensitivity, may significantly impact the daily lives of older adults, by limiting the amount of extractable information. Inevitable changes in the central system further affect information processing, working memory, attention, and retrieval (Humes & Christopherson, 1991). While perceptual decline is inevitable for everyone, the pace, severity, and specific perceptual processes affected often vary by individual (Guerreiro et al., 2013).

The interpretation of age-related changes in integration of AV speech becomes particularly complex due to changes in unimodal processing (Sommers et al., 2005). Not surprisingly, there are reports of lower (Gordon & Allen, 2009; Tye-Murray et al. 2008; 2010; 2011), equal (Gordon & Allen, 2009; Helfer, 1998; Middelweerd & Plomp, 1987; Sommers et al., 2005), and higher (Dias et al., 2021; Sekiyama et al., 2014; Setti et al., 2013) integration of AV speech in older adults compared to younger adults.

Two approaches can be employed to address individual perceptual differences. First, stimuli could be individually adjusted, irrespective of age, to ensure uniform unimodal performance across all participants. When doing so (Helfer, 1998; Sommers et al., 2005), similar AV integration is reported between younger and older adults. A second approach is to track performance longitudinally. Voss (2016), for example, reported a continuous decline in auditory performance when aging, whereas visual decline was more stable, at least after the age of 65. This implies that the visual modality becomes relatively more reliable during the aging process (Brooks et al., 2018; Freiherr et al., 2013).

As the auditory modality loses reliability in the aging process, a shift in perceptual weights in AV integration can be expected. Indeed, older adults direct their attention more to the speaker's mouth than younger adults (Thompson & Malloy, 2004), regardless of their lower speechreading performance (Sommers et al., 2005). And AV integration has been measured to be especially lower in older individuals with degraded visual skills (Legault et al., 2010) or who are presented with degraded visual stimuli (Gordon & Allen, 2009; Huyse et al., 2014; Tye-Murray et al., 2010, 2011). Furthermore, during assessment of the AV illusion, the McGurk Effect (McGurk & Macdonald, 1976), older adults are more likely to report the visual alternative (Ballingham & Cienkowski, 2004; Cienkowski & Carney, 2002; Sekiyama et al., 2014; Setti et al., 2013).

This visual preference can be translated to temporal synchrony perception and spatial localization in older adults. Older adults show a larger tolerance for temporal asynchrony (wider TBWs) in comparison to their younger peers (Hay-McCutcheon et al., 2009; Stevenson et al., 2018). Various factors could account for the observed variations in synchrony perception. First, older adults tend to have a more conservative response bias (Chan et al., 2014), and as the wider TBWs are mainly reported for auditory-leading stimuli (Hay-McCutcheon et al., 2009) – which do not occur in real-life situations, and definitely not in conversational speech – the different interpretation of temporal coherence might just represent this response bias. Second, age-related hearing loss disrupts the ability to code temporal information represented by the speech envelope (Moore, 2014). A change in temporal synchrony detection is therefore not surprising. However, the fact that this is not measured for visual-leading stimuli might actually mean that older adults, who put heavier weight on these visual cues, can actually benefit from visual speech to an extent that their synchrony perception for visual-leading stimuli is similar. The conservative response bias in older adults then translates to higher decisional thresholds, and might therefore also explain why older adults show no difference in

the visual-leading TBW. Similarly, in the spatial domain Jones and colleagues (2019) showed that older and younger adults show comparable localization decisions in an AV spatial localization task. Introducing incongruent AV trials showed that aging had no influence in the integration decision; however, responses were significantly slower for incongruent trials in older adults. Therefore, older observers preserved AV localization performance, by sacrificing response speed, due to a longer accumulation of noisier auditory representations (Jones et al., 2019), and a heavier reliance on visual cues (Zou et al., 2017).

1.4. Linguistic factors in multisensory integration

1.4.1. Adults

After identifying the target AV speech stream, we can use more advanced speech-specific cues of AV speech to improve the quality (Eskelund et al., 2011; Tye-Murray et al., 2007a) and the speed (Reisberg et al., 1987) of speech understanding (speech recognition).

At the most basic phoneme-viseme level auditory and visual signals provide both redundant and complementary information, as voicing and manner of articulation are mainly presented in the auditory modality, and place of articulation is represented in both modalities (Grant & Seitz, 1998). Complex scene analysis via AV integration occurs exactly because of redundancy of information in both streams. The complementary information in both speech streams can help to "fill in the gaps" and improve the quality of speech perception (Grant et al., 1998; Walden et al., 1977).

On a word-level, word frequency and word density each play a unique role in AV speech recognition. Even when the redundancy level of two words is identical, if a word is more frequent in a language it will have higher recognition rates unimodally, and therefore benefit less from AV speech (van de Rijt et al., 2019). In contrast, high-density words, meaning words with a high number of phonological or

visual neighbors, are often harder to recognize because there are more items to choose from (Bradlow & Pisoni, 1999). However, when multiple modalities are present, competitors of each modality will be restricted until only competitors from both auditory and visual speech will remain. The size of this intersection density is a good predictor for AV performance and explains improved accuracy, potentially higher processing speed, and lower effort that has been observed during multisensory speech perception (Tye-Murray et al, 2007b).

Yet, conversational speech is not just a syllable or a word, it is presented in a meaningful context. This semantic context improves auditory and AV speech intelligibility in challenging conditions (Bradlow & Alexander, 2007). However, the benefit observed from semantic context in the auditory modality cannot be simply extended to multimodal speech. For example, Van Engen and colleagues (2014) showed that semantic context was beneficial for AV speech for both fluctuating (2-,4-,8- talker babble) and stationary maskers (speech-shaped noise), but auditory speech only benefited from semantic context in speech-shaped noise. Therefore, the combination of added visual information and semantic context can have an increased impact on speech perception, providing larger benefits (Ma et al., 2009; Tye-Murray et al., 2007b).

How well and easily you can access this context is determined by the linguistic knowledge of the listener. Linguistic knowledge includes the individual's vocabulary or lexicon, knowledge of semantics and syntax, use of context and memory processes, and strategies for lexical access (Grant et al., 1998). Top-down processes driven by this linguistic knowledge have in turn an influence on our perceptual weighting of the AV information stream. This is individually and culturally defined (i.e., language dependent), and therefore can be learned. Multiple cross-lingual studies showed that visual cues in AV speech signals of different languages were weighted differently in the perceptual decision process, even when very similar items between languages were used (e.g., English vs. Japanese:

Sekiyama & Tohkura, 1993; Spanish vs. German: Fuster-Duran, 1996; German vs. Hungarian: Grassegger, 1995). This points to differences in learning processes between cultures, both linguistically (e.g., differences in categorization of speech – Kuhl, 1994), and perceptually (e.g., differences in gaze patterns/attendance to the face – Sekiyama & Tohkura, 1993).

1.4.2. Infants

Presenting speech to infants, and extracting meaning from speech is crucial for infants to learn a language. So, in the context of AV speech perception and more specifically defining the role of linguistic information in AV integration, it is really interesting to study infants. At birth, infants are sensitive to coherent multisensory information (Guellaï et al., 2016), yet they have not learned a language. Studying AV speech in infants therefore informs us about (1) the role of AV features in language learning, and (2) the contribution of linguistic information in the differences of AV integration reported across development.

First, we will address the role of AV features in language learning. In the context of language acquisition, newborn infants have boundless abilities. A newborn infant has no language preference, so whatever language is consistently offered to them, is what will grow into their native language(s). However, this broad ability to learn confines when infants grow older. Early language development contains a phase of perceptual narrowing (Werker & Tees, 1984), meaning that auditory speech sensitivity in early infancy is broadly tuned (i.e., not language specific), but when infants grow older this sensitivity limits, until they are only sensitive to native language cues (Kuhl, 2004; Kuhl et al., 2006). A similar process happens for multisensory speech perception, where children of 6 months are still sensitive to matching AV speech of multiple languages, yet at 10-12 months AV sensitivity of speech becomes specific to their native tongue (Lewkowicz & Pons 2013; Kubicek et al., 2014; Pons et al., 2009).

Interpreting developmental differences regarding the use of phoneme-viseme connections in AV speech before (<6 months-old), during (6-12 months-old), and after (children > 12 months-old, and adults) this phase of perceptual narrowing would be especially informative. However, one has to be cautious in widely comparing AV speech perception literature across development. As Lalonde and Werner (2021) pointed out, studies across different age groups often have different methods and different cognitive requirements. Infant studies often use indirect measures, resulting in a sensitivity measure, whereas child and adult studies use direct measures to quantify the size of AV speech perception benefits. Importantly, as highlighted by Shaw and Bortfeld (2015) associating and integrating AV cues is not the same. Nonetheless, there are some conclusions to be made from the current body of literature.

Higher exposure to (9- to 12-month-olds: Dodd, 1972), and being more attentive of (6-months-olds: Young et al., 2009) AV speech (i.e., watching their mother's mouth) early in life is correlated to better expressive vocabulary skills, as infants (Dodd, 1972) and later in life (Young et al., 2009). Interestingly, the infants in these studies fall exactly within the phase of perceptual narrowing. Between four and eight months of age infants' attention shifts from the eyes to the mouth when listening to a talker, regardless of the language of the talker (Lewkowicz & Hansen-Tift, 2012). This shift helps attune infants to phoneme-viseme connections of their native speech forms. For example, Teinonen and colleagues (2008) reported that infants in the phase of perceptual narrowing (6-months-olds) use visual cues to distinguish between phonetic categories in a phoneme categorization task and subsequently even translated this to unimodal auditory speech categorization. Indeed, around 12-months of age infants can use visual salience of phoneme-viseme connections. Weatherhead and White (2017) showed that infants (12-months-old) could detect mispronunciations (i.e., incongruent phoneme-viseme pairs) in words, when the mispronunciation was visible (change in place of articulation), but not when it was

only an auditory change (change in voicing). After this phase of perceptual narrowing (around 12 months) attention shifts back from the mouth to the eyes (for native speech only), which allows infants to increase native-language expertise by including social cues (Lewkowicz & Hansen-Tift, 2012).

To our second question, how linguistic information contributes to differences in AV integration across development, we conclude that AV detection benefits between infants and adults are very similar (6-8-months-old: Lalonde & Werner, 2019). However, infants might use AV features more flexibly than adults (Baart et al., 2014). For example, infants (5-15 months) can detect coherence between the auditory and visual modality for both regular speech and SWS, whereas adults needed to be informed that SWS contained (language-specific) phonetic information to use it (Baart et al., 2014). Furthermore, adults mainly rely on spectrotemporal information in the decision-making process of segregation versus integration for AV speech (Jones & Munhall, 1997), whereas infants show more flexibility as they also use redundant prosodic information (Kuhl et al., 1991). It had been argued that for fluent speech infants only detect AV coherence in late infancy (Lewkowicz et al., 2015), pointing to large maturational differences based on extensive linguistic knowledge for complex speech. Yet, when infants are presented with exaggerated prosodic information, like in infant-directed speech, they did show sensitivity to multisensory coherence of fluent speech early on (2-6-month-olds: Englund & Behne, 2020; 7.5-month-olds: Hollich et al., 2005; 12-month olds: Kubicek et al., 2014). So, although underdeveloped linguistic knowledge has its role in AV speech perception of infants, infant-directed speech is thought to have evolved as a species-specific adaptation (Fernand, 1995) where parents intuitively adjust their speech to the infant's developmental stage (Shapiro et al., 2021), and therefore facilitate the use of AV cues already in infancy.

Beyond sensitivity to AV speech, infants (6 - 8.5 months-old) showed AV discrimination benefits of AV speech in noise (Lalonde & Werner, 2019). However,

these benefits were significantly smaller than what has been observed in adults. For infants this discrimination benefit of the full speech signal was similar to a temporal onset/offset cue, whereas adults obtained larger benefits when presented with the full speech signal (Lalonde & Werner, 2019). Therefore, it is plausible that infants only used temporal onset information in their discrimination decision. Yet, there is evidence that infants benefit from more complex spectrotemporal cues like the speech envelope (7.5-month-olds; Hollich et al., 2005) in discriminating AV speech from competing sources. Furthermore, basic linguistic information also plays a role in AV speech perception early on. Even in the absence of differences in temporal information infants could distinguish between matched and mismatched AV stimuli (vowels) at a very young age (Less than 2 days old: Aldridge et al., 1999; 4-5 month-olds: Kuhl & Meltzoff ,1982; 1984; 4.5-months-old: Patterson & Werker, 1999; 2-months-old: Patterson & Werker, 2003; 4.5-months-old: Yeung & Werker, 2013), and at 12-months of age infants could discriminate between correct and incorrect phoneme-viseme representations (Weatherhead & White, 2017). A noteworthy caveat is that these phoneme-viseme representations needed to be visually salient for infants, but not for adults, confirming that infants are still developing these linguistic features.

Thus, there is evidence for both AV speech detection and discrimination in infants very early on. Nonetheless, literature showing integration of AV speech signals in infants stays very limited. To provide some extra insights we will briefly discuss the McGurk Effect (McGurk & MacDonald, 1976) in infants. We have minimized the implementation of this well-known AV illusion into this review so far, as increasing neural and behavioral evidence shows that the McGurk effect taps into different mechanisms than AV integration from face-to-face speech (for review: Alsius et al., 2018; Van Engen et al., 2022). Yet, when presenting infants with congruent and incongruent AV stimuli that potentially induce an illusionary effect, we can learn something about how they use Causal Inference (Körding et al., 2007) in the

integration process of AV speech stimuli. Burnham and Dodd (2004) exposed infants (4.5-months-old) to either incongruent A/ba/V/ga/ or congruent A/ba/V/ba/ stimuli, followed by auditory stimuli /ba/, /da/ or /ð/a/. Visual fixation measures showed that /da/ or /ð/a/ were only perceived as unfamiliar syllables by the group that was never presented with the incongruent stimuli. Therefore, they concluded that the group that was exposed to the incongruent stimuli perceived /da/ or /ð/a/ during the incongruent McGurk stimulus habituation period and thus integrated AV speech. However, Tomalski and colleagues (2013) tried to replicate this in 6-9-month-old infants but added a third condition, an incongruent AV stimulus (A/ga/V/ba/) that could not trigger an illusionary percept. They reported similar visual fixation responses for both incongruent stimuli, suggesting that infants did not experience the illusion and relied less on phonetic cues in the integration process than adults. Furthermore, Kushnerenko and colleagues (2008) used similar stimuli, but neurophysiological measures instead of fixation times, and reported that only these non-fusible incongruent stimuli led to differential brain responses in 5-month-old infants. This in combination with the findings from Burnham and Dodd (2004) suggests that infants (at least before the start of the perceptual narrowing phase), are susceptible to AV speech illusions, and incorrectly integrate incongruent AV stimuli based on linguistic information like phoneme-viseme connections.

1.4.3. Children

Large individual variability in the development of language skills (Bates et al., 1995), in combination with the developmental maturation process of language learning (Chomsky, 1965) and perception (Moore, 2002) until adolescence, and the interconnected relationship of speech perception and language (Golestani et al., 2009), stress the complexity of AV speech perception across development.

As a result, the influence of linguistic context on AV speech perception might be more pronounced in children than adults. Taitelbaum-Swead and Fostick (2016)

investigated the unique contribution of phonetic context in AV speech compared to auditory-only speech, from age 4-years-old and up. They report, similar to Tye-Murray and colleagues (2007b), that compared to auditory speech fewer phonemes are needed to perceive the whole word in AV speech, and more interestingly that this is not different depending on the age. This suggests that (at least in Hebrew) linguistic phoneme-viseme information is readily available at age four. This is in line with the finding of Gijbels and colleagues (See Chapter 3), who showed that children at age four used phoneme-viseme connections in a speech intelligibility task in English. Lalonde and Holt (2015) extended the use of AV phoneme-viseme connections to three-year-olds for discrimination tasks, but not for speech recognition tasks, indicating a difference in user-effectiveness in this youngest group. Furthermore, they reported that although 4-year-olds showed AV speech recognition benefits based on phoneme-viseme connections, they were smaller than the ones reported in adults and they only perceived benefits for visually salient cues, whereas adults can use both more and less salient cues.

In a broader linguistic context, lexical information can be discussed. In adults, consonant phonemes were better recognized in words than in nonsense words, and this was significantly better in the AV modality (Fort et al., 2010), demonstrating that visual cues in AV speech helped gain lexical access, especially for low-frequency words where lexical access is difficult. Fort and colleagues (2012) repeated this task with children (5-10-years-old), but changed consonant recognition to vowel recognition to make the task easier. Similar to adults, at least from the age of 5, children showed a word-superiority effect (i.e., better recognition for words than nonsense words). From age 6 and up children also showed a clear AV speech recognition benefit in noise, that was equal for all ages tested. However, in contrast to adults, no interaction was reported between modality and word-superiority-effect for children. Fort and colleagues (2012) therefore stated that visual cues did not aid in lexical processing in children (age 5-10-years-old). However, a remarkable

difference between both studies (Fort et al., 2010; 2012) was that the lexical decision in children was based on vowels, whereas adults based their decision on consonants. Cutler and colleagues (2000) showed that, independent of the language, vowel information constrains lexical selection less tightly than consonant information, which potentially explains the absence of an interaction-effect with the word-superiority task.

In sum, AV speech recognition benefits have been reported as early as in 2-year-olds (Król, 2018). While Lalonde and Holt (2015) argued that the AV speech recognition benefit in three-year-olds was solely based on temporal information (i.e., prelinguistic), and that linguistic information only contributes after age 4, Król (2018) reported that AV speech recognition benefits strongly correlated with with vocabulary scores arguing linguistic processing already plays a role at 2-years-old. However, they used a look-while-listening paradigm to measure AV speech perception benefits and their findings differ from other studies that report a later start of AV speech recognition benefits (3-years-old: Desjardins et al., 1997; Lalonde & Holt, 2015; 4-years-old: Erdener & Burnham, 2018, 6-years-old: Fort et al., 2012; Maidment et al., 2015; 9-years-old: Wightman et al., 2006). Task differences tapping into (1) stimulus complexity (Wightman et al., 2006), (2) attentional resources (Jerger et al., 2014), or (3) speech production and articulatory skills (Yeung & Werker, 2013; Desjardins et al., 1997) might play a significant role in these contradictory findings.

As language, perception, and cognition develop across childhood, they mutually influence each other positively. Speech perception predicts later language development (Kushnerenko et al., 2013), and the better the language skills, the more room there is for improved AV integration (Erdener & Burnham, 2018). Furthermore, AV speech guides, orients, and manipulates attention and learning (Bahrick & Lickliter, 2000), and can even compensate for limited attentional resources (Salinas-Marchant & Macleod, 2021).

1.4.4. Older adults

In order to reconcile the presence or absence of an aging effect in AV speech perception performance studies, it is valuable to clarify the type of linguistic information addressed by the task. Stevenson and colleagues (2015) were able to characterize age-related AV speech perception benefits at the fundamental linguistic level (phoneme-viseme connections) as well as at a lexical level. Measured via a phoneme analysis of a word recognition task, individuals of all ages exhibited comparable AV speech perception benefits. In contrast, on the word-level older adults showed greater AV benefits at intermediate SNRs but reduced benefits at low SNRs, suggesting elementary linguistic processing skills (phoneme-viseme connections) stayed preserved in a healthy-aging population, but older adults benefited less from lexical information in noisy environments. This work specifically showed that solely perceptual changes – age-related hearing loss in combination with lower SNRSs – are not sufficient to reduce AV integration in older adults. Only in combination with the need to access lexical information do older adults show reduced AV integration skills. Visual cues aid AV speech perception for a lexically complex signal by limiting its competing lexical neighbors (see Neighborhood Activation Model; Tye-Murray et al., 2007b). Age-related hearing loss reduces the ability to select lexical neighbors, but this is especially apparent in noisy environments as older adults have lower inhibitory abilities that are needed for both lexical activation access and AV object selection of speech-in-noise (Dey & Sommers, 2015). Furthermore, there is a point at which individuals cannot extract any useful information anymore and this might be reached earlier (less decline of SNR) for individuals with more impaired perceptual skills, like older adults (Stevenson et al., 2015).

Interestingly, age did not affect AV speech perception performance in sentences as long as semantic context was provided (Smayda et al., 2016). However, older adults did show larger AV speech perception benefits more from high-context sentences

compared to low-context sentences (Gordon & Allen, 2009). Therefore, these higher-order predictable linguistic cues can help older adults to tolerate more difficult sensory circumstances, as older adults use compensation strategies (e.g., linguistic context) and stronger top-down regulation of attentional focus to keep behavioral responses relatively stable (Cabeza et al., 2002).

1.5. Conclusion

Integration of AV speech is a complex phenomena uniquely benefiting communicational needs and language development. Approaching if and how AV integration occurs through a prelinguistic and linguistic lens allowed us to review AV speech perception from infancy to older adults. Accumulating behavioral evidence suggests that both mechanisms have their role starting in infancy, constantly influence each other, and are further shaped throughout the lifespan. Selective attention to coherent multisensory events in complex scenes is triggered immediately when infants are exposed to both auditory and visual information in the world and plays a major role in making sense of the complex world throughout the lifespan. How AV speech signals are integrated is highly dependent on the information present and the prior knowledge approaching this information, from the perspective of the signal, the context, and the individual.

1.6. Outline of the dissertation

The research conducted in this dissertation adds to the current understanding of the field of AV speech integration in four ways. These studies are of significance to our further understanding of (1) the applicability of remote assessment of psychophysical AV speech perception tasks in both children and adults (Chapter 2 & 5), (2) the developmental trajectory of AV speech integration (Chapter 3), (3) the role of AV speech integration in developmental disorders such as developmental dyslexia (Chapter 4), and (4) the intricate details of both prelinguistic and linguistic information in the process of AV speech integration in adults (Chapter 5).

Considering that these research projects commenced in 2020, a period marked by a significant surge in remote assessments, with limited understanding and guidelines pertaining to the intricacies of (1) remotely testing children and (2) encountering a substantial loss of methodological control in this context, our initial focus is on outlining a set of considerations for remote work, particularly with children (Section 2.1; published as Gijbels, Cao et al., 2021). Additionally, we examined the utility of employing comparison measures, as opposed to absolute measures, to address the challenges associated with the lack of control inherent in (unmoderated) remote testing scenarios (Section 2.2; published as Gijbels & Lee, 2023).

After establishing feasibility of remote AV speech perception measures (Chapter 2), this dissertation outlines three studies involving a total of 261 unique participants between the age of 4 and 15 years old. The objective is to enhance understanding of the developmental trajectory of AV speech perception benefits.

The initial study (Section 3.1, N = 37, published as Gijbels et al., 2021) serves as a proof-of-concept for an AV speech recognition task in 6-year-olds, which aims to limit cognitive and linguistic developmental influences on task performance. This work investigates whether the visual salience of phonemes influences AV speech perception benefits in first graders and explores whether individual differences in AV speech enhancement can be attributed to vocabulary knowledge, phonological awareness, or general psychophysical testing performance.

Building on this proof-of-concept work, the second study (Section 3.2, in revision Gijbels et al., 2024) expands the task to a larger group of children (N = 161) aged 4 to 15. This work focuses on describing the developmental trajectory of AV speech perception benefits concerning the use of linguistic information presented by phoneme-viseme connections, the perceptual weighting of individual modalities in predicting AV speech perception benefits on an individual level, and potential differences in the Causal Inference process.

In the fourth chapter, the same task in a different population (N = 135, with 63 unique individuals compared to Section 3.2; age 6 - 15) aids in understanding AV speech perception in children with developmental dyslexia (published as Gijbels et al., 2023).

The combined insights from working with children, along with a better understanding of recent literature developments, highlight the need to comprehend integration of AV speech by delineating both prelinguistic and linguistic information processing of the AV speech signal.

Chapter 5 encompasses four experiments conducted remotely in 46 young adults (age: 21 - 40). These experiments aim to:

1. Determine the influence of different levels of linguistic complexity on temporal synchrony perception of AV speech signals, providing unique insights into the contributions of various levels of language processing on perceptual decisions based on variations of prelinguistic temporal coherence.

2. Define the test-retest reliability of temporal synchrony perception, and more specifically study both the inter- and intrasubject variability on temporal synchrony measures in relation to different linguistic categories.

3. Assess perceived loudness of the speech signal at different levels of linguistic complexity and different temporal asynchronies, offering information about a continuous orthogonal measure to aid in the understanding of prelinguistic and linguistic components at different stages of integration.

4. Compare AV speech intelligibility performance to earlier acquired measures (i.e., temporal synchrony perception and loudness perception) at different levels of linguistic complexity and temporal synchrony. This allows interpretation of how prelinguistic information like temporal synchrony relates to speech intelligibility, how linguistic complexity relates to speech

intelligibility, how an orthogonal feature like loudness perception relates to speech intelligibility, and how individual differences in AV speech perception benefits can be explained by performance on any of these previous tasks.

Finally, Chapter 6 provides a future outlook of work that should be investigated, following from this dissertation work.

Chapter 2

Setting the Stage: Considerations for Remote Measures of AV Integration

In recent decades, technological advancements have significantly expanded the scale and breadth of academic research (Hewson et al., 1996; Reips, 2001, 2002; Duffy, 2002; Kraut et al., 2004). Notably, the year 2020, marked by the COVID-19 pandemic, witnessed a substantial increase in the exploration of remote research methods. However, methodological adaptation procedures, particularly in developmental research, for the adoption and validation of remote assessments, are often described as immature (Scott and Schulz, 2017; Nussenbaum et al., 2020; Rhodes et al., 2020). In light of this, we delve into the complexities of remote work, especially concerning children, by delineating a set of considerations (Section 2.1), and we investigate the effectiveness of utilizing comparison measures, in contrast to absolute measures, to tackle challenges arising from the inherent lack of control in unmoderated remote testing scenarios (Section 2.2).

2.1. Designing virtual, moderated studies of early childhood development

Section 2.1. is part of a larger publication (Gijbels, Cai et al., 2021) with co-authors Ruofan Cai, Patrick M. Donnelly, and Patricia K. Kuhl, and is integrated here with approval of all co-authors and according to the copyright agreement of Frontiers in Psychology.

2.1.1. Introduction

Remote, digital modalities have been recognized as viable substitutions for in-person research settings (Reips, 2001, 2002; Sheskin and Keil, 2018) and advantages

associated with online research, in comparison to lab-based research methods, have been reported. Some benefits include reduced operating costs, increased access to diverse populations, and reduction of experimenter effects (Reips, 2002). Accompanying the recent rising trend of remote research practice, these advantages make it possible to envision a future of advanced remote methodologies for developmental work.

However, shifting from in-person to remote modalities is not without its challenges. For example, Reips (2002) noted that experimental control may be difficult for certain study designs, and that attrition is often a concern for online research. Although potential solutions have been proposed in response to these challenges (Reips, 2002), observations and conclusions regarding virtual studies remain to be systematically documented. The success of remote adaptations of the three studies described in Gijbels, Cai, and colleagues (2021) shared certain commonalities regarding informed consent, study designs that attract and hold children's attention, and valid data collection procedures. Through this work, we suggest guiding principles for future facilitation of online developmental research.

2.1.2. Considerations for remote developmental assessments

2.1.2.1. Informed consent and privacy

All research involving human subjects needs to be concerned with ethical decisions such as obtaining informed consent, ensuring confidentiality, and establishing methods for data security (Whitehead, 2007). Even in traditional laboratory settings, consent procedures require deliberate and thoughtful actions. In language suitable for the intended individual, informed consent/assent should communicate the study's purpose and procedures, associated benefits and risks, confidentiality, safety, etc. Additionally, when appropriate, the researcher or caregiver may verbally communicate the informed consent to the child, as ethics of non-therapeutic research

involving children are a delicate issue as children are vulnerable and would likely not benefit directly from participation (Lambert and Glacken, 2011).

In general, remote consent procedures can take place over secure online portals. But the downside of solely obtaining (electronic) signatures online, is the lack of explicit opportunity for participants/caregivers to raise questions and/or concerns. We recognize that it is important to consider consent acquisition as a process rather than a product (Whitehead, 2008), especially when children are involved. Therefore, we posit that a valuable step to take is to ensure participants' and caregivers' understanding of informed consent through researcher moderation. This can serve to supplement written consent procedures or can occur as a separately documented process to replace text-based consent forms.

2.1.2.2. Caregiver involvement

For child studies in laboratory settings, caregiver involvement is often minimized, and equipment manipulations and task instructions are offered by the researcher in a standardized way, equal for all research participants. The degree of caregiver involvement for remote research is typically determined by the age group and the complexity of equipment manipulation. Involving caregivers of younger participants required intentional efforts to ensure that they followed the research protocol closely to avoid introducing unwanted interference. Clear communication of research protocols prior to the appointment is crucial in establishing desired caregiver involvement. Additionally, we experienced that it was helpful to provide families visualizations of experimental procedures or scripts of approved caregiver encouragements. Therefore, in addition to a carefully designed protocol, we believe that these steps could help minimize the confounding risk of caregiver interference. Although the level of caregiver involvement differed by age, technical support was critical for all three studies described in Gijbels, Cai et al. (2021). When active manipulation of technical devices (e.g., mouse clicks) is required by the children, it

can be helpful to objectively assess technical proficiency of the child during a training session, and based on the outcome, decisions can be made regarding caregivers' assistance in technical manipulations.

2.1.2.3. Control of stimulus quality

Moving our studies online required substantial adjustments in stimulus presentation and experimental setup. During data acquisition, it is crucial to generate and deliver consistent stimuli across subjects. However, in remote studies containing visual and auditory stimuli, it is more complicated to ensure this. Generating and delivering testing materials using experiment builders would be a favorable option as the automation of stimulus delivery has been reported to reduce the workload of the researcher during the task, lowering the chance of human error (Rhodes et al., 2020).

Another helpful measure to ensure experimental control for online developmental studies is researcher moderation. Although most online behavioral procedures can be automated, it is beneficial to control for unexpected changes in the environment, allowing for impromptu adjustments and extra technical support. We suggest from the findings in the AV studies that moderation could help improve participants' attention. The researcher can be aware of any decline in participants' attention and suggest a break or introduce adequate motivators. Additionally, researcher moderation allowed participants and their caregivers to ask questions during the consent procedure and ensured that no data would be lost due to invalid consent/assent procedures. Finally, we believe that the personal connection we established with the participants through moderation was beneficial to lowering the attrition rate and helped sustain participants' attention.

2.1.2.4. Validation of remote assessments

Last but not least, due to the variability and complexity of study designs in developmental research, validation of online methods in this field often stays specific to each study. We believe a potential solution may be to carry out a study design

both in person and remotely during the initial pilot phase, and assess the validity of the online study design by comparing pilot results. Moreover, when designing an online study or converting an in-person study to virtual environments, it is consequential to identify areas of adaptation and define the purpose of each adaptation. Meticulous deliberation and systematic documentation of such decisions would maximize the comparability between data collected in person and remotely and could benefit future replications of the study within or between labs .

2.1.3. Conclusion

We believe that by adjusting developmental research methods from traditional in-person settings to an online format and by acknowledging all the changes needed to be made, our developmental work was equally valuable to in-person work. All children could participate from a familiar environment at a time that worked for both them and the researcher, without having to make concessions. Testing from home can positively impact general attention and comfort for children. In our observations, many of our participants wanted to share their world (e.g., toys, pets) with the researcher, and were highly motivated to participate. Additionally, we recognize that part of the reason for the ease of our recruitment and the high compliance from our participants could be that we had established strong rapport with most of the participants and their caregivers from previous in-person studies.

All experimental control that would be routine in a lab environment had to be reevaluated and adjusted for online testing, which led to carefully considered and documented protocols. As our observations and results showed consistency over participants' and home environments, we believe we succeeded in tackling what we initially observed as the most challenging parts of remote developmental work. This goes from finding platforms and technical support to move the experiment online, to control of the participant's environment and even logistical issues.

2.2. How moderation affects remote psychophysical tasks with children.

Section 2.2. is published as Gijbels & Lee, 2023 in JASA Express Letters. The integration of the full manuscript into this dissertation was approved by both authors and the editorial team.

Abstract

The increasing use of remote platforms for auditory research necessitates more in-depth evaluation of assessment protocols, especially when working with children. This work investigates the influence of the presence of a moderator on remote audiovisual speech perception studies, by assessing how moderation impacts children's understanding and performance of the psychophysical tasks, as well as their attention on these tasks. In sum, both moderated and unmoderated methods can reliably assess audiovisual speech perception benefits. However, regardless of similar error patterns between both studies, unmoderated online studies with children are prone to more general attention lapses as suggested by higher overall error rates.

2.2.1. Introduction

Internet-based research is not new. The dawn of the internet ushered the first wave of remote data collection (Reips, 2002). Initially, online surveys gained popularity (Sills & Song, 2002), followed by increased interest for online assessments of psychology-related topics with therapeutic purposes (Stasiak et al., 2016). Other disciplines remained more attached to in-person assessments. For auditory-related research, the recent COVID-19 pandemic accelerated interest in online research (Peng et al., 2022).

Remote research can be beneficial when executed thoughtfully. It allows data collection of large sample sizes (Rhodes et al., 2020) in real-world environments while potentially overcoming geographical limitations, with opportunities for increased ecological validity and generalizability. Participating from the comfort of the individual's home can reduce stress and costs (Reips, 2002), as well as enhance opportunities for longitudinal or training studies (Whitton et al., 2017).

Nonetheless, conducting research online is not without challenges – calibration of stimuli can be difficult and attrition is often a concern for remote research (Reips, 2002). There is even more hesitance when adapting studies with children for remote testing, perhaps due to the limited peer-reviewed methodological reports benchmarking the success of these remote adaptations in developmental research protocols. Behavioral research involving child participants requires age-specific considerations related to task design (Desjardins et al., 1997), attention span (Thillay et al., 2015), immature auditory attention skills (Buss et al., 2011), cognitive load (Gibson & Twycross, 2008), language development (Bates et al., 1995), and general perceptual development (Moore, 2002). Limited face-to-face interaction, minimal relationship building with a remote researcher, the abstractness of online rewards, and the limited opportunity for physical breaks in combination with longer screen times makes it even more complex to control all these aspects when moving developmental research online.

The presence of a researcher moderating a study can also influence the results. For example, a researcher can unintentionally or unconsciously present unwanted verbal or non-verbal cues, or even create stress for the participants that lowers performance just by their presence. This phenomenon, known as the researcher effect, can be easily reduced in remote studies (Rhodes et al., 2020). However, limited presence also confines aspects the researchers can control. For example, readily assessing engagement of the participant or understanding of the task becomes more complicated when the researcher is not present. Therefore, when designing a

developmental study with remote data collection, the degree of moderation may need to be carefully considered. This degree of moderation can range from completely unmoderated (i.e., no researcher is present and all the instructions and guidance are provided by computer prompts), to completely moderated (i.e., the researcher is constantly present via a video-conferencing tool giving the instructions as they would for in-person research and guiding the participants through the tasks), to anything in between.

Unmoderated studies can increase across-subjects consistency and remove potential experimental bias due to the instructions being delivered in a fully automated fashion. The data collection phase is also less work-intensive and can be easily scaled to a greater number of subjects. However, it can increase the number of invalid data points, especially if the child does not sufficiently understand the task. Additionally, the absence of researcher supervision can increase technical errors (e.g., corrupted recording files – Scott & Schulz, 2017), and in general it does not create space to address questions or concerns of the participants and their caregivers (Gijbels, Cai et al., 2021 – See Section 2.1). In contrast, moderation allows for improved task understanding, and for immediate experimental adjustments to better suit the testing environment. Children often reside in sensory stimulating rooms that can be distracting during data collection. The presence of a remote researcher can anticipate and mitigate these problems. However, especially when working with young children who struggle with unfamiliar surroundings, introducing a stranger (i.e., the moderator), even online, can be intimidating and potentially interfere with the validity of data collection (Rhodes et al., 2020). Additionally, not only the child's but also the caregiver's behavior can unwillingly or unconsciously change by the presence of a researcher. Moderated or unmoderated remote studies therefore might introduce different additional confounds to the original research question that warrants care in the interpretation of the study results, and especially comparing across studies.

The current work quantifies how moderation can affect remote psychophysical tasks with children. We (Chapter 3, Section 3.1 and 3.2) completed two highly similar remote AV speech perception studies in children (age 6-7.5 years)—one moderated and one unmoderated—to address how children use AV speech perception cues to improve their speech understanding in noise. Specifically, the paradigm was designed to answer the question: "How does it benefit children when visual cues (mouth movements) are added to a degraded auditory speech signal (speech-in-noise)?" Both studies looked at improvement of speech recognition scores, and whether the use of speech-specific visual articulatory cues was necessary to accomplish this AV speech benefit. Cognitive and linguistic influences were minimized by developing an age-appropriate task that requires very little cognitive load, and all vocabulary used in these studies should be acquired by age 3 – 5 years (Holt et al., 2011). The details of the specific research questions in both studies, as described in Chapter 3, varied but are not relevant for the further analysis described here.

Both studies included measures to look at different aspects of attention. This allowed the researchers to investigate how attentional factors influence these perceptual skills and whether this is different based on the level of moderation. Attention skills are generally immature for children between 6 - 7.5 years old (Buss et al., 2011) and the level of attention varies depending on how appropriately designed these tasks are for the age group (Desjardins et al., 1997). By monitoring both their general attention level (i.e., the task engagement within every trial) as well as their cross-modal attention (i.e., the children's ability to attend to the visual modality while they are performing this speech task), different aspects of attention influenced by the researcher being present for this age group can be studied.

In sum, this work aims to address three questions when comparing moderated and unmoderated studies:

- What is the influence of moderation on performance (audio-only, visual-only, and audiovisual) of an online speech perception task with children?
- Does moderation influence general attention for children performing an online task?
- Does moderation influence cross-modal attention for children performing an online task?

2.2.2. Methods

A description of the study design is provided below, highlighting the differences between the moderated and unmoderated studies. For a more detailed explanation of participants, stimuli, recordings, procedures, and analysis methods we refer to Section 3.1.

2.2.1.1. Participants

All participants were children, with reported typical speech, language, auditory, and visual (normal or corrected to normal) development. They were recruited through a pre-existing database at the Institute for Learning & Brain Sciences,or through the Communication Studies Participant Pool at the University of Washington. Parents of all participants provided written informed consent under a protocol that was approved by the University of Washington's Institutional Review Board.

A. Moderated study

The participants (N = 37) of this remote moderated AV speech perception study (Section 3.1), were a subset of the 48 children who participated in a previous study at the University of Washington (Yeatman et al., 2022). This subset of participants (M = 15, F = 22) had English as a first language, were in first grade and were between 6.29 and 7.36 years of age (μ = 6.74 years, SD = 0.27), at time of participation.

B. Unmoderated study

For the purpose of comparing both moderated and unmoderated studies we selected a subgroup of 46 (M = 23, F = 23) children with a similar age range (6.04 – 7.44 years, mean = 6.73, SD = 0.44) from a group of 161 English-speaking children (M=79, F=82), age 4 to 15, who participated in the remote unmoderated audiovisual speech perception study (Section 3.2).

2.2.1.2. Stimuli and Materials

The psychophysical tasks used in both moderated and unmoderated studies measured auditory, visual, and AV speech perception. All target words were one-syllable consonant-vowel-consonant (CVC) words from the Lexical Neighborhood Test (LNT – Kirk et al., 1995) recorded by a female native English speaker (Holt et al., 2011), embedded in speech-weighted noise, presented at a comfortable suprathreshold level (for more details: Section 3.1).

We used a one interval four alternative-forced-choice (1I-4AFC) task, consisting of the CVC word (1I), followed by four images (4AFC; presented 2 x 2): the target (e.g., sun) and three carefully chosen foils: (a) a "minimal pair" foil, which differed in first or last consonant from the target word (e.g., run), (b) a "vowel" foil, which shared only the same vowel with the target word (e.g., gum), and (c) a "random" foil, which had no shared phonemes with the target word (e.g., pink). This "random" foil was added to measure general attention as all children should receive sufficient information from the audio-only trials to not pick this answer option. All the minimal pairs were auditory minimal pairs; thus, they are distinctively different acoustically but not necessarily visually (i.e., articulatory movement of the mouth).

Audiovisual trials were a combination of the auditory recording of the target word, with the accompanying video of the speaker producing the word. Audio-only trials were auditory recordings of the target word in noise paired with a still image of the same female speaker with a neutral expression, and visual-only trials were

essentially silent videos. Speech-weighted noise was present in all three modalities. There was no visual noise present in any modality.

A. Moderated study

All participants completed two speech-in-noise recognition sessions, which differed by their target stimulus set sizes (i.e., 25 vs 10). As the small stimuli set was a subset of the large set and our analysis (Section 3.1) yielded no distinction between these two sessions, we will only discuss the combined results. The stimuli were tested in two noise levels, -2 dB signal-to-noise ratio (SNR) and -8 dB SNR.

B. Unmoderated study

The participants completed one speech-in-noise recognition session, with 20 different spoken target words. In this unmoderated study, the target words were presented in three different noise levels: -5 dB, -8 dB, and -11dB SNR. These levels were chosen based on the expanded age range for this study and analysis of the moderated study (Section 3.1).

Although the two studies were not identical and consisted of different conditions (i.e., noise levels and set sizes), to accurately compare both studies we only used trials from both studies that were similar. This means we limited our comparison for all further analysis and discussion to (congruent) audiovisual speech trials and audio-only trials that were tested in -8 dB SNR. The visual-only trials were very similar in both studies. The number of trials presented in each study for each condition slightly differed (i.e., 170 vs 200 trials).

2.2.1.3. Procedures

In both study types, children were seated at a table facing the computer screen, and a caregiver was asked to sit next to the child participant during testing. Depending on each child's computer fluency as assessed during the training session, answers were collected via mouse clicks either completed directly by the participant or by the

parent when the participant pointed to the screen. Before testing and training, participants were instructed to set the computer's audio output to a comfortable level based on a repeating speech fragment of a cartoon character, and auditory stimuli were presented in free field through the computer's speakers. Participants completed a training session consisting of two phases (for more details: Section 3.1). In phase 1, all children would be exposed to all used words (target and foils) and in phase 2 children would complete some example trials of the actual task.

Three trial types (audio-only, visual-only and AV) were assessed at the different SNRs, in random presentation order. There was no time limit to respond, and no feedback was provided. Participants were offered a break after every block, and they could continue when ready. On average task completion would take up to 20-30 minutes, depending on the duration of breaks and response times.

A picture of a cartoon character was introduced with a double function; i) it was a fun motivator for the participants throughout the studies and ii) served as a catch-trial stimulus to measure cross-modal attention throughout the experiment. In these catch-trials, the cartoon character would randomly pop up on the screen and children had to respond verbally (in the moderated study) or manually by pressing an additional button on the screen (in the unmoderated study), to confirm that they noticed the cartoon and were still visually engaged to the task. For these trials no response to the presented speech stimuli from the children was measured.

A. Moderated study

The moderated study was created via a free online study builder. The tasks were emailed to the caregiver, and they were directed to load the tasks on their computer directly. The study was moderated via a videoconferencing session that was simultaneously set up to guide caregivers through the experiment (e.g., loading the task file, collecting the data, and sending the output file back), and to make notes of any unforeseen circumstances (e.g., internet issues, sibling intervening). The

computer screen was shared with the moderator. The child's understanding of the training was assessed with the moderator to assure all words were part of the child's vocabulary (phase 1) and the child had an understanding of the task setup itself (phase 2). At the beginning of the training the caregiver was informed by the moderator about their role during the experiment.

Children completed 200 trials, broken up in a longer (110) and shorter (90) session, with a total of 10 blocks of trials. Randomly dispersed throughout all trials, the catch-trial cartoon character appeared 20 times in total. Participants could take breaks between blocks in every session, but critically the moderator had the opportunity to suggest additional breaks in the middle of trial blocks when they noticed interest in the task declined.

B. Unmoderated study

The unmoderated study was created via the same free online study builder; however, the experiment was instead hosted via an online platform created to independently run remote studies. Parents would navigate to the experiment via a web browser link and responses were automatically stored online. In this unmoderated study, parents were provided with clear instructions via email with no researchers present during the assessment of the task. Like the moderated study there were two training phases; however, image understanding from phase 1 was not explicitly assessed. Children completed 160 trials, broken up into 8 blocks. The remaining 10 trials were catch-trials.

2.2.3. Analysis, results and discussion

2.2.3.1. Moderation and Overall Task Performance

The first step was to validate that our findings in both remote studies (moderated and unmoderated) are comparable to results obtained in other AV speech perception tasks in children obtained in traditional in-person settings. Accordingly, these

children were expected to i) score above chance level even on the most difficult visual-only trials (Kyle et al., 2013) and ii) to perform best in the AV trials, followed by audio-only then visual-only modality. Performance was analyzed in all modalities with a one-tailed t-test against performance at chance level (i.e., > 25% in these 1I-4AFC tasks), for both the moderated and unmoderated studies, with Bonferroni correction for each modality per study. Results (Table 2.1) show that children scored significantly above chance level for all modalities in both studies (all $p < .001$) and that they performed best in the AV trials, followed by the audio-only and visual-only trials (Fig 2.1)

TABLE 2.1. Comparison of perceptual performance in a moderated and unmoderated AV speech perception task.

Study	Modality	Mean score (% correct)	t-value	p-value
Moderated	AV	84	61.30	< 0.001
Unmoderated	AV	79	37.53	< 0.001
Moderated	Auditory	81	44.10	< 0.001
Unmoderated	Auditory	71	36.95	< 0.001
Moderated	Visual	44	7.82	< 0.001
Unmoderated	Visual	38	4.79	< 0.001

One-tailed t-test of auditory, visual, and AV modality for both the moderated and unmoderated task, against a 25% chance performance.

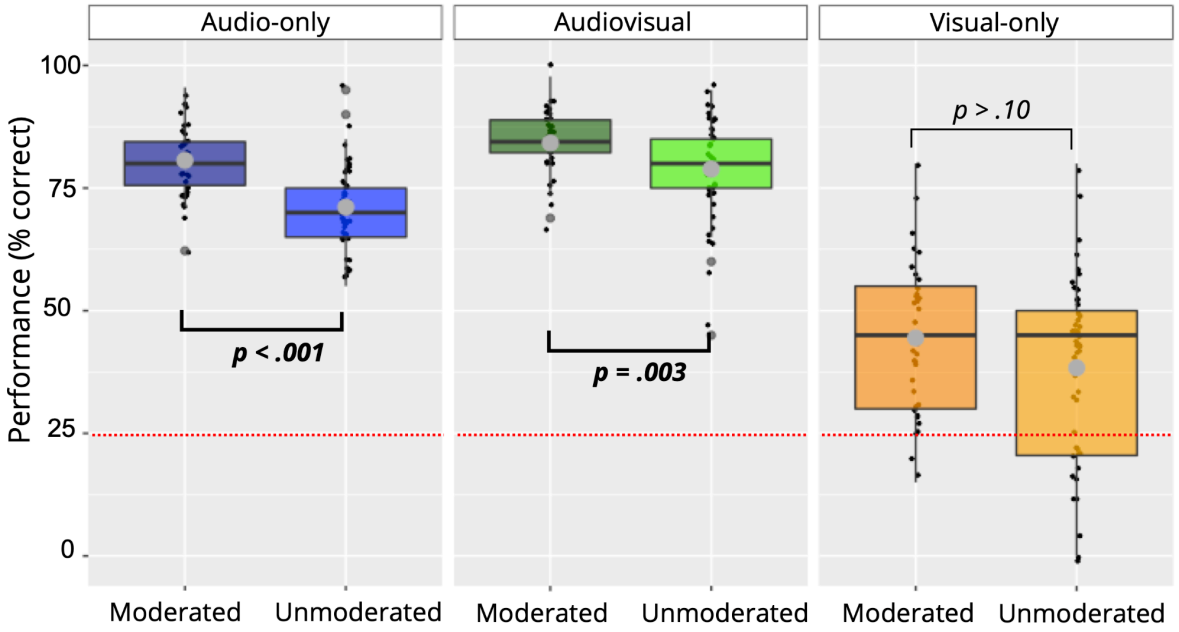
Children this age are expected to perform significantly better in AV trials than in audio-only trials (Ross et al., 2011). These AV speech perception benefits can be expected for monosyllabic words presented in moderate background noise (e.g., -8 dB SNR – Ross et al., 2011). A one-tailed t-test shows that for both moderated (μ = 3.54, t_{36} = 3.11, p = 0.001) and unmoderated studies (μ = 7.74, t_{45} = 4.62, $p < 0.001$), there are significant AV speech perception benefits. Results so far are in line with

findings of other in-person studies (Kyle et al., 2013; Ross et al., 2011), indicating the validity of both moderated and unmoderated online studies with young children.

Next, we explored how moderation influenced the online performance of these studies with children. As visualized in Fig 2.1, the mean performance is lower in the unmoderated study for every modality. Two-tailed t-tests comparing both studies per modality show significant differences for the audio-only ($t_{79.80} = 5.37$, $p < .001$) and AV ($t_{75.60} = 3.08$, $p = .003$) trials, but not for the visual-only trials ($t_{80.99} = 1.63$, $p > .10$). A Type III ANOVA of our linear mixed model (*Model 2.1.: Performance ~ modality * study + (1|subject)*), describes a significant main effect of modality ($F_{175.44} = 361.38$, $p < .001$), a significant main effect of study ($F_{200.75} = 17.14$, $p < .001$), but no significant interaction effect ($F_{175.44} = 0.96$, $p > .10$). Therefore, we conclude that performance is overall lower in the unmoderated study, but the relationship between modalities is not significantly different for the moderated or unmoderated studies.

Overall, we find that children in the moderated study tend to do better in all modalities (Fig 2.1). This difference in performance between both studies is significant in both the auditory and audiovisual modalities. This suggests that a moderator's presence significantly influences the study results (Rhodes et al., 2020), where children perform better when a moderator is present. However, the AV speech perception benefit — a difference score in performance ([(AV-A)/(1-A)]), and our a priori measure of interest—is not significantly different for the moderated or unmoderated studies. This is the result from a similar decline in audio-only and AV performance in the unmoderated study and suggests that whether moderation influences remote studies depends on the chosen outcome measure (i.e., performance vs. difference in performance).

FIGURE 2.1. Boxplots of performance (in % correct) per modality (Audio-only (-8 dB SNR), Audiovisual (-8 dB SNR), Visual-only) per study (Moderated, Unmoderated).



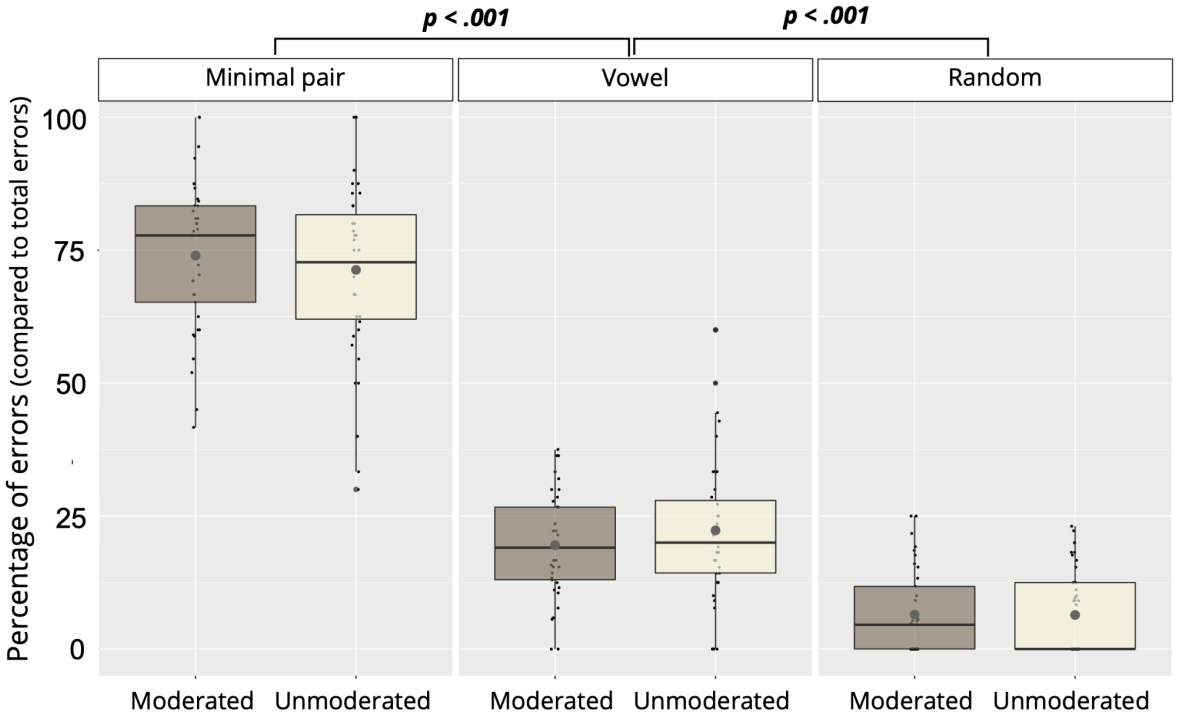
The red dotted line indicates chance performance (25%). Differences between studies are indicated with p-values. Thick horizontal lines represent medians, gray dots represent means, boxes represent interquartile ranges, and whiskers represent the range, excluding outliers. Individual scores are visualized as small black dots.

2.2.3.2. *Moderation and General Attention*

Analyzing the errors made in both moderated and unmoderated studies provides an indication of general attention during the online studies. Both moderated and unmoderated studies use the same 1I-4AFC tasks. As explained in the Methods section, a random foil word was chosen as one of the four alternatives to monitor the participant's general attention. Given the moderate SNR level used (-8 dB), it would be expected for both the audio-only and the AV trials that children receive sufficient auditory information to avoid choosing the random foil. However, when working with children on studies consisting of 170 to 200 trials, limited attention span (Thillay et al., 2015) resulting in some random responses is not surprising. Since all participants responded above chance level, we interpret that children would pick the random answer only due to a lapse of general attention, and not due to them not understanding the task. For both moderated and unmoderated studies, we find that,

as expected, when errors are made in the audio-only or AV trials, most of them are closest to the target option (minimal pair errors), followed by vowel errors and only a limited number of random errors (Fig 2.2).

FIGURE 2.2. Boxplots of the percentage of errors compared to the total amount of errors made, for all types of error options (Minimal pair, Vowel, Random) by study (Moderated or Unmoderated).



Thick horizontal lines represent medians, gray dots represent means, boxes represent interquartile ranges, and whiskers represent the range, excluding outliers. Individual scores are visualized as small black dots.

To address whether these moments of inattention are significantly different when assessing a moderated or unmoderated online study with children, the percentage of errors are compared to the total errors. There are limitations in just comparing the number of errors itself, since children in the unmoderated study made significantly more errors than in the moderated study. An ANOVA of the linear model (*Model 2.2.: Percentage of errors to total errors ~ study * error type*) was used to quantify the main effects of study and error type, and to see if an interaction effect is present indicating a relationship between the error types and moderation used in the study. There is a main effect of error type ($F_{1,234} = 617.59$, $p < .001$), with having significantly

more minimal pair errors than vowel errors (t_{234} = 18.74, p < .001) and significantly more vowel errors than random errors (t_{234} = -4.49, p < .001; see Fig 2.2). There is no main effect of study, and no interaction effect. This suggests that although children made overall more errors in the unmoderated study, when looking at the types of errors compared to the total errors made, there was no difference in error patterns between the two studies. More specifically, when we look at the random errors, the mean performance in the moderated and unmoderated studies were the same (~ 6%).

Thus, when looking at absolute error rate, children made more random errors associated with a lapse of general attention in the unmoderated study, but despite this lower performance, the error pattern stays the same indicating that general attention is not different when assessing an online audiovisual speech perception study that is moderated or unmoderated.

2.2.3.3. Moderation and Cross-modal Attention

Our third measure of interest is cross-modal attention. Cross-modal attention is described by measuring how often (in %) a child finds the visual catch-trial throughout the task. Although children were trained to find this catch-trial in phase 2 of the training for both studies, there was a difference in task setup. There are some intrinsic differences between the two studies: the number of catch-trials (20 vs 10), in how children reported the catch-trial (verbally vs manually) between studies, and the moderated study allowed for a repetition of instructions because there were two sessions, whereas the unmoderated study has just one.

Descriptive statistics of these catch-trial counts show that there is a discrepancy in mean performance between the two studies (μ_{mod} = 95.68%, μ_{unmod} = 67.39%). As shown by a two-tailed t-test, the mean score of the catch-trial count is significantly lower in the unmoderated study ($t_{46.72}$ = 4.45, p < .001). Interestingly, no participant finds less than 80% of the catch-trials in the moderated study, whereas a fifth of the

participants (N = 9/46) do not find any catch-trials (0 %) in the unmoderated study, and six participants find 50% or less. This discrepancy suggests that it might be more meaningful to compare the medians ($Q2_{mod}$ = 95%, $Q2_{unmod}$ = 90%) than means (μ_{mod} = 95.68%, μ_{unmod} = 67.39%). Comparing the median catch-trial counts (by using a Wilcoxon rank sum test) in both studies indicates no statistically significant difference (W = 653, p > 0.05), and shows there is a big discrepancy in catch-trial performance in the unmoderated study (either scoring very low, or very high).

We explored further into this group of children (N = 14) who had low catch-trial counts (< 50%). By comparing the performance on the task itself (audio-only, visual-only and AV trials) of the children who have a catch-trial count <50% and >50% in the unmoderated study we did not find any significant differences in performance in any of the modalities (audio-only: $t_{19.45}$ = -0.57, p > .10; visual-only: $t_{11.27}$ = 0.73, p > .10; AV: $t_{17.47}$ = 0.23, p > .10).

From both moderated and unmoderated studies, we conclude that the cross-modal attention measure is not sufficiently sensitive. In the moderated study, the catch-trial count was too high to correlate with performance, and in the unmoderated study the catch-trial count did not correlate with performance on the speech perception task at all. This is surprising since visual task-irrelevant catch-trials are often used to check attention in-person studies (Yeatman et al., 2022). Adding extra focus on the utility of these catch-trials as well as on task understanding for children in online studies in future work could clarify these findings.

2.2.4. Conclusion

The interest in remote studies is burgeoning in the auditory field, but validity measures of remote studies remain scarce. To limit confounds such as attention span, cognitive load and general perceptual development, remote moderation has been suggested especially when working with children (Gijbels, Cai et al., 2021). The current work evaluates the influence of moderation on i) remote AV speech

perception studies, ii) potential differences in attention, and iii) task understanding in moderated and unmoderated psychophysical tasks with children.

First, we conclude that both moderated and unmoderated remote audiovisual speech perception studies can be valid. Nonetheless, there is a significant influence on performance from the virtual presence of a researcher. Selecting the outcome measure as a difference score, rather than absolute performance could limit moderation effects. Second, a higher number of general attention errors are found in the unmoderated study, but they do not influence the observed error pattern. Therefore, both studies provide us similar information regarding the use of auditory cues to make decisions about the presented stimuli. Finally, cross-modal attention measures might not be as sensitive when assessed online as for in-person studies, and the presence of a moderator is important for task understanding. This points to the need to find other valid ways to assess task comprehension as there is no room for asking questions or re-iterating instructions in unmoderated studies.

Chapter 3

Exploring Influential Factors in the Developmental Trajectory of AV Speech Integration

Following the establishment of the feasibility of remote AV speech perception measures in Chapter 2, this chapter delves into two studies encompassing a total of 198 unique participants age 4 -15. The primary objective is to advance our understanding of the developmental trajectory of AV speech perception benefits. The initial study (Section 3.1, N = 37) functions as a proof-of-concept for an AV speech recognition task in 6-year-olds, designed to minimize cognitive and linguistic developmental influences on task performance. This investigation explores the impact of the visual salience of phonemes on AV speech perception benefits in first graders and examines whether individual differences in AV speech enhancement can be linked to vocabulary knowledge, phonological awareness, or general psychophysical testing performance. Building upon this proof-of-concept, the second study (Section 3.2) extends and expands the task to a broader cohort of children (N = 161) aged 4 to 15-years-old. The emphasis is on clarifying the developmental trajectory of AV speech perception in relation to (1) the use of linguistic information presented by phoneme-viseme connections, (2) the perceptual weighting of individual modalities in predicting AV speech perception benefits on an individual level, and (3) potential differences in the Causal Inference process.

3.1. Audiovisual speech processing in relationship to phonological and vocabulary skills in first graders

Section 3.1 is published as Gijbels et al. (2021) in the Journal of Speech, Language, and Hearing Research with co-authors Jason D. Yeatman, Kaylah Lalonde, and Adrian KC Lee. Permission to integrally re-use the published manuscript was

requested and obtained via the American Speech-Language-Hearing Association Copyright Clearance Center and all co-authors agreed.

Abstract

It is generally accepted that adults use visual cues to improve speech intelligibility in noisy environments, but findings regarding visual speech benefit in children are mixed. We explored factors that contribute to audiovisual gain in young childrens' speech understanding. We examined whether there is audiovisual (AV) benefit to speech-in-noise recognition in children in first grade and if visual salience of phonemes influences their AV benefit. We explored if individual differences in AV speech enhancement could be explained by vocabulary knowledge, phonological awareness, or general psychophysical testing performance.

Thirty-seven first graders completed online psychophysical experiments. We used an online single-interval, 4-alternative forced-choice picture-pointing task with age-appropriate consonant-vowel-consonant (CVC) words to measure auditory-only, visual-only, and AV word recognition in noise at −2 dB and −8 dB SNR. We obtained standard measures of vocabulary and phonological awareness and included a general psychophysical test to examine correlations with audiovisual benefits.

We observed a significant overall AV gain among children in first grade. This effect was mainly attributed to the benefit at −8 dB SNR, for visually distinct targets. Individual differences were not explained by any of the child variables. Boys showed lower auditory-only performances, leading to significantly larger AV gains.

This study shows AV benefit, of distinctive visual cues, to word recognition in challenging noisy conditions in first graders. The cognitive and linguistic constraints of the task may have minimized the impact of individual differences of vocabulary and phonological awareness on AV benefit. The gender difference should be studied on a larger sample and age range.

3.1.1. Introduction

Daily listening environments are often noisy, and the presence of background noise degrades or slows down our speech understanding (Leavitt, Javitt, & Foxe, 2006; Mattys et al., 2012; Ross et al., 2006). Speech in noise is a frequently occurring problem in children for two reasons (Erickson & Newman, 2017; Neuman et al., 2010). First, children have immature perceptual, cognitive, and linguistic skills (Leibold & Buss, 2019; McCreery, Buss, & Leibold, 2020; McCreery et al., 2017). Second, they constantly interact in noisy backgrounds, e.g., in the classroom, on the playground, in the park, or at home (Knecht et al., 2002; Nelson & Soli, 2000). These frequently occurring experiences in noise can lead to lower performances in the classroom (Mealings et al., 2015).

Fortunately, oral communication is most often multisensory. Early pioneers such as Sumby and Pollack (1954) and Erber (1969) showed that seeing faces and accompanying articulation movements improves speech intelligibility significantly. This finding in adults has continuously been supported over time (see Grant & Bernstein, 2019 for a review). Audiovisual (AV) enhancement or alteration occurs for both non-speech (Hirst et al., 2020) and speech stimuli, including syllables (Lalonde & McCreery, 2020; Mcgurk & Macdonald, 1976), words (Ross et al., 2006) and sentences (Grant & Seitz, 1998; Lalonde & McCreery, 2020). It has been studied in multiple domains, namely temporal (Hillock-Dunn & Wallace, 2012; Lalonde & Werner, 2019; Maddox et al., 2015; Stevenson et al., 2012), spatial (Bishop & Miller, 2011), as well as manipulation of the visual (Campbell & Massaro, 1997; Rosenblum & Saldaña, 1996) or the auditory stimuli (Grant & Walden, 1996).

3.1.1.1. Audiovisual benefit in speech perception in adults and children

Audiovisual speech benefit is especially prominent when the auditory signal is degraded (Sumby & Pollack, 1954). Audiovisual enhancement in adults with normal-hearing thresholds can result in 6–15 dB threshold improvements (MacLeod

& Summerfield, 1987) or 30–50% increased accuracy of speech recognition (Binnie et al., 1974; Ross et al., 2006, 2011). However, the exact relationship between AV enhancement and increasing noise levels has been debated over time. Early studies in adults (Stein & Meredith, 1993; Sumby & Pollack, 1954) found that the more the auditory signal is degraded by noise (i.e., the lower the signal-to-noise ratio; SNR), the larger the effect of AV enhancement. More recent findings in adults (Ma et al., 2009; Ross et al., 2006) and children (Ross et al., 2011) show that enhancement of AV speech perception is the greatest at intermediate SNRs (−8 to −12 dB, depending on age).

The fact that AV benefit differs greatly as a function of SNR and age could explain some of the variability in AV gain in the literature, especially in children. AV speech improvement has been observed across the lifespan (Lalonde & Werner, 2021): in infants (e.g., Hollich & Jusczyk, 2005; Lalonde & Werner, 2019), children (e.g., Fort et al., 2012; Lalonde & McCreery, 2020; Ross et al., 2011), young adults (e.g., Barutchu et al., 2010; Ross et al., 2006), and older adults (Winneke & Philips, 2011). AV speech perception plays an important role in early language development. The language learning process takes place in complex listening conditions, but this does not prevent a child from successfully learning a language with immature auditory attention skills (Bargones & Werner, 1994; Buss et al., 2011). Infants and children use visual speech to make decisions about competing auditory signals (Knowland et al., 2016) and to learn phonetic categories in a language (Teinonen et al., 2008).

Compared to adults, children's AV speech processing is less dominated by visual input (Sloutsky & Napolitano, 2003), but how AV enhancement changes throughout development is still up for debate. On the one hand, AV enhancement has been found in 3- to 4-year-olds (Lalonde & Holt, 2016) with continuous increases over age (Fort et al., 2012). Other studies report AV benefits only starting later, around 9 years of age (Wightman et al., 2006). Jerger et al. (2009) even reported a U-shaped

relationship where AV enhancement presents in young (4-year-olds) and older (10- to 14-year-olds) children, but not in between (5- to 9-year-olds).

Ross et al. (2011) showed that this difference across ages is dependent on the SNR used. Specifically, at SNRs as low as −4 dB, there was no difference between 5- to 7-year-olds, 10- to 11-year-olds, and adults. However, at more negative SNRs, the difference between age groups increased in favor of the adults. Overall, the youngest group showed similar benefit across SNRs (−3 to −15 dB), whereas the older children and adults clearly showed a peak benefit at −12 dB SNR.

3.1.1.2. Intrinsic and Extrinsic factors in children's AV enhancement

Differences in AV enhancement between children and adults—and the large variability in AV enhancement among children—are likely explained by a combination of developmental factors (intrinsic) and experimental design factors (extrinsic). First, these intrinsic differences could be explained by phonological skills (Fort et al., 2012; Jerger et al., 2009, 2014). Jerger et al. (2009) suggested their observed U-shaped curve of AV benefit over age could be explained by the "Dynamic System Theory." This theory states that reorganization of phonological knowledge demands a disproportionate share of a child's limited processing capacity, to the extent that overloading available information processing resources can create an obstacle to processing visual speech (Jerger et al., 2009). This process of phonological reorganization can be expected in 6- to 9-year-olds, when children learn how to read (Jerger et al., 2009, 2014).

A second intrinsic child factor that could play a role is language development. Smaller AV benefits in children could also be explained by less linguistic experience compared to adults (Elliott, 1979; Fort et al., 2012; Jerger et al., 2009). A number of studies support this assertion. Fort et al. (2012) found evidence that children perform better on AV tasks when vowels are embedded in words compared to nonwords; therefore, lexical knowledge improves the children's AV performance. Davies et al.

(2009) found a significant correlation between receptive vocabulary and speechreading in young children. Sekiyama and Burnham (2008) showed the impact of language and language development on AV speech processing. They found that visual impact on speech perception was nearly absent in both Japanese and English 6-year-olds, stayed constant for older Japanese children, and increased for older English children. Cognitive skills, such as working memory and attention, may also account for some individual and age-related differences in AV enhancement, as working memory is correlated with individual differences in children's speechreading acuity (Lyxell & Holmberg, 2000). Speechreading is more cognitively demanding for children as they have not developed their cognitive skills to the same level as adults and therefore have to devote more of their limited processing capacity to the speechreading tasks (Lyxell & Holmberg, 2000).

Extrinsic factors as related to aspects of experimental design (Lalonde & Werner, 2021), such as procedures, stimuli, and cognitive and linguistic task demands (Bjorklund, 2005; Desjardins et al., 1997; Lalonde and Holt, 2015) could also explain variability in AV benefit in children. Task demands are a particularly important design parameter to consider for young children, as their performance differs between indirect (e.g., looking time) and direct tasks (e.g., formulating a response; Jerger et al., 2009), and between recognition and discrimination or detection tasks (Lalonde and Holt, 2016). Task demands are important, as different tasks may require different underlying mechanisms of AV enhancement (Lalonde & Holt, 2016; Lalonde & Werner, 2021).

The extrinsic factor of experimental procedure as it relates to using stimuli words of an open- or closed-set needs further considerations especially for studies in children. Speech recognition scores have been shown to be worse using an open-set response. When a closed-set task is used, results are dependent on the number of choices (Yu & Schlauch, 2019). Children are more often presented with closed-set tasks because these tasks are easier to understand and to execute, but they are fundamentally

different in terms of their information processing demands, especially when considering the potential responses. In an open-set task, the performance is determined by the size of the mental lexicon, whereas in a closed-set task, it will mainly be determined by the provided alternatives (Clopper et al., 2006). But the impact of the different amount of word stimuli used in a closed-set task in children, who are still developing their language skills, has not been well studied. A second important aspect of using a closed-set task is the number of alternative forced choices provided. Clopper et al. (2006) showed that spoken word recognition performance in an alternative forced-choice task is determined by the confusability of the foils used.

Finally, there is an interaction between extrinsic and intrinsic experimental factors. If psychophysical tasks in general require cognitive (and linguistic) skills (Witton et al., 2017), then developmental and individual differences in these skills will influence performance on any psychophysical task. Therefore, comparing psychophysical tasks that are similar in extrinsic experimental set-up but measure different intrinsic values would allow us to better focus on the experimental factors of interest.

3.1.1.3. Purpose of the current study

The purpose of the current online study is to explore these intrinsic and extrinsic factors that impact AV processing in first graders, which is a narrow age group at the bottom of the suggested U-curve (Jerger et al., 2009). The first question we asked was whether first graders show significant AV enhancement of speech in noise. As AV benefit differs over varying SNRs (Barutchu et al., 2010; Ma et al., 2009; Ross et al., 2006; Stevenson et al., 2015; Sumby & Pollack, 1954), we chose two SNR conditions: −2 dB and −8 dB SNR. We expected maximal benefits for children this age around the −8 dB SNR condition (Barutchu et al., 2010; Ma et al.,2009; Ross et al., 2006). While this question has been addressed in the past, test stimuli and methods ranged over a wide spectrum of cognitive and linguistic demands. In this study, we aimed to limit

the influence of articulation, language, task attention, and cognitive demands on AV speech perception performance in order to isolate children's AV enhancement skills. For example, multiple previous results were often obtained by using a word/sentence repetition task. Although an open-set task has benefits, we instead used a closed-set (four-alternative forced-choice; 4AFC) picture pointing task. The goal of this closed-set task is to constrain cognitive and linguistic (Jerger et al., 1968) task demands for these children. More importantly, this closed-set picture pointing task would not be impacted by any articulatory issues. Previous studies using open-set tasks could have had confounds because children do not always have fully developed articulation skills (Vance et al., 2005). We used consonant-vowel-consonant (CVC) words rather than multisyllabic words or sentences for two reasons. First, we picked a word set that is well known by typically developing children of this age (Holt, Kirk, & Hay-McCutcheon, 2011), therefore minimizing the impact of their linguistic abilities. Second, CVC words facilitated the use of specific foils, which allowed us to closely investigate visual salience of phonemes at this age (Baart et al., 2014; Lalonde & Holt, 2015; Lalonde & Werner, 2021). To probe whether the vocabulary set size used in an AV task interacts with our choice of using a closed-set response in this task, we tested two target stimulus set sizes.

We hypothesized that children at the bottom of the U-shaped curve (6- to 7-year-olds) would show significant AV speech enhancement on a task low in cognitive and linguistic demands. Further, we expected AV speech benefit to be larger in the −8 dB SNR condition, similar to results in older children and adults (Ross et al., 2011), and independent of the target stimulus set size, when accounting for vocabulary knowledge in the task.

Our second question focused on explaining intrinsic and extrinsic differences in AV enhancement. By limiting the cognitive and linguistic demands of the task, we aimed to isolate individual differences in AV enhancement. Earlier work suggests not only

relationships between AV benefit and intrinsic factors such as phonological awareness (Jerger et al., 2009, 2014), linguistic skills (Elliott, 1979; Jerger et al., 2009), and attention (Lyxell & Holmberg, 2000; Tye-Murray et al., 2011), but also extrinsic factors like task complexity (Bjorklund, 2005; Desjardins et al., 1997; Lalonde and Holt, 2011; Lalonde & Werner, 2021). Here, we tested whether these factors are related to AV enhancement in a closed-set paradigm, exploring whether the development of AV speech enhancement is necessarily tied to other developmental skills such as vocabulary and phonological awareness. By using target words that are typically acquired several years younger (3- to 5-year-olds; Holt, Kirk, & Hay-McCutcheon, 2011) than the current age of the children, we constrained linguistic demands of the tasks. This indicates that any relationship between individual differences in vocabulary and AV enhancement is due to a fundamental relationship between these underlying constructs, rather than due to difficulty of the vocabulary of the target stimuli.

Finally, we wanted to explore the correlation between phonological awareness skills and AV gain. Based on the Dynamic Systems Theory (Jerger et al., 2009, 2014), one would hypothesize a positive correlation between these two factors. However, a null result in this correlational analysis would suggest a need to revisit the model suggesting that a lack of linguistic experience (Elliot, 1979; Jerger et al., 2009) causes a dip in AV processing skills at this age.

In general, auditory psychophysical testing performance often improves with increasing age because these tasks rely on attention and short-term memory skills (Witton et al., 2017). We included an auditory psychophysical task that had the same response structure (i.e., pictures presented in a 4AFC format) without visual or speech stimuli. This task served to account for general psychophysical test performance. Given the relatively low cognitive and linguistic demands of this auditory-only task, we hypothesized that we constrained these tasks enough so that it would not show a common variance between the speech and non-speech task, and

thus, better performance in psychophysical tasks would not be a main factor in explaining AV performance.

3.1.2. Methods

3.1.2.1. Participants

A group of 37 English-speaking children (M = 15, F = 22) participated in this study. Although this study was initially planned in-person, it moved online due to COVID-19. These participants were a subset of the 48 children that participated in a ten-day camp in the summer of 2019 at University of Washington's Institute for Learning & Brain Sciences before they entered kindergarten. Parents of all participants provided written informed consent under a protocol that was approved by the University of Washington's Institutional Review Board. All participants showed typical speech, language, and hearing development, measured in their previous participation in 2019. This information was acquired both by parental report and by a behavioral task battery, including the Peabody Picture Vocabulary Test (PPVT-4), Comprehensive Test of Phonological Processing (CTOPP-2), and Test of Preschool Early Literacy (TOPEL). All participants demonstrated normal or corrected-to-normal vision, measured with the Snellen eye chart. At the time of recruitment, in 2019, according to parental report, no participants showed a history of neurological or auditory disorders. In January 2020, one participant was officially diagnosed with autism spectrum disorder, and therefore, excluded from the current study. The other ten participants dropped out after May/June 2020.

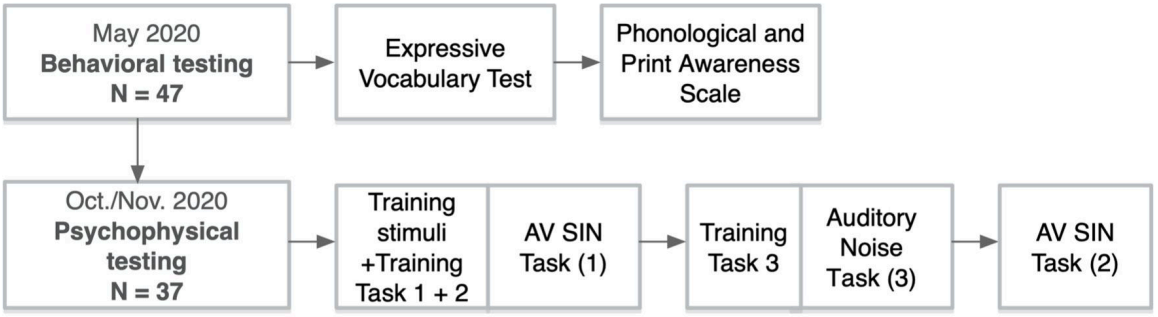
Participants were tested online in May and June 2020 on phonological awareness (Phonological and Print Awareness Scale; PPA; Williams, 2014) and expressive vocabulary skills (Expressive Vocabulary Test; EVT-3; Williams, 2018). The PPA is a 3AFC task testing initial sound matching (e.g., Which one begins with the same sound as…?), final sound matching, and phonemic awareness (e.g., How many sounds do we hear in the word …?). The EVT-3 is a vocabulary task in which

responses are expected to be expressed based on a picture with an accompanying question (e.g., What is this? Tell me another word for …?). These standardized and norm-referenced tests were administered by a trained research assistant. The psychophysical tasks were completed online between October and November 2020. All participants were in first grade and were between 6.29 and 7.36 years of age (M = 6.74 years in October/November 2020). All 37 children completed all psychophysical tasks and the training session.

3.1.2.2. Experimental Protocol

Fig 3.1 shows an outline of the study. Vocabulary and phonological awareness tasks (EVT-3 and PPA) were collected between May and June 2020 via videoconferencing. These tasks were administered as similar as possible to in-person testing. The participant (with parent) and research assistant both sat at a table in front of a computer with audio and video turned on. All original materials (PPA and EVT-3) were presented in accordance with the test manual but were delivered as a PowerPoint presentation over the videoconferencing platform. Since the PPA is a 3AFC task, participants were asked to name the word or the accompanying number (1, 2, or 3) for each stimulus. The participant could also opt to point at the screen, in which case the parent verbalized the response. Three psychophysical tasks were collected over a 1-hour online moderated session between October and November 2020. Each task took about 10–15 minutes when completed without breaks. While up to 5 breaks per task and a break between every task were offered to the participants to encourage their focus, most participants only took breaks between tasks. No participant took more than 2 breaks during a task.

FIGURE 3.1. Experimental protocol of behavioral and psychophysical test sessions.



The diagram describes the training and testing procedures. Task 1 + 2 represent the AV Speech-in-Noise (SIN) tasks, where Task 1 is the large stimulus set (N = 25) and Task 2 the small stimulus set (N = 10). Task 3 represents the non-speech auditory task that assesses general psychophysical testing performance. N denotes the number of participants in each test session.

Task 1 and 2: (AV) Speech-in-Noise recognition

Stimuli and materials:

All participants completed two speech-in-noise (SIN) recognition tasks. These two psychophysical tasks both used AV, audio-only, and visual-only stimuli. They differed only by their target stimulus set sizes. In Task 1, 25 spoken target words were used; in Task 2, 10 spoken target words were used. All target words were CVC words. The 10 target words in Task 2 were a subset of the 25 target words in Task 1. These words were drawn from the CHILDES database (MacWhinney & Snow, 1985) and used by Kirk et al. (1995) in the Lexical Neighborhood Test (LNT). All stimuli were professional recordings (audio + video) of a female native English speaker created by Holt, Kirk and Hay-McCutcheon (2011) and used by Lalonde and Holt (2016). All CVC words used were of low complexity, in both meaning and word form. They were imageable and are considered to be part of the vocabulary of 3- to 5-year-olds. We generated a 3-minute noise matching the long-term average speech spectrum (LTASS) of 100 LNT CVC-word recordings from the same female talker, by using PRAAT (Boersma & Weenink, 2021). Random 2-second fragments of the noise were mixed with target words at two SNR levels: −2 dB or −8 dB. These noise

fragments were added to a relatively constant speech signal. During development of the stimuli the speech signal was roved between 55 dB and 65 dB SPL and the noise was added, 2 dB or 8 dB more intense than the stimulus signal. The SNRs were chosen to make sure the conditions were not too easy (−2 dB) and would show maximum AV benefits (−8 dB; Barutchu et al., 2010; Ma et al., 2009; Ross et al., 2006). To minimize the effect of the absolute intensity level of the auditory stimuli, we roved the intensity of the CVC word presentation within a 10 dB range, +5 dB /−5 dB around 60 dB SPL (i.e., randomly drawn from a uniform distribution). Preliminary analysis of the data (in Task 1) confirmed no relationship between presentation levels (per target word and modality) and percentage correct scores (Pearson correlation: $R_{98} = -.005$, $p = .961$). During the training phase, SNR of 0 dB and −5 dB were chosen to make it easier for the participant to focus on learning the task.

All 38 pictures used for this one-interval 4AFC paradigm were retrieved from an open-source clipart database. Every set of pictures consisted of the target picture, and three foil pictures: 1) a "minimal pair" foil picture —target and foil picture names only differed in the first or last consonant, 2) a "same-vowel" foil picture—target and foil picture names shared only the same vowel, and 3) a "random" foil picture—target and foil picture names had no shared phonemes. Most targets served as foils for other targets (e.g., "hold" was a target, a minimal pair foil for target "cold", and a vowel foil for target "goat"). The word lists and associated foils can be found in the Appendix Table A of the original manuscript (https://doi.org/10.1044/2021_JSLHR-21-00196). All of the minimal pairs used were auditory minimal pairs; they are distinctively different acoustically, but not necessarily visual. For example, you can clearly hear the difference between "hold" and "cold" in a quiet environment, but it is hard to notice visual differences between someone articulating "hold" and "cold", in absence of any auditory input. In order to capture individual variability in the participants' benefit from visual speech, minimal pair words were scored according to their visual similarity to the target

(Appendix, Table A; https://doi.org/10.1044/2021_JSLHR-21-00196). Similarity was defined based on speechreading consonant confusion errors in previous studies with adults (e.g., Owens & Blazek, 1985), as there is little literature about visual identification of phonemes in children (Kishon-Rabin & Henkin, 2000). For example, word-initial /g/ and /θ/ are typically not confused during speech reading, so "gum" and "thumb" are low in visual similarity. In contrast, word-initial /h/ and /k/ are often confused, so "hold" and "cold" are high in visual similarity (Binnie et al. 1974).

AV stimuli were generated offline using ffmpeg software (Python 3.7) to ensure synchronicity between the auditory file and its corresponding silent video file. Video-only stimuli were essentially silent videos but with an LTASS noise presented in the background. Audio-only stimuli were paired with a still image of the same female speaker with a neutral expression. The task was designed in the free online study builder, Lab.js (Henninger et al., 2021).

Procedure:

Parents were directed to load the tasks on their computer via a website and email output file back to the experimenter. A videoconferencing session was active to guide parents through the experimentation setup (e.g., loading the task file, collecting the data, sending the output file back, etc.) and to make notes of any unforeseen circumstances (e.g., internet issues, sibling intervening, etc.). Children were seated at a table facing the computer screen. A parent sat next to the participant during testing, and their computer screen was shared with the researcher. Auditory stimuli were presented in free field through the computer's speakers. Answers were recorded via mouse clicks. Depending on each child's computer fluency as assessed during the training session, mouse clicks were either initiated directly by the participant or by the parent when the participant pointed to the screen. The remote testing did not allow us to control for the absolute audio intensity. Instead, before testing, participants set the computer's audio output to a comfortable level, based on

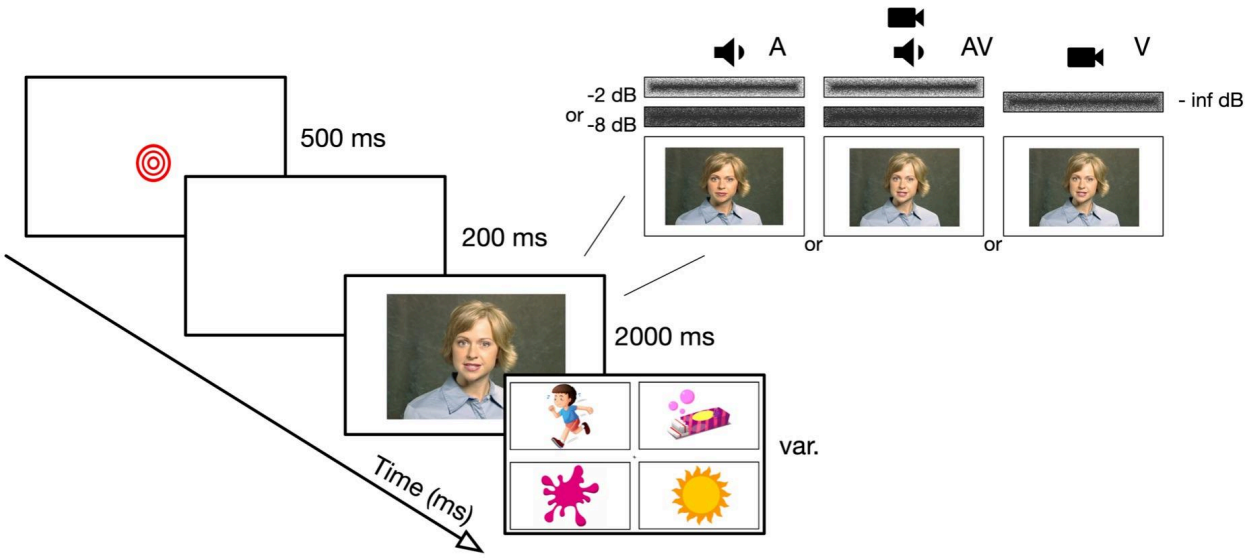
a repeating speech fragment of a cartoon character saying, *"Let's go play some games"* (recorded in quiet with a mean intensity of 60 dB SPL), and kept at the same level throughout the tasks.

All participants completed a training session prior to the AV tasks. First, we exposed the participants to all 38 clipart pictures—25 target pictures and 13 foil pictures that were never used as a target. Every picture was shown for 1.6 s, accompanied by a related word (i.e., picture name) spoken by an adult female native English speaker (different from the speaker of the stimuli for the actual tasks). After all pictures had been shown, they had to name five randomly selected pictures. If the participant made any mistake, the training started over until all five pictures were named correctly. Eight children had to repeat the familiarization picture phase once. Although we only tested five random words per participant, a pilot study confirmed that children this age could correctly identify all pictures and that these pictures were appropriate for the target words. Next, we familiarized the participants with the AV speech task by exposing them to 8 trials with feedback. These stimuli were reserved for training only and were not used in the actual task. Task 1 contained 110 trials (50 AV, 50 audio-only, and 10 visual-only) and Task 2 contained 90 trials (40 AV, 40 audio-only, and 10 visual-only), each having five blocks. All target stimuli words were presented in each task an equal number of times in both AV and audio-only modalities.

Five conditions with three modalities (2 AV, 2 audio-only, and 1 visual-only, see Fig 3.2) were tested in each task. The presentation order of the modality and noise level was randomly assigned for each participant. Every trial started with a fixation target (500 ms), followed by a blank screen (200 ms). Then, the 2-second-long noise was presented with the word stimulus appearing 500 ms after the beginning of the noise. Finally, a screen with four pictures (arranged in a 2x2 block) appeared, and the participant selected a response by clicking an image. No feedback was provided and

there was no time limit to respond. The position of the pictures was randomized on each trial.

FIGURE 3.2. Visualization of set up for Task 1 and 2.



The task started with a 500 ms fixation marker 500 ms, followed by a 200 ms blank screen. Then the auditory, AV or visual stimulus was presented (in different SNRs, 2000 ms), followed by the four answer choices. There was no time limit in the response period.

In order to monitor participants' visual attention, a picture of a cartoon character, that also narrated the instructions, was randomly shown on the screen. This character served as a catch-trial stimulus to measure cross-modal attention and was a fun motivator for the participants throughout the tasks. During the two AV tasks, this cartoon appeared 20 times in total, spread throughout all conditions. The participants were instructed to identify this character by saying its name. One child opted to raise their hand for this catch-trial response instead of providing a verbal response.

Task 3: Tone-in-Noise counting task

Stimuli and materials:

Task 3 was an audio-only psychophysical task. Three harmonic complexes were generated, with each complex consisting of the first five harmonics with the

following fundamental frequencies: 200 Hz, 400 Hz, and 600 Hz. Each complex was 300 ms long. One to four auditory events (or "beeps") were generated (with an interstimulus interval, or ISI, of 100 or 200 ms, presented isochronously) by repeating one of the three complexes. Variability in harmonic complexes and ISI was added to keep the child engaged and to increase the complexity of the task. This resulted in 24 different stimuli combinations (3 harmonic complexes x 1-4 "beeps" x 2 ISIs). These stimuli were presented in the presence of a LTASS noise at 3 different SNR levels (no noise, −2 dB, and −8 dB). Similar to Task 1 and 2, the free online study builder, Lab.js (Henninger et al., 2021), was used to build this task.

Procedure:

The set-up of this task was very similar to the previous two tasks. Participants were first trained in a 1I-4AFC task with feedback. The participants were presented with the same 500 ms fixation marker, followed by a 200 ms blank screen. The participants heard 1, 2, 3, or 4 "beeps" while seeing a fixed cartoon "listening" character. This was followed by four choices (arranged in a 2x2 block) with pictures showing 1, 2, 3, or 4 dots (with the accompanying number). These choices stayed in the same order throughout Task 3, in contrast to the AV tasks. Answers were recorded via mouse clicks.

The training consisted of 8 practice items with feedback, followed by 60 test trials without feedback, presented in 3 blocks of 20. At the end of each block there was a verbal confirmation on how far along the participant was in the task, similar to the previous tasks. In this task, no speech sounds or visual stimuli were provided.

3.1.3. Statistical analysis and results

General and cross-modal task attention

First, we wanted to verify whether the children were attending to our visual stimuli—an experimental variable that could be challenging to control especially

when the tasks were carried out at home and online. All children showed that they were attending to the visual stimuli as they noticed almost all of the 20 catch trials, with a mean count of 19.14 (SD = 1.06, min = 16, max = 20). The random answer options in the 4AFC task also served as a mechanism to account for general attention. We expected that the random foil option would only be chosen when the participant was not attending to the stimulus or was guessing because there was not enough information to make a measured decision. In the audio-only and AV modalities, we expected the stimuli were informative enough that a response to the random foil should be interpreted as moments of inattention. In both the audio-only and AV modalities only 1% of the total responses were random foils. Thus, these data revealed that the participants were attending to the stimuli. Neither the rate of random responses (R_{35} = −.29, p = .087) nor the rate of catch-trial responses (R_{35} = −.19, p = .245) correlated with performance in the AV tasks.

Audiovisual enhancement

Statistical analyses were performed using the lme4, lmerTest, stats and psych packages in R (Bates et al., 2015) and RStudio (version 1.3.1093). The ANOVA function provided F-statistics for the models generated. We used linear (mixed effects) models to test two main hypotheses; (1) whether children this age show AV gain and (2) whether individual differences between participants could be explained by vocabulary knowledge, phonological awareness skills or psychophysical testing performance.

First, we tested the hypothesis that children in first grade use visual cues in conjunction with the auditory stimulus to improve CVC word perception when presented in noise. The dependent variable was the combined percentage of correct trials in Tasks 1 and 2. In *Model 3.1: Test score ~ Modality*SNR+Set+(1|Participant)+ (Set|Participant)*, we examined the fixed effects of the presentation modality (i.e., AV vs. audio-only), the SNR level (i.e., −2 dB SNR and −8 dB SNR), and the impact of the

number of stimulus set size (i.e., small vs. large). The audio-only modality, −8 dB SNR, and the large target set size served as references in each category. An interaction effect of modality and SNR was modeled to account for the possibility that children might only attend to the visual cues in the −8 dB SNR condition. The model included a random intercept for participant, which estimates a variance component for the fixed factors such that the model fits an intercept for each participant. A random slope for participant depending on the stimulus set, was included to account for individual slope differences between Tasks 1 and 2. Gender was initially included as a variable in the model, but no significant effect was observed and thus, this variable was deleted from the model. Table 3.1 summarizes how children benefited from visual salience of the phonemes (Model 3.1) and their corresponding task performance is shown in Fig 3.3 The random intercept had a standard deviation of 0.01, suggesting that performance across participants was quite consistent.

TABLE 3.1. Summary of linear mixed model 3.1.

Predictor	**Estimate**	**SE**	t-score	p-value
Intercept	0.797635	0.009551	83.517	**< 2e-16 *** **
Modality ~	0.03473	0.010644	3.263	**0.00128 ** **
SNR°	0.107973	0.010644	10.144	**< 2e-16 *** **
Set^	0.020135	0.01028	1.959	0.05794
Modality ~: SNR°	-0.029459	0.015053	-1.957	0.05162

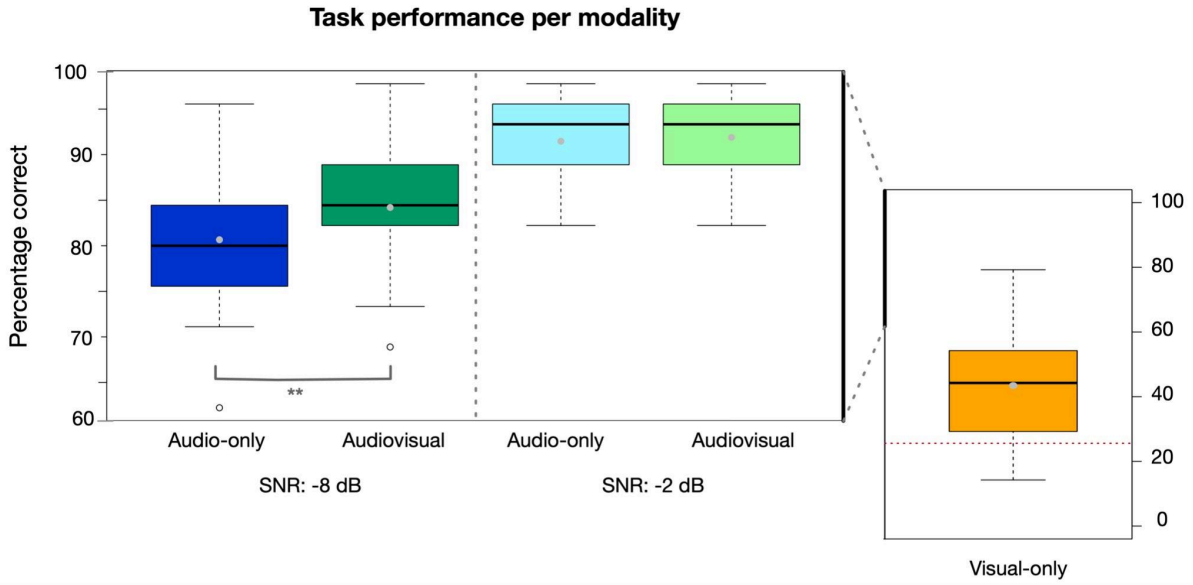
Model 3.1: Test score (%) ~ Modality * SNR + stimuli set + (1 | participant) + (stimuli set | participant). Regression estimates, standard errors, t-scores and p-values are presented in the table. The model predicts the test performance. Bold p-values represent statistically significant findings. Significance: * p < .05, ** p < .01,*** p < .001. ~ Reference is audio-only. ○ Reference is -8 dB. ^ Reference is large stimulus set.

An analysis of variance (ANOVA) showed a statistically significant main effect of modality ($F_{1, 219}$ = 7.0609, p = .008), suggesting that children performed better in the AV modality than in the audio-only modality regardless of the SNR level and the set-size of their responses. The significant main effect of SNR ($F_{1, 219}$ = 153.4744, p <

.001) suggests that participants performed better overall in the −2 dB SNR condition than in the −8 dB SNR condition. A trend for a different relationship between audio-only and AV in the −2 dB and −8 dB SNR was observed, but the interaction between modality and SNR was not statistically significant ($F_{1, 219} = 3.8299$, $p = .052$). There was no main effect for target stimuli set size.

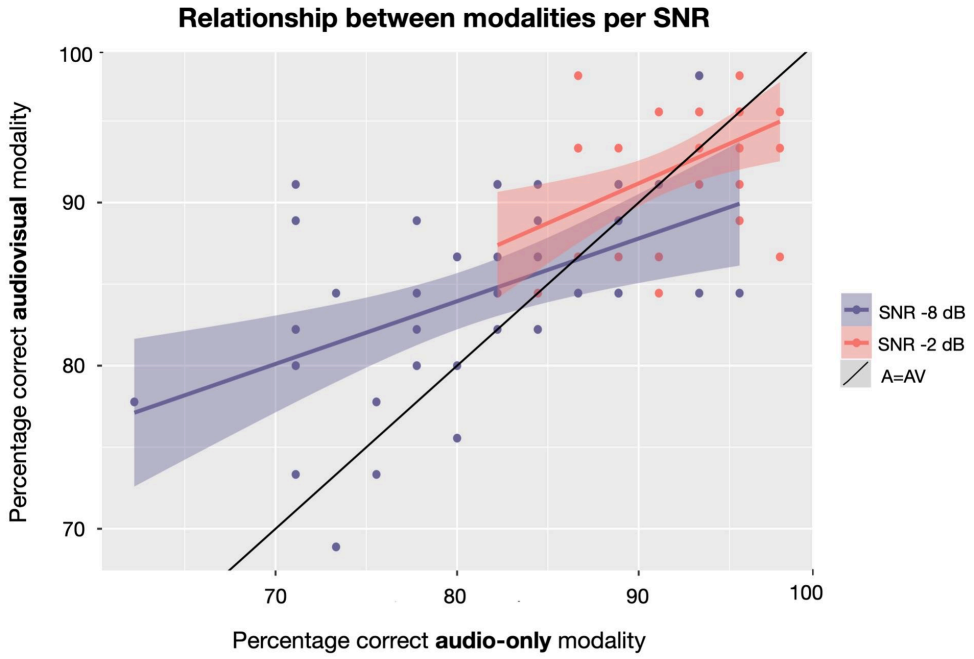
Post-hoc analysis showed no significant differences (in mean or median) between audio-only and AV performance in the −2 dB SNR condition. It is also of note that the data in the −2 dB SNR condition were skewed, revealing a potential ceiling effect (Fig 3.3). We further explored this relationship between audio-only and AV performance at the two SNR levels (Fig 3.4). All data points above the equal-performance line (black diagonal) showed better performance in the AV modality for each participant. Regression lines for each SNR condition were plotted, within the range of the data. Two important observations were revealed in this figure. First, the ceiling effect in the −2 dB SNR condition (red in Fig 3.4) was evident with all these data points clustered in the top right corner. Second, we found that the higher the score for the audio-only performance, the less children benefited in the AV modality. Exploratory analysis showed, in both −2 dB and −8 dB SNR conditions, a strong negative Pearson correlation $R_{35} = -.67$, $p < .001$ (−8 dB SNR condition) and $R_{35} = -.48$, $p = .002$ (−2 dB SNR condition) between the audio-only performance and the AV benefit. The higher performance in the audio-only modality, the lower the gains (Fig 3.4). For the highest (> 90%) audio-only scores, we found that more than 60% of the children showed no AV benefit.

FIGURE 3.3. Boxplots of percent correct test scores per modality (audio-only, AV, visual-only) and SNR (-8 dB or -2 dB).



Thick horizontal lines represent medians, gray dots represent means, boxes represent interquartile ranges, and whiskers represent range, excluding outliers. Outliers are defined as values falling more than 1.5 x below or above the 25th and 75th percentiles, respectively, and are shown as circles. Significance based on model 1b: * p < .05, ** p < .01, *** p < .001. The red dotted line in the visual-only modality represents chance level.

FIGURE 3.4. The relationship between percent correct scores on the audio-only and AV modalities per participant at each of the SNR levels.



Regression lines are plotted with 95% confidence intervals. The black line shows where audio-only and AV performance are equal. Thus, all dots above the black line represent AV benefit.

Speechreading performance

We analyzed performance in the visual-only modality to examine whether participants could use visual cues explicitly to do the tasks. We performed a one-sided t-test against the children performing at chance level (i.e., > .25 in these 4AFC tasks). We found a mean score of 44% correct (SD = 15%), with a range between 15% and 80% (Fig 3.3, right). Importantly, children scored significantly above chance (t_{36} = 7.6754, p < .001). Only 5 of the 37 participants scored at chance level or lower.

Child factors

Next, we tested the hypothesis that individual differences in the amount of AV benefit could be explained by children's vocabulary skills, phonological awareness, or performance on a control psychophysical task. By using *Model 3.2: Relative AV Gain ~ Set * Vocabulary scores + Task3 (control psychophysical task) + Phonological Awareness scores + Gender*, we examined the relationship between the AV gain (only in the −8 dB SNR condition) and these child factors. We also modeled the interaction between vocabulary score and set size because we reasoned that vocabulary skills may relate more strongly to performance with the larger stimulus set. Because there was a strong negative relationship between the AV benefit and the auditory scores, we did not use the simple difference score between AV and audio-only modalities. This bias that high audio-only scores necessarily lead to low benefit scores could be avoided by using a relative AV gain. Here we divided the AV gain (described as the amount of speech recognition improvement relative to a baseline modality, audio-only) by the difference between the total response information (100% correct) and the audio-only performance: *relative AV gain = (AV−A)/(1−A)*. This would answer the question: What is the added visual contribution relative to the possible available contribution in the absence of visual cues? (Alsius et al., 2016; Grant & Seitz, 1998; Sumby & Pollack, 1954). We included gender as a factor to account for potential AV

enhancement differences between boys and girls. The large stimulus set and the female gender group served as reference groups in each category.

After checking the assumptions for normality, linearity, and homoscedasticity for Model 3.2, we excluded four scores based on extreme residual values (i.e., 4 SD from the mean, with the values of the relative AV gain varied between 1 and −4, and all excluded values smaller than −2). Results were similar when outliers were included. The model summary (Model 3.2) exploring the relationship between AV processing, vocabulary, phonological awareness, general psychophysical test performance, and gender is presented in Table 3.2. Variability of both dependent and independent variables are described in Table 3.3.

TABLE 3.2. Output of linear model 3.2 with regression estimates, standard errors, t-scores and p-values.

Predictor	**Estimate**	**SE**	t-value	p-value
Intercept	-1.295493	1.05849	-1.224	0.2255
Set ^	0.14172	0.974434	0.145	0.8848
Vocabulary	0.012885	0.007588	1.698	*0.0944*
Gender +	0.319262	0.134055	2.382	**0.0203 ***
PA	-0.033742	0.01754	-1.924	*0.0589*
Psychophysical task	0.763777	0.765357	0.998	0.3221
Set ^ * Vocabulary	-0.001892	0.00994	-0.19	0.8496

Model 3.2: AV Gain ~ stimuli set * voc + PA + psych + gender

Output of linear model 3.2 with regression estimates, standard errors, t-scores and p-values. Bold p-values are significant, AV Gain: adjusted (AV-A /1-A), for SNR -8 dB, Vocabulary = vocabulary score, psychophysical = psychophysical testing score, PA = phonological awareness score.

^ Reference is large stimulus set, + Reference is female

The analysis of variance showed a statistically significant main effect of gender ($F_{1, 1}$ = 8.7832, $p < .01$), suggesting that boys have significantly higher normalized AV gain than girls. A trend for phonological awareness was observed ($F_{1, 1}$ = 2.8244, $p = .09$). Higher phonological awareness scores were associated with lower AV gain. No other effects reached statistical significance.

TABLE 3.3. Descriptive statistics, showing mean, median, standard deviation, min and max scores of dependent and independent variables.

Variable	Mean	Median	SD	Min	Max
Audio-only (A)	0.80	0.80	0.10	0.56	0.96
Audiovisual (AV)	0.84	0.85	0.09	0.60	1.00
AV Gain (AV - A)	0.04	0.05	0.10	-0.20	0.35
Relative AV Gain (AV - A / 1 - A)	0.10	0.25	0.56	-1.50	1.00
Vocabulary scores	97.24	99.00	12.68	67.00	124.00
Phonological Awareness	19.23	20.00	4.10	11.00	26.00
Control Psychophysical task	0.89	0.92	0.09	0.58	1.00

Descriptive statistics, showing mean, median, standard deviation, min and max scores of dependent and independent variables. Audio-only, Audiovisual, AV gain and relative AV gain statistics are calculated of the -8 dB SNR stimuli. Vocabulary and Phonological awareness scores are expressed in Raw scores. Audio-only, AV, AV Gain and the control psychophysical task are expressed in percentages.

Post-hoc analysis showed that the difference in relative AV gain between boys and girls resulted from lower performance for the auditory-only modality in boys. Performance in the AV modality was equal between boys and girls. Although we did not find a relationship between normalized AV gain and our control psychophysical task, a small, non-significant, positive Pearson correlation (R_{35} = .266, p = .112) was found between overall performance on the AV tasks and the control psychophysical task. Further exploratory analysis showed no significant relationships between the individual modalities (AV, audio-only, or visual-only) and the phonological awareness or vocabulary scores.

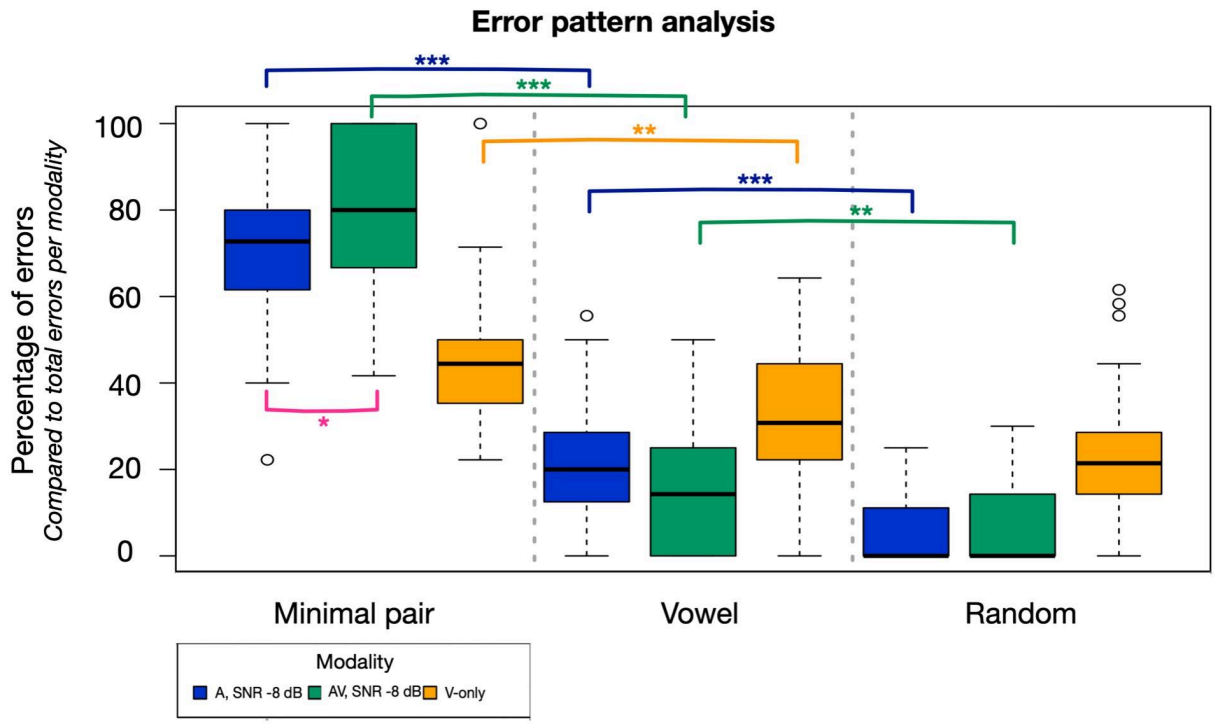
Other Planned Parametric Tests

To explore the impact of both answer options in our 4AFC task and the impact of the added visual cues in the tasks, we performed an error pattern analysis (the total errors per error category) per modality (AV, audio-only, visual-only) in the −8 dB SNR condition. We classified our errors into three categories: minimal pair (e.g.,

confusing “run” and “sun”), same-vowel (e.g., confusing “run” and “gum”) and random (e.g., confusing “run” and “pink”). We expected a descending error pattern of choosing the minimal pair followed by same-vowel and random foil. Furthermore, we expected that relative to the audio-only modality, participants would make more minimal pair errors in the AV modality, confirming the use of visual cues in the tasks. This error analysis was conducted with pairwise comparisons using Wilcoxon rank sum test and false discovery rate corrected using Benjamini–Hochberg method (Benjamini & Hochberg, 1995).

Fig 3.5 shows that error patterns were as expected in the audio-only and AV modalities: significantly more minimal pair errors than the same-vowel foil errors ($p < .01$) as well as more same-vowel errors compared to random responses ($p < .01$). The error pattern in the visual modality was less distinct. There were significantly more minimal pair errors than same-vowel foil errors ($p < .01$) but did not reach significance when comparing the same-vowel foils with random responses ($p = .052$). The less polarized error pattern in the visual-only modality was expected, given that speechreading without auditory input is much more difficult than audio-only or AV judgements, especially for children this age. Although there were fewer errors in the AV modality than in the audio-only modality, a significantly greater portion of errors were minimal pairs in the AV modality the audio-only modality ($p = .047$). This suggests that even when children made errors, the use of visual cues led to answer choices closer to the presented stimuli. Detailed p-values can be found in the original manuscript in Appendix, table B (https://doi.org/10.1044/2021_JSLHR-21-00196).

FIGURE 3.5. Error pattern analysis.

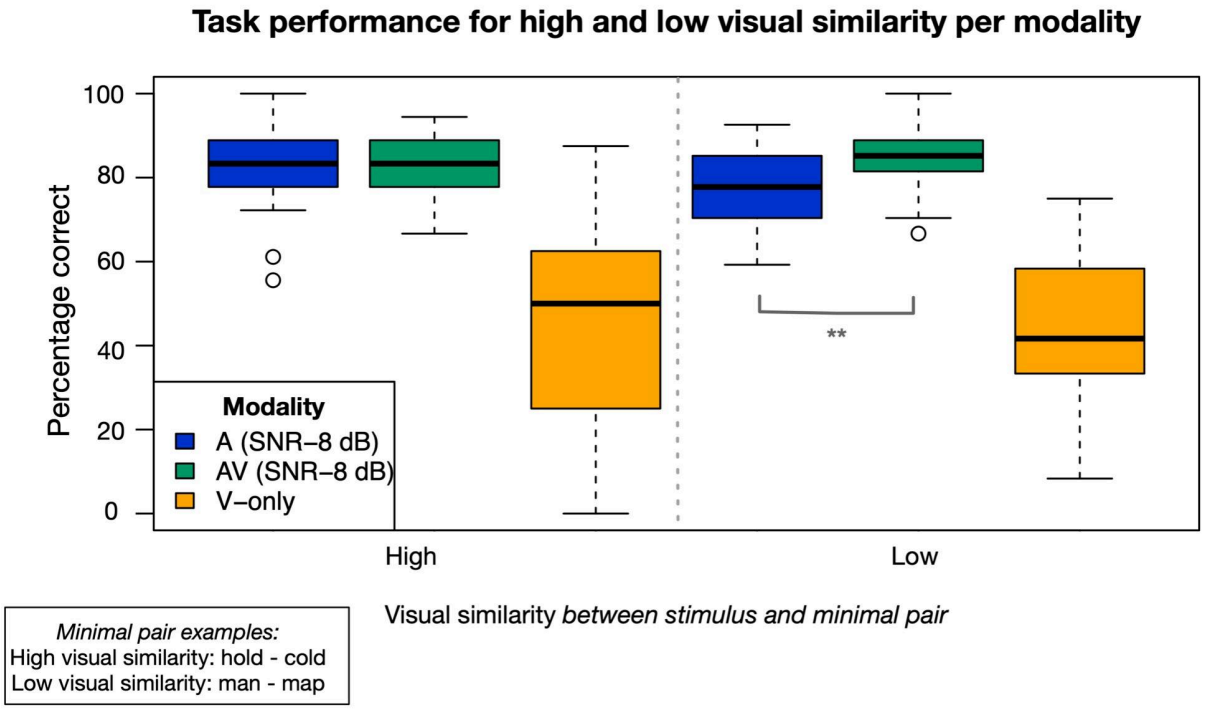


Percentage of errors by error type (minimal pair, vowel, random) and modality (audio-only, AV, visual-only). The percentage on the Y-axis is the percentage of errors in relation to total errors per modality. Significance: * p < .05, ** p < .01, *** p < .001.

All our minimal pairs were auditory minimal pairs. We therefore classified the minimal pairs by visual similarity (Appendix, Table A (https://doi.org/10.1044/2021_JSLHR-21-00196); e.g., high visual similarity: "hold" - "cold"; low visual similarity: "man"-"map") to further analyze the utility of the visual information. Visual similarity analyses were conducted with pairwise comparisons using Wilcoxon rank sum test and false discovery rate corrected using Benjamini–Hochberg method (Benjamini & Hochberg, 1995).

The difference between audio-only and AV performance, and therefore the AV benefit, was only significant (p = .002) for stimuli with low visual similarity. This suggests that the children in this age range only show AV benefit for stimuli that are visually more distinctive. We found no statistical differences between low and high visual similarity in the visual-only modality (p = .811).

FIGURE 3.6. Task performance for high and low visual similarity minimal pairs per modality.



The percentage on the Y-axis is the percentage correct in relation to different visual similarity, presented per modality (audio-only, visual-only or AV), significance: * $p < .05$, ** $p < .01$, *** $p < .001$. Examples of a minimal pair with high visual similarity: "hold" – "cold" and low visual similarity: "man" –"map".

3.1.4. Discussion

The aim of this study was to examine the factors that impact AV processing in first graders and to look into individual differences found in AV enhancement (Barutchu et al., 2010; Fort et al., 2012; Jerger et al., 2009; Lalonde & Holt, 2016; Ross et al., 2011). Specifically, we designed an experiment that constrained cognitive and linguistic demands (McCreery et al., 2010). By using a closed-set 4AFC task with simple CVC words, we ensured all children knew the presented words and did not have to use potentially undeveloped articulatory skills in their responses. We further explored the relationship between AV benefit, vocabulary knowledge, phonological awareness, psychophysical testing skills and attention (general and cross-modal) in these first graders. Previous findings were extended by exploring the use of salient visual cues in children in these AV and visual-only tasks.

An overall AV gain was found in both Tasks 1 and 2. Post-hoc analysis showed a ceiling effect of the −2 dB SNR condition and therefore all further analysis only used the −8 dB SNR condition. Error analysis indicated that when errors were made, minimal pair errors dominated, followed by vowel errors, then random errors. This pattern was more clearly expressed in the AV modality than in the audio-only modality. The AV benefit was mainly found for minimal pairs with clear visual distinction (i.e., low visual similarity). Individual AV enhancement differences could not be explained by vocabulary skills or phonological awareness skills. A non-significant trend towards higher AV gain with lower phonological awareness skills showed, but this was mediated by significant gender differences in the audio-only modality.

AV benefit in first graders and use of visual cues

Results showed an overall AV benefit for spoken CVC words in noise in first graders. The average AV enhancement was 4% for the total group and 8% (N = 24) for the subset of children that actually showed benefit. This seemingly small increase in performance was not surprising with a high average audio-only performance (81%), yet our error analysis showed this was of significant impact to influence decision making in the current 4AFC task and therefore could have been experienced as a significant improvement. Although we did not measure processing speed, small benefits in accuracy could accompany larger improvements of speech processing speed and effort (Holt, Bruggeman & Demuth, 2020) and might have a significant impact in the classroom (Mealings et al., 2015). We still found some variability in individual results, which indicates that our constraints on the cognitive and linguistic demands alone were not enough to completely exclude variability in AV performance in children. This could indicate that some of the variability found in AV enhancement is not related to these cognitive and linguistic skills.

Although the presence of an overall AV benefit in young children has been described before (Barutchu et al., 2010; Fort et al., 2012; Lalonde & Holt, 2015; Lalonde & McCreery, 2020), we put particular focus on the use of salient visual cues in this population. Therefore, we also looked into speechreading performance, without any auditory input. The studies that addressed speechreading performance in children have had mixed findings. The differences described in these studies might be explained by experimental design differences. Although children were able to choose the correct answer from a variety of answer options in a closed-set task (mean = 44% correct), they might not have developed speech reading skills required by the use of an open-set task. Holt at al. (2011) found that typically developing children (3- to 6-year-olds) were not able to speech read via an open-set visual-only task, where others (Davies et al., 2009; Heikkilä et al., 2017; Kyle et al., 2013) reported above chance performance of speechreading in closed-sets in typically developing children of a similar age group. Kyle et al. (2013) found similar results to the current study, with an average performance on a 4AFC word task around 45–50% correct, significantly above chance. This suggests that children in first grade begin to form knowledge of the relationship between sounds and the visual component.

Previous work by Buss et al. (2016) showed that attributes of the foils presented at this age influence the outcome of speech-in-noise tasks. Their 4AFC task was harder when foils were phonetically similar to the target than when they were phonetically distinct. Our results are consistent with this finding. Responses followed the hypothesized order with mainly correct answers, followed by minimal pair alternatives, then vowel alternatives and lastly, a small amount of random answer choices in the three different presented modalities (AV, audio-only, and visual-only). The differences in error types were smaller in the visual-only modality. This is not surprising as children overall performed significantly worse in this modality. This error pattern suggests that for the majority of trials, children approached this task

more like a quasi two-alternative rather than a true four-alternative forced-choice task.

It is interesting to note that out of all errors made per modality, there is a significantly higher percentage of minimal pair errors in the AV modality. This suggests that children use visual information in the AV modality to aid their response. We marked each minimal pair by visual similarity (Owens & Blazek, 1985) and found a significant AV benefit in the low, but not high, visual similarity group. This confirms that added distinct visual salience of the phonemes can actively be used as a decision-making tool for children this age, therefore increasing intelligibility.

Relationship between AV gain and other factors

Exploratory analysis showed, in both −2 dB and −8 dB SNR conditions, a strong negative relationship between audio-only performance and AV benefit. As sensory processing will be mainly determined by auditory stimuli in situations where the noise level is not highly deleterious (Barutchu et al., 2010; Ma et al., 2009; Ross et al., 2006), one could expect that children with high audio-only scores barely use the visual component in AV modalities and therefore show very limited AV gain.

Earlier research (Elliott, 1979; Fort et al., 2012; Jerger et al., 2009) suggested that linguistic knowledge moderates AV performance. In this study, we set out to explore whether AV performance could be decoupled from other linguistic factors. Specifically, we chose simple CVC words that were mastered by children this age, and two stimuli sets with a different number of stimuli, to examine whether vocabulary knowledge and retrieval are necessarily tied to AV performance at this age. We found that vocabulary skills do not necessarily serve as a good predictor of AV gain if the task does not demand high vocabulary skills. This finding is consistent with results from Lalonde and McCreery (2020), who found no relationship between AV benefit and vocabulary in somewhat older children (6- to 13-year-olds), for

sentences with similarly early acquired target words. It is important to note that children with higher vocabulary skills may still perform better integrating visual information in daily settings when performing the more complex task of comprehending spoken language. We also found that there is no relationship between the vocabulary set size used in a 4AFC task and AV benefit for a task low in cognitive and linguistic demands.

Another suggested explanation for differences in AV enhancement was phonological awareness performance. Dodd et al. (2008) showed that children with phonological impairments rely more on the auditory component in AV illusions, suggesting speechreading skills are better for children with good phonological awareness. Heikkilä et al. (2017) also found an association between phonological knowledge and speechreading skills in children. Exploratory analysis showed no significant relationship between phonological awareness skills, and speechreading (visual-only) or AV performance. Model 3.2, however, showed a negative trend, ($p = .09$) between phonological awareness and AV gain, suggesting that the better the phonological awareness scores, the smaller the AV gain. Given that phonological awareness is not a significant predictor in our model with our current data, further studies with a greater sample size may help in further exploring the relationship between AV enhancement and phonological awareness.

We also wanted to explore whether AV gain can be explained by general task performance or other attentional factors. We introduced an extra psychophysical task with a similar set-up but no visual component to confirm that any individual differences we observed were due to more than task performance. Although there was, as expected, a small but non-significant, positive Pearson correlation coefficient ($R = .27$) between performance on the two tasks, no relationship was found between the normalized AV gain and this extra task. Therefore, the AV enhancement could not be explained by psychophysical task performance. Finally, neither the rate of random responses ($p = .087$) and catch-trial responses ($p = .245$) correlated with their

performance in these AV tasks. General (random response) or cross-modal (catch trials) attention were not predicting factors in this study.

Gender differences

Although gender was not a significant factor in Model 3.1, it was a significant predictor for the relative AV gain ((AV−A)/(1−A), Model 3.2). Gender differences in AV processing are understudied, with a limited set of findings reported. Lalonde and McCreery (2020) found no gender differences in children (6- to 13-year-olds, listeners with normal-hearing thresholds and hearing impaired). Ross et al. (2015) found that typically developing female children (8- to 17-year-olds) outperformed male children in both auditory and AV speech perception, but they did not find these results for adults. They proposed that the development of AV integration is delayed in male children. Interestingly, our data in this study showed that boys had a lower performance on the audio-only task, but a similar performance on the AV task. This is in line with the findings of better auditory speech perception in female adults (McFadden, 1998; Yoho et al., 2018) and babies (Newmark et al., 1997). It is possible that the lower performance in the auditory-only modality for male children in our study would be caused by distraction of the unexpected "still face" in contrast to the AV and visual-only modality, as performance was equal across gender for those two modalities. Both boys and girls mentioned, and therefore noticed, that the face was not moving in the audio-only modality, but it might impact boys more because they showed lower inhibitory control to "oddballs" (Yuan et al., 2008). This interpretation is speculative, and future studies should further tease apart the origin of the gender differences.

Study limitations and future directions

The ability to test in the laboratory was hampered due to pandemic limitations at the time of this study. While our online test set-up (complemented with videoconferencing) allowed us to account for some of the experimental variables

(e.g., SNR, general and cross-modal attention, behavior, computer used, table set-up, role of adult), we could not control for the absolute stimulus presentation levels. Instead, participants were instructed to adjust their computer's audio output to a comfortable level, based on a speech stimulus presented before the experimental tasks. In daily listening environments, speech loudness varies considerably; factors such as the relative distance between the speaker and the listener, as well as how the speaker is oriented relative to the listener, contribute to these natural variations (Monson et al., 2019). Thus, one could argue that our results are more generalizable than if data were collected in a more controlled setting. Nevertheless, future studies should investigate whether AV gain is affected by the absolute intensity level of auditory stimulus.

In this study, we focused on a limited age range, to target that group where AV performance is least showing (Jerger et al., 2009). We hoped to reduce individual variability by using a strict age range (i.e., first graders) and by constraining cognitive (closed-set, 4AFC task, with highly imageable targets) and linguistic demands (vocabulary acquired by 3 to 5 years of age) of the task. This would allow us to predict individual differences related to vocabulary and phonological awareness that would not be caused by the specific demands of the task. Although individual variability was still present in our group, it would be interesting to look at the impact of this more constraining set-up over a wider age range. If we test younger children (3- to 5-year-olds) or clinical populations (e.g., children with autism spectrum disorder or developmental language disorder) who may have more recently acquired the words used in our task, we might find different results. Future studies should also extend to lower SNR ranges to explore how that influences overall performance and the AV gain.

3.1.5. Conclusions

The current study showed that children in first grade can use visual speech to improve accuracy of speech-in-noise perception at low SNRs. Their performance gain was dependent on the salience of the visual cues in speech-in-noise tasks. This was evident from their speechreading performance as well as the different error patterns they made, especially between words with low and high visual similarity. Children in first grade could perform above chance on a 4AFC closed-set speechreading task with early-acquired CVC words. A constrained AV speech task (closed-set, stimuli of low complexity) as used in this experiment showed no relationship of individual differences of AV speech enhancement with first-graders' vocabulary knowledge. This could suggest that the relationship between vocabulary knowledge and AV performance may be mediated by task demands. More research is needed to understand what underlies gender differences in AV speech enhancement.

3.2. Audiovisual speech perception benefits are stable from preschool through adolescence

Section 3.2 is under review at the journal Multisensory Research and is written with co-authors Jason D. Yeatman, Kaylah Lalonde, Piper Doering, and Adrian KC Lee.

Abstract

The ability to leverage visual cues in speech perception—especially in noisy backgrounds—is well established from infancy to adulthood. Yet, the developmental trajectory of audiovisual benefits stays a topic of debate. The inconsistency in findings can be attributed to relatively small sample sizes or tasks that are not appropriate for given age groups. We designed an audiovisual speech perception task that was cognitively and linguistically age-appropriate from preschool to adolescence and recruited a large sample (N=161) of children (age 4-15). We found that even the youngest children show reliable speech perception benefits when provided with visual cues and that these benefits are consistent throughout development when auditory and visual signals match. Individual variability is explained by how the child experiences their speech-in-noise performance rather than the quality of the signal itself. This underscores the importance of visual speech for young children who are regularly in noisy environments like classrooms and playgrounds.

3.2.1. Introduction

Speech perception is difficult in noisy environments like classrooms and playgrounds. Fortunately, the visual system significantly adds to the ease and quality of speech understanding in these noisy situations through mechanisms of audiovisual (AV) speech perception (Grant & Seitz, 2000; Ross et al., 2006; Sumby & Pollack, 1954). These benefits occur for two reasons. First, visual cues help to identify the source of the speech signal by providing temporal and spatial cues that we can

use to direct attention to the target speaker (Bernstein et al., 2004; Fiscella et al., 2022; Peelle & Sommers, 2015). Second, we can use more speech-specific AV cues, like leveraging the correspondence between the visible movement of articulators (e.g., the mouth and lips) and acoustic-phonetic cues in the speech signal (e.g., place of articulation /b/ vs. /g/) to improve the accuracy (Eskelund et al., 2011; Tye-Murray et al., 2007a) and the speed (Reisberg et al., 1987) of our speech understanding. Our linguistic competence, language background, word and world knowledge (Grant et al., 1998), as well as cognitive skills like attention or working memory (Alsius et al., 2005; Frtusova et al., 2013) play a large role in AV benefits of conversational speech in complex listening environments. Language and cognition can help improve our AV speech perception performance (Golestani et al., 2009; Taitelbaum-Swead & Fostick, 2016), and in turn, AV speech perception improves language and cognitive skills.

This multisensory interaction in all its complexity becomes even more interesting when looking at it from a developmental perspective. Both bottom-up perceptual skills and top-down language and cognitive skills are less developed in children than adults, and complex auditory and visual speech perception performances only mature in adolescence (Leibold & Buss, 2019; McCreery & Stemachowicz., 2011; McMurray et al., 2018; Wightman et al., 2006). Developmental changes in AV speech perception are therefore not surprising (Fort et al., 2012; Jerger et al., 2009; Lalonde & Holt, 2016; Maidment et al., 2015; Ross et al., 2011), as many factors of the individual (e.g., development of the auditory and visual systems, linguistic skills, phonological skills, attention, etc.) or the AV speech perception task (e.g., required verbal or motor responses, adapting to different signal-to-noise ratios and linguistic complexity of the speech stimuli, etc.) change across development (Holt et al., 2013; McCreery et al., 2017). We know that children are already sensitive to AV speech shortly after birth (Aldridge et al., 1999), but AV speech perception benefits that improve speech recognition (beyond detection and discrimination), especially in noise, might start

later. Specifically, infants and young children mainly use general mechanisms like temporal and spatial cues to create AV speech perception benefits, whereas the use of advanced speech-specific AV cues to improve the quality of speech understanding is still more limited (Lalonde & Holt, 2016; Lalonde & McCreery 2020; Lalonde & Werner, 2019). When in the developmental trajectory both general and speech-specific cues are combined to improve speech understanding is still unclear as large variability in task performance between children and large variability across different tasks might play a significant role in previous findings. Audiovisual speech recognition benefits have been measured as early as three years for syllables (Desjardins et al., 1997; Lalonde & Holt., 2015), words (Holt et al., 2011), and sentences (Holt et al., 2011), but some report a later start, at the age of four (Erdener & Burnham, 2018), six (Fort et al., 2012; Maidment et al., 2015), or even nine years old (Wightman et al., 2006).

Interestingly, general (i.e., temporal and spatial) and speech-specific AV correspondences affect multisensory speech processing in complex listening situations at different stages (Fiscella et al., 2022). General cues are especially important in early processing to help an individual work out which pair of sensory inputs to integrate in order to separate the overlapping speech signals of individual talkers (Aller & Noppeney, 2019; Fiscella et al., 2022; Jones & Noppeney, 2021). This is often referred to as solving the causal inference problem and it becomes more challenging when (1) there are more competing sources present (Cao et al., 2019), like in classrooms or on playgrounds or when (2) the auditory and visual signals are not equally reliable. For example, visual information gets lost when the teacher wears a face mask in the classroom, or auditory information is affected when the signal-to-noise ratio (SNR) of the auditory signal is degraded due to children yelling on the playground. Therefore, each incoming auditory and visual signal must be weighted based on the uncertainty of whether they come from the same source and

the uncertainty of the signal itself in each modality over time (Aller & Noppeney, 2019; Jones & Noppeney, 2021).

When a modality is less reliable either due to degradation in input signals (e.g., background noise) or sensory perception (e.g., hearing loss), less weight is assigned to make a final decision about the AV percept (Ma et al., 2009; Schwartz, 2010). However, there is a lot of individual variability in how these modalities are weighted and how large a benefit can be gleaned by leveraging visual information to support speech perception. There are indications that when younger children apply weights to individual perceptual modalities of AV speech, performance is weighted more heavily by the auditory component. Adults give relatively more importance to the added visual cues in AV speech, especially for conflicting information (Hockley & Polka, 1994; Massaro et al., 1986). This, in combination with children's need for higher SNRs for similar performance on auditory speech-in-noise tasks (Corbin et al., 2016; McCreery & Stelmachowicz, 2011) and lower speechreading skills in children (Wightman, 2006; Knowland et al., 2016; Kyle et al., 2013; Tye-Murray et al., 2014), makes it not surprising that children retrieve less visual phonological knowledge needed for speech-specific mechanisms used to benefit from AV speech (Desjardins et al., 1997; Jerger et al., 2009).

Many researchers have endeavored to measure variability in weighting of modalities, as how to optimally combine modalities via the implementation of incongruent stimuli (e.g., the McGurk illusion, McGurk & MacDonald, 1976), or the flash-beep illusion, Nava & Pavani, 2013; Petrini et al., 2015). By presenting incongruent auditory and visual stimuli, the weight individuals assign to each modality can be quantified by looking at how often individuals report either the auditory or the visual stream (Magnotti & Beauchamp, 2017). However, as postulated by the causal inference problem (Körding et al., 2007), the use of incongruent stimuli can lead to different speech perception decisions (i.e., segregation vs integration). Therefore, the question arises to what extent weighting

of the incongruent streams explains the individual differences found in AV speech perception benefits for congruent speech signals.

This work contributes to the existing literature by focusing on different levels of multisensory processing in a large group of children (N = 161, age = 4-15 years old), while minimizing individual and age-related differences attributed to cognitive and linguistic complexity of the presented task. We used a simple game-like online closed-set task with spoken words acquired by age 4 years. We expect all children tested here to have equal opportunities to accurately complete the presented speech perception tasks, to be engaged throughout this fun task without introducing confounds of new environments (i.e., the lab) or new people (i.e., the researcher) while limiting influences of linguistic development (i.e., phonological processing, Jerger et al., 2009; articulatory development, Vance et al., 2005). By using multiple SNRs we allow both younger and older children to reach comparable performance in the auditory/AV modality, observe their AV speech perception benefits, and look more specifically at how individual performance in each modality contributes to the overall AV gain (i.e., weighting of the perceptual streams). The use of congruent AV stimuli allows us to focus specifically on the use of speech-specific cues in children, while the added introduction of incongruent AV stimuli informs us about the causal inference process in children, while listening to AV speech in background noise.

This set-up allowed to investigate the following questions: i) Do AV speech perception benefits differ across ages for children when the task is low in cognitive and linguistic demands, ii) can observed variability of AV speech perception benefits be explained by the weights of auditory and visual modalities, and iii) are there developmental differences in how children in this age group approach the causal inference problem (Körding et al., 2007) as they were introduced to both congruent and incongruent AV stimuli? We hypothesized that when using age-appropriate linguistic stimuli and a behavioral task low in cognitive demands, speech-specific phonological knowledge is already accessible to create AV speech perception

benefits across this entire age group. Second, younger and older children might weigh auditory and visual stimuli of the AV speech signal in a different way, and especially the use of incongruent stimuli could point to developmental differences in solving the causal inference problem.

3.2.2. Materials and Methods

The methods described below to measure audiovisual speech perception via this remote psychophysical assessment were validated in Gijbels and Lee (2023).

3.2.2.1. Participants

A group of 161 English-speaking children (M=79, F=82) participated in a remote study and completed an online AV speech perception task. The age range varied from 3.87 to 14.92 years old (μ = 8.39, SD = 2.78). All children were recruited via a pre-existing database or the Communication Studies Participant Pool at the University of Washington. Parents of all participants provided written and/or video-recorded informed consent under a protocol that was approved by the University of Washington's Institutional Review Board. All methods were performed in accordance with the relevant guidelines and regulations of this Institutional Review Board approval. Typical speech, language, hearing, and vision (normal/corrected to normal) development was reported via a parental questionnaire. Children with any diagnosis of a developmental disorder were excluded. All children had English as their first language. The sample in this study was aimed to be representative of the US/Washington State population.

3.2.2.2. Stimuli and materials

The paradigm of this AV speech-in-noise recognition task was based on the task used in Gijbels et al. (2021). Twenty of the 25 previously determined stimuli were used in the current experiment. The number of stimuli was limited to 20 (out of 25), to keep a reasonable total number of trials (N_{trials}=160, plus 10 catch-trials) for

children in this age range (4 to 15 years old), while the number of different conditions increased compared to the previous experiment, and to remove items that were likely to cause ceiling performance (Gijbels et al., 2021). As shown in Gijbels et al. (2021) the number of different stimuli used (10 vs 25) did not result in a significantly different outcome, and therefore the change to 20 stimuli should not have a direct effect. All target and foil words used for this experiment were consonant-vowel-consonant (CVC) words. The target words originated from the CHILDES database (MacWhinney et al., 1985) and are acquired at ages 3 to 5 (Holt et al., 2011). All stimuli were recorded by a female native English speaker by Holt et al. (2011) and previously used by Lalonde and Holt (2016) and Gijbels et al. (2021). All words were of low complexity, in both meaning and word form. For the background noise we used PRAAT (Boersma & Weenink, 2021) to create random 2-s fragments of a 3 minute long-term average speech spectrum (LTASS) of 100 LNT CVC-word recordings from the same female talker. The selection of three SNR levels: -5 dB, -8 dB, or -11 dB allowed us to compare over multiple SNRs and to give children of all tested ages success experiences; where an SNR of -5 dB might be easy for the oldest children, an SNR of -11 dB might be difficult for the youngest children (Elliott, 1979; Massaro et al., 1986).

The stimuli were presented in 1 of 4 modalities (auditory-only, visual-only, congruent AV, or incongruent AV). The congruent AV modality consisted of the target word in noise with a synchronous matching video of a woman speaking the target word (N_{trials}=60). The visual modality was only the video and added noise, there was no auditory signal of the target word (N_{trials}=20). The stimuli in the auditory modality (N_{trials}=60) consisted of the target word in noise accompanied by a picture of the woman's face, in a neutral position. This still image was a frame selected from one of the target video recordings. Twenty trials were assigned to a second, incongruent AV condition. Here the target word in noise was combined with a synchronous silent video of a different word (semantic incongruency). This was a

video of a word of the same duration, with the same onset time, that had no phonemes or visemes in common with the target word and that was not part of any of the presented stimuli (e.g., audio: bath vs. video: song).

Picture selection, generation of stimuli, response layout, and the relationship between target word and foils were all similar to Gijbels et al. (2021). A one-interval four-alternative forced choice paradigm (1I-4AFC paradigm) was used, and all pictures were retrieved from an open-source clipart database. The four images displayed presented either the target word, a minimal pair foil (first or last consonant differed), a foil with the same vowel as the target word (but different consonants), or a random CVC foil (with no phonemic similarity to the target word). Most targets served as foils for other targets (e.g., "sit" was a target, a minimal pair foil for target "sing," a vowel foil for target "pink", and a random foil for the target "gum").

The minimal pairs were auditory minimal pairs, meaning they were distinctively different acoustically but not necessarily visual. For example, you can easily hear the difference between "hold" and "cold" in a quiet environment, but it is hard to notice visual differences between someone articulating "hold" and "cold," in the absence of any auditory input. All minimal pair words were scored according to their visual similarity, ranging from 1 (most similar) to 4 (least similar), based on confusion matrices of previous research (Owens & Blazek, 1985). The task was designed in the free online study builder, Lab.js (Henninger et al., 2021) and hosted via Pavlovia; an online hosting platform created by Psychopy (Peirce et al., 2019) to independently run online studies.

3.2.2.3. Procedures

Parents were directed to load the task in their web browser (in Google Chrome or Firefox) via a link created by Pavlovia and a uniquely identifiable code. Children were instructed to be seated at a table in a quiet environment, facing the computer

screen, and parents were asked to stay close to the child during the task. Stimuli were presented in free field through the computer's speakers (no headphones were allowed) and answers were recorded via mouse clicks. This could be initiated by the child, but parents were allowed to click the picture the child pointed at if the child was not fluent in handling the computer.

Testing via the participants' web browser did not readily allow them to control the absolute audio intensity. To make stimuli as similar as possible across participants, they were instructed to set the computer's speakers to a comfortable level at the beginning of the task and to not change it during the task. This level was based on a repeating English speech fragment and played for 7,000 ms. Due to the uncontrollable nature of the exact sound level of online testing we acknowledge that our SNRs might not exactly be -5 dB, -8 dB, or -11 dB. Nevertheless, we will keep referring to them like this for convenience of understanding.

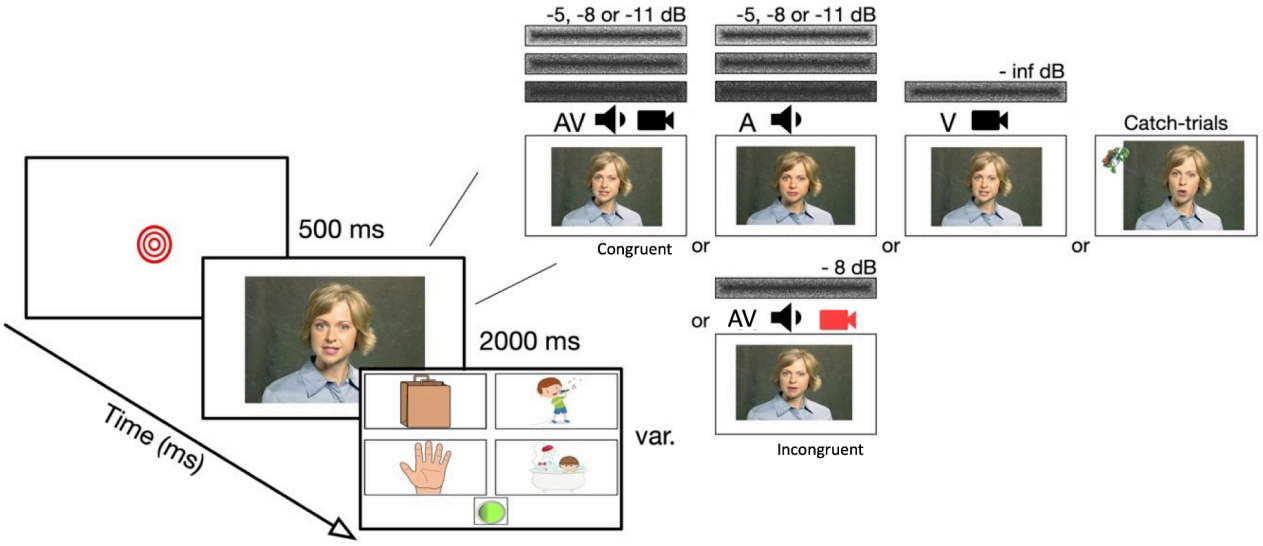
A two-part training session preceded the speech-in-noise task. First, all 30 clipart pictures (target words and foils) used during the experiment were shown to the participants for 1,600 ms, accompanied by the picture name (audio). These words were recorded by a female native English speaker, different from the speaker of the stimuli for the actual tasks. Second, the participants were familiarized with the AV speech task by exposing them to eight trials with feedback. Four stimuli (different from the actual task) were used, with an SNR of 0 dB and -5 dB to focus on learning the task. The full task, including training, could be completed in 20-25 minutes. Both the training and the task itself were narrated by a cartoon character. This character would give the instructions verbally, while they were also presented in writing. It provided motivational words during the breaks and served as an attention foil during the experiment.

The experimental task consisted of 160 trials divided into 8 blocks, with 10 catch-trials randomly added within the blocks. Participants were offered a break

after every block and could continue whenever they felt ready. All trials were randomized per participant, before the start of the task and presented in a mixed sequence. They consisted of 60 congruent AV trials (20 for each SNR level), 60 auditory trials (20 for each SNR level), 20 visual trials, and 20 incongruent AV trials (only presented at -8 dB SNR). The remaining 10 trials were catch-trials, in one of three modalities (auditory, visual or congruent AV). In these catch trials, an image of the cartoon character would appear in front of the video/picture and serve as a control for visual attention to the screen at all times. Every trial started with a fixation target (500 ms) followed by a 2-second-long noise with the word stimulus appearing 500 ms after the beginning of the noise. Then, a screen with four pictures (arranged in a 2x2 box) appeared. The participant selected a response by clicking one of four (positionally randomized) images. Whenever the character showed up during the catch trials, the participant had to click a green button at the bottom of the screen. This was taught during the training session. No feedback was provided throughout the task and there was no time limit to respond. A visualization of the experiment is shown in Fig 3.7.

To establish other potential factors influencing individual variability of AV speech perception benefits like phonological processing (i.e., Jerger et al., 2009), we also collected phonological awareness scores (Comprehensive Test of Phonological Processing, CTOPP-2; Wagner et al., 2013) via a video conferencing call. This task was assessed as similar as possible to in-person testing. The Phonological Awareness Score was determined by 3 subtests: elision, blending, and phoneme isolation/ sound matching (depending on the age).

FIGURE 3.7. Visualization of the Speech-in-Noise task.



The task started with a 500 ms fixation marker, then the Auditory, congruent AV, incongruent AV, or Visual stimulus was presented (in different SNRs, 2,000 ms), followed by the four answer choices. There was no time limit in the response period.

3.2.2.4. Analysis methods

Statistical analyses were performed using the lme4, lmerTest, stats, and psych packages in R (Bates et al., 2015) and RStudio (version 1.3.1093).

Planned analysis

We used a linear mixed effects model; *Model 3.3: Test score ~ Modality * Age + SNR + (1|Participant)*, to test our hypothesis that children between 4 and 15 years old perform significantly better in the AV modality than the auditory modality, and to test whether age is a significant effect. The dependent variable was the task performance in the congruent AV and auditory trials, for all SNRs. We examined the fixed effects of the modality (i.e., congruent AV vs. auditory), age (as a continuous variable, in years), and SNR level (i.e., −5 dB, −8 dB, and -11 dB). Age was included as an interaction term since we expected older children to score higher in all modalities (Kyle et al., 2013; Massaro et al., 1986; Taitelbaum-Swead & Fostick, 2016; Tye-Murray et al., 2014). Gender was initially added to the model as an interaction

term since in our previous work (N=37; Gijbels et al., 2021), we found a significantly lower performance in 6-7-year-old boys in the auditory modality, compared to equal performance for both boys and girls in the congruent AV modality in an almost identical task. However, as our results here showed no effect of gender, we excluded this variable to simplify the model and make it easier to interpret. We included the different SNRs as a main effect, since we expected performance in individual modalities to be better for higher SNRs (Erber, 1969, 1975; Ross et al., 2011; Sumby & Pollack, 1954). We did not include SNR as an interaction term since we had no specific hypothesis about these specific SNRs. The auditory modality, and −8 dB SNR served as references in each category. The model included a random intercept for participant, which estimated a variance component for the fixed factors such that the model fits an intercept for each participant. To look more specifically at individual modalities (auditory, visual, and congruent AV) in relation to age we calculated Pearson correlation coefficients.

Model 3.4. focused more specifically on AV speech perception benefit by age, and other potentially influencing factors such as gender, phonological awareness skills, and cross-modal attention scores. As suggested by Alsius et al. (2016), Grant & Seitz (1998), and Sumby & Pollack (1954), we did not use the simple difference score between AV and auditory modalities to calculate the AV speech perception benefit. This would create a bias that high auditory scores necessarily lead to low benefit scores. In order to also account for the well-known observations that children do better in speech-in-noise tasks as they get older (Leibold & Buss, 2019; McCreery & Stelmachowicz, 2011; McMurray et al., 2018; Wightman et al., 2006, we used the normalized AV gain [(AV−A)/(1−A)] (Alsius et al., 2016; Grant & Seitz,1998; Sumby & Pollack, 1954).). *Model 3.4: Normalized AV gain ~ Age + Gender + Phonological awareness scores + cross-modal attention,* with normalized AV gain as the dependent variable, age, gender, phonological awareness scores, and attention as independent variables. Cross-modal attention was determined by counting how often (out of 10 catch-trials)

the participant noticed the cartoon character popping up on the screen. Four participants were excluded from this model since they had no phonological awareness scores.

We calculated the Bayes factor of the normalized AV gain to quantify the support for the null hypothesis that AV benefit is stable across the sampled age. A Bayes Factor of < .33 yield evidences for the null hypothesis (H_0) being favored over the alternative hypothesis (H_1), whereas .33 - 3 shows data that is too weak to show favor for H_0 or H_1, and > 3 favors H_1 (Biel & Friedrich, 2018).

The last factor of interest was the impact of the incongruent AV modality. Since the incongruent AV modality was only assessed in SNR -8 dB, and we expected performance to be similar to auditory performance, we compared performance (in %) in the auditory modality of SNR -8 dB, with the performance of the incongruent AV modality. A linear mixed-effects model, *Model 3.5: Test score ~ Modality(A_{-8dB} or $AV_{incongruent\text{-}8dB}$) * Age + (1|Participant)*, was used to predict performance based on the modality (auditory or incongruent AV) by age. The model included a random intercept for participant, since there were 2 test scores for each participant. For both modalities, a linear model was created (*Modality ~ Age*). An ANOVA was used to determine if both conditions were significantly different across ages.

Pairwise comparisons using Wilcoxon rank sum test and false discovery rate using Bonferroni correction were used to look at speech-specific cues that might contribute to AV benefits when comparing performance and visual similarity between minimal pairs in different modalities.

Post-hoc analysis

Our results generated the hypothesis that individual auditory performance, relative to the range of the SNR performance, determines the size of the AV speech perception benefit. To further explore these findings, we created a new continuous variable, the relative auditory performance. By calculating individual z-scores of the

auditory scores, per SNR (e.g., [% correct auditory score of participant $X_{SNR\ -5\ dB}$ − mean (auditory $scores_{SNR\ -5\ dB}$)]/sd (auditory $scores_{SNR\ -5\ dB}$)), we could look at low and high performance within each range of performance rather than looking at overall performance. For example, where 50% correct would be low auditory performance for SNR = -5dB, it would be in the middle range for SNR = -11dB. By normalizing our auditory score, low auditory performance would get a value of -3, whether this was 50% correct for -5dB or 20% correct for -11dB.

As our interest here is specifically in the AV gain compared to the relative auditory performance, we used the normalized AV gain. Furthermore, we calculated the AV gain per SNR. A linear mixed-effects model, *Model 3.6: Normalized AV gain (per SNR) ~ Relative auditory performance + SNR + Age + visual performance + (1|participant)* was used to predict normalized AV gain based on individual relative auditory performance, SNR, age, and visual performance. The size of the effect of each factor was determined by standardized β coefficients. The model included a random intercept for participant, which estimated a variance component for the fixed factors such that the model fits an intercept for each participant.

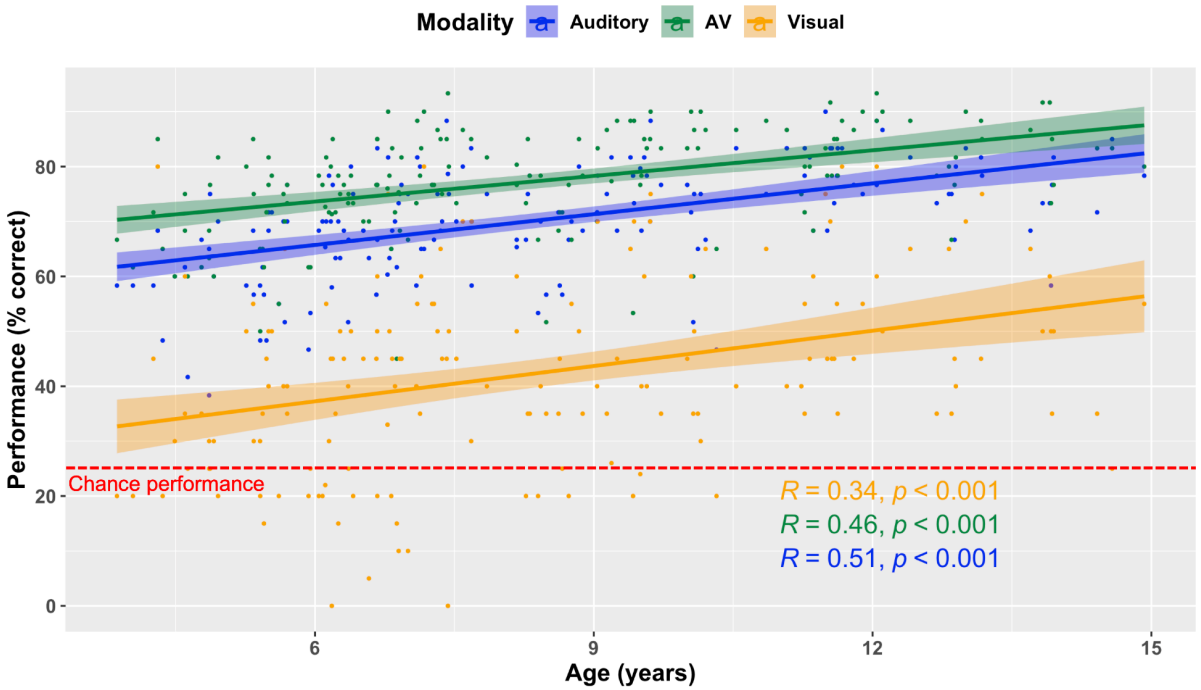
3.2.3. Results

3.2.3.1. Development of AV speech perception

A group of 161 typically developing English-speaking children (3.87-14.92 years old) completed an online AV speech perception task (Gijbels et al., 2021), and we were interested in the AV speech perception benefit as captured by the normalized AV gain measure (henceforth also referred to as AV gain) as the children were tested at different SNRs and how this gain varied with age. We tested the hypothesis that children between 4-15 years old already perform significantly better in the AV modality than in the auditory modality, given the task used here is low in cognitive and linguistic demands (Model 3.3, Fig 3.8). Overall, these children show significantly better performance in the AV modality ($F_{1,\ 801}$ = 26.30 $p < .001$), across

SNRs and age. As expected, main effects of age ($F_{1, 159} = 62.85$ $p < .001$) and SNR ($F_{1, 801} = 294.02$, $p < .001$) were found, indicating that speech-in-noise performance improves with age and SNR levels across all modalities (Nittrouer & Boothroyd, 1990; Wightman et al., 2006). No significant interaction effect was found, indicating that the observed trends between modalities were not different for different ages.

FIGURE 3.8. Performance by age (4 – 15 years) for auditory-only, visual-only, and AV speech perception.



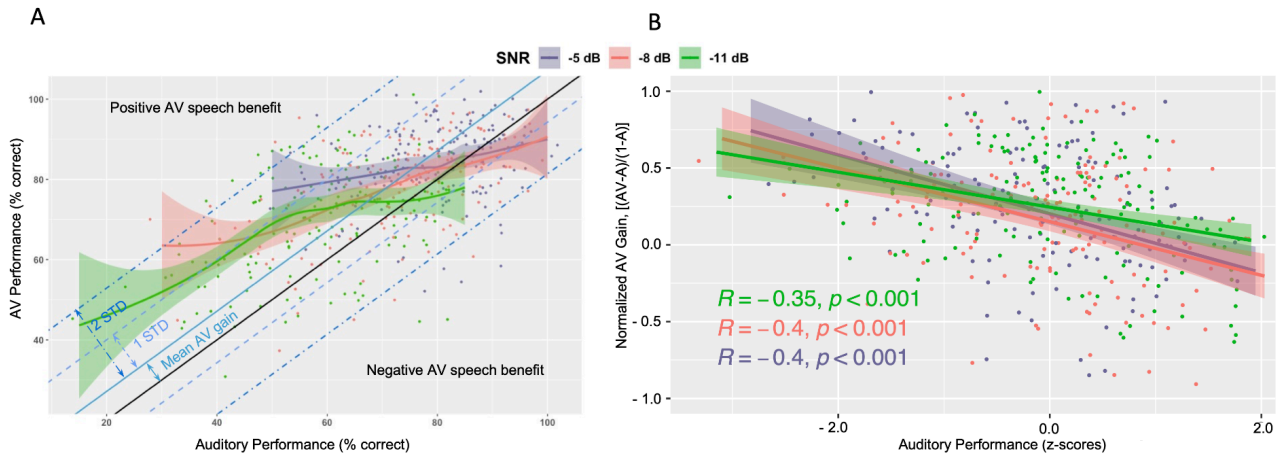
Regression lines of task performance per modality (auditory (blue), visual (orange), and AV (green)) is shown, with shaded regions representing 95% confidence intervals. Performance in every modality shows to be significantly above chance level (red dotted line - 25%) and significantly improves with increasing age ($p<0.001$), as shown in the Pearson-correlation coefficients (right bottom). Performance in all modalities is significantly different with the highest performance in the AV modality and the lowest performance in the visual modality. Individual performance is presented by dots.

To quantify the AV speech perception benefit, we used the normalized AV gain measure (c.f., Model 3.4). The absence of an age effect ($F_{1,152} = .0011$, $p = .924$), in combination with moderate proof (van Doorn et al., 2021) for the null hypothesis (Bayes Factor = 0.17) indicates that although younger children perform lower on

speech-in-noise tasks in both auditory-only and AV trials, their AV gain created by adding visual cues to the speech signal is not significantly different across ages.

Varying the SNRs allowed us to parametrically manipulate the reliability of the auditory stream in the AV speech signal. Most AV speech perception models factor in the reliability of the auditory and visual signals as a determinant for the weight assigned to information available in each modality (e.g., Ma et al., 2009; Schwartz, 2010). Thus, it is reasonable to assume that the less reliable the auditory input (i.e., lower SNR), the poorer the auditory performance and the less the auditory input stream will be weighted in the AV signal. Smaller weights assigned to the auditory modality in the AV signal will therefore mean the auditory stream explains a smaller part of the observed AV gain. Interestingly, this was not the case for the current study (Fig 3.9A). We did find, as expected, the lower the SNR, the worse the performance in auditory-only trials (data not shown). Additionally, in some cases, performance on AV trials was even worse than on auditory-only trials (Fig 3.9B, normalized gain below 0). This would not be surprising when participants are presented with relatively good SNRs (e.g., -2 dB, in Gijbels et al., 2021), especially for children, who in general have limited attention span (Dye & Bavelier, 2010). The ceiling effect of auditory trials for some participants, as described in Gijbels et al. (2021), could explain some of the AV performances at higher SNR levels (i.e., at -5 dB, and even -8 dB for older children) in the current study, however it does not explain the absence of an AV gain at low SNRs (i.e., -11 dB). For example, individuals in this study who are only at 75% correct for the auditory-only trials (at -11 dB SNR) more often fail to show an AV gain, regardless of the fact that this performance leaves sufficient room for improvement in the AV trials. Instead, we observed that individuals showed AV gains within a range of auditory and AV performance at every SNR (blue dotted lines in Fi. 3.9A). Contrary to what we expected from AV speech perception models, AV gain for children at this age is not primarily driven by the quality of the auditory signal.

FIGURE 3.9. Impact of relative auditory performance.



The relationship between relative individual auditory performance (x-axis) and AV performance (Panel A, y-axis) or Normalized AV Gain (Panel B, y-axis), for each SNR (-5, -8, and -11 dB), with shaded regions representing 95% confidence intervals and dots representing individual performance.

Panel A: AV speech perception performance in relation to relative auditory performance per SNR. This plot shows AV speech perception benefit for each SNR, by looking at the relationship of the AV performance to the Auditory performance to the black diagonal line. Loess regression lines are shaded to indicate 95% confidence intervals. The black line represents equal auditory and AV performance. The blue line represents the mean AV gain, with the dashed lines 1 STD and the dot-dashed lines 2 STD from the mean. The blue lines show the scope of the scores per SNR and every score along a blue line has similar AV gains.

Panel B: Normalized AV gain in relation to the relative auditory performance in z-scores. This relationship is very similar for the different SNRs and shows that the normalized AV gain is significantly larger for individuals with relatively low auditory performance, in contrast to individuals with relatively high auditory performance.

Instead, we found that an individual's auditory performance relative to the group (i.e., the auditory performance within the performance range of each SNR; henceforth referred to as relative auditory performance) correlates with AV gains for these children. Post-hoc analyses (Model 3.6) showed the correlation of relative auditory performance with AV gain at each SNR (Fig 3.9B). Model 3.6 showed a significant main effect of relative auditory performance ($F_{1,483}$ = 144.29, p < .001) meaning that AV gains were significantly larger for individuals with lower auditory performance within each SNR. A significant effect of visual performance ($F_{1,483}$ = 28.03, p < .001) suggests, as expected, that higher speechreading skills lead to more

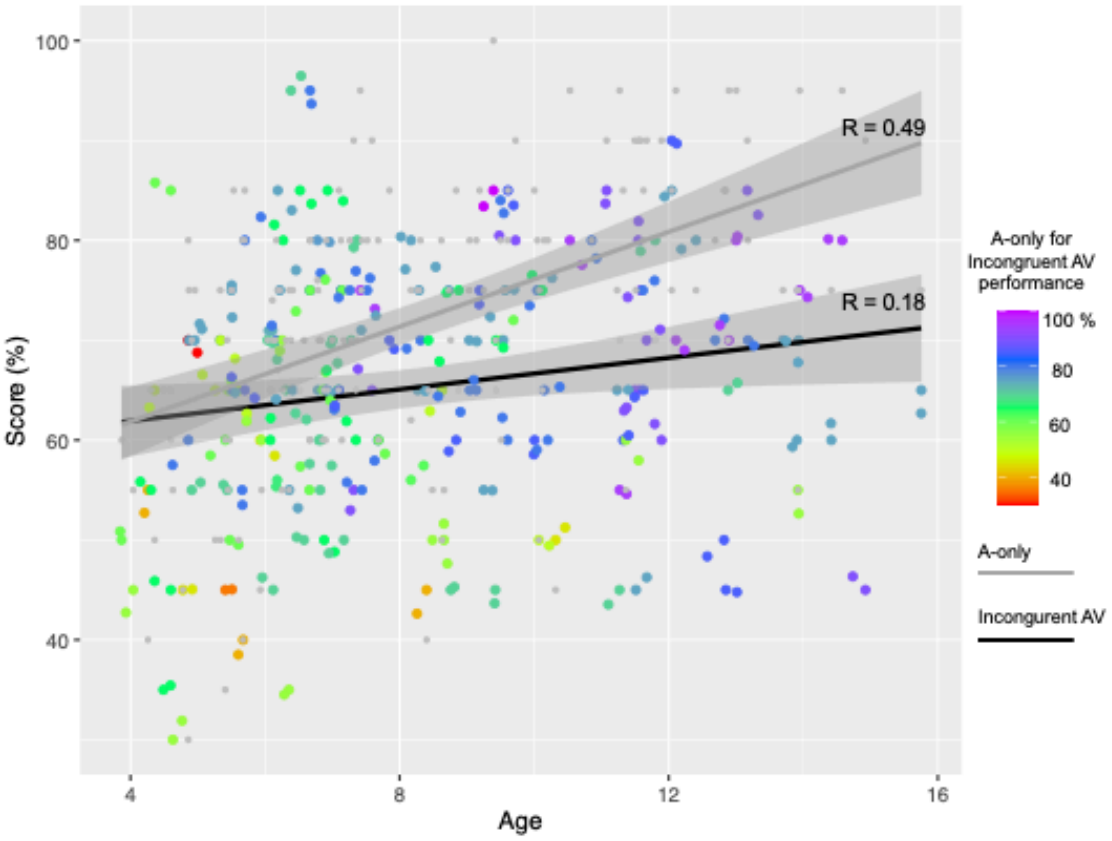
AV gain, and a significant effect of SNR ($F_{1,483}$ = 5.83, p = .003) suggests that larger AV gains were found for lower SNRs. There was no main effect of age ($F_{1,483}$ = 2.74, p > .05), again indicating that AV gains are similar for children between ages 4 and 15. Additionally, by comparing the beta values of these standardized variables in our model, we found that AV gains improve with increasing visual performance, but that the individual auditory performance is more important to determine the size of the AV gain ($\beta_{auditory}$ = -0.33 > β_{visual} = 0.15).

Thus, AV speech perception benefits increase with increasing speechreading performance, but individuals with low speechreading skills and low relative auditory performance still show larger AV benefits than individuals with high speechreading skills and high relative auditory performance, suggesting relative auditory performance within each SNR plays a more important role in explaining individual variability of AV benefits in children.

3.2.3.2. Causal inference problem and development

Finally, to explore how children solve the causal inference problem (Körding et al., 2007), we used congruent and incongruent AV stimuli to see how they are differentially impacted across ages. According to the causal inference model (Körding et al., 2007), the coherence of multimodal streams based on their temporal, spatial, and semantic cues would determine whether information should be processed together (i.e., integrated) or not. By introducing several randomly presented trials (20 out of 160) that are temporally aligned (in onset and duration) but are semantically incongruent (i.e., different CVC words presented in the auditory and visual stream), at -8 dB SNR, we created trials where integration of the auditory and visual stream should not occur, leaving the auditory stream the most reliable for children.

FIGURE 3.10. Effect of incongruent stimuli presentations.



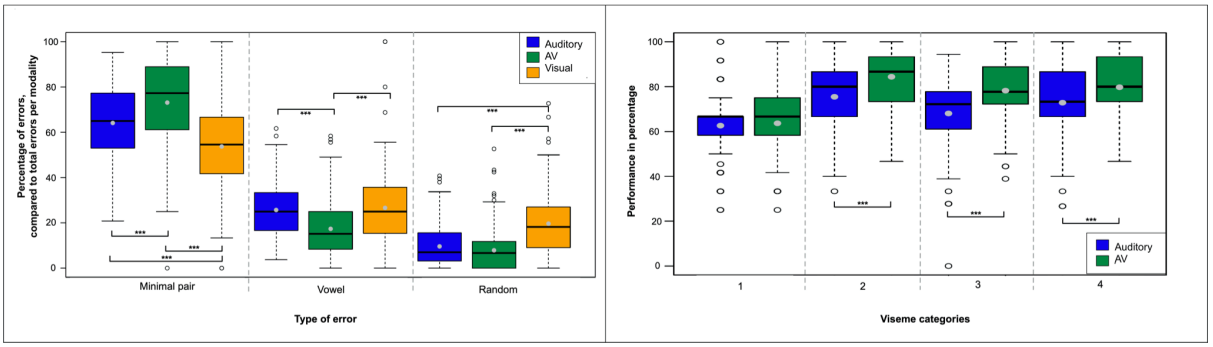
Performance (in % correct) by age (years) for 2 different stimulus conditions in SNR -8 dB. Linear fit to the data is represented by lines (auditory modality;gray vs incongruent AV modality;black) with shaded regions representing 95% confidence intervals. Individual performance is presented by dots (gray for A-only, colored by A-only for incongruent AV). The Pearson correlation coefficient shows more limited improvement with age for the incongruent AV modality, in contrast to what we saw in the auditory and congruent AV modality (c.f., FIG 10). The rainbow colors show that although younger children often have lower A-only performance, the incongruent AV performance is spread out over the whole range of performance, regardless of age.

Our linear mixed model (Model 3.5) showed a main effect of age ($F_{1,159}$ = 31.84, p < .001), suggesting both modalities significantly improved with age, and an interaction effect of modality by age ($F_{(1,159)}$ = 16.74, p < .001). This means that the age improvement in task performance that we found for auditory, visual, and AV speech perception (Fig 3.8) is significantly smaller for the incongruent AV stimuli (Fig 3.10) and suggests that there may be fundamentally different mechanisms of development involved in AV speech processing when congruent and incongruent stimuli are used.

3.2.3.3. Speech-specific cues and development

We chose the alternative choices in this 1I-4AFC task such that we could perform error analyses. These analyses showed similar error patterns to Gijbels et al. (2021) in that when children made mistakes, they more often made minimal pair errors, followed by vowel errors and then random errors. Furthermore, it was shown that this pattern was more extreme in the AV modality than in the auditory modality, indicating the visemes help to make decisions closer to the target word (Fig 3.11A).

FIGURE 3.11. Error and viseme analysis.



A B

A: Percentage of errors compared to total errors per modality (y-axis), by error type (x-axis) for Auditory (blue), AV (green), and Visual (orange) modalities.

B: Performance in percentage (y-axis) as a function of viseme categories (x-axis) for Auditory and AV modalities. Category 1 being minimal pairs with most visual similarity and category 4 being minimal pairs with least visual similarity. Significance: $p<0.05$ *, $p< 0.01$ **, $p < 0.001$ ***.

We further quantified the viseme similarity in the minimal pairs. Gijbels et al. (2021) showed that AV gain was only present for minimal pairs that had low visual similarity (e.g., “man”-“map”). Words with high visual similarity (e.g., “hold”-“cold)” did not benefit from the visual aspect in the AV trials. We now expanded our sample size, age range, and viseme categories. We used four visual similarity categories to analyze the data (with 1, highest visual similarity, and 4, lowest visual similarity between minimal pair words). This enabled us to verify that when there are little to no speech-specific visual cues available (category 1) to distinguish the minimal pair words (e.g., “hold”-“cold”), children do not show AV gain ($p = .41$; Fig 3.11B). Whereas any visual distinction (i.e., category 2,3,4) is

sufficient for children age 4 to 15 to have an AV benefit ($p < .001$) beyond the temporal congruence of the auditory and visual signals.

3.2.4. Discussion

We collected a large representative sample (N = 161) of data via remote assessment in order to clarify the developmental trajectory of AV speech perception benefits in children. Findings from previous studies have been hard to reconcile since, collectively, they have not accounted for factors related to the task difficulty or individual differences, like language and cognitive load, associated with AV speech perception performance in children at different ages. A task designed for children, low in cognitive and linguistic demands (Gijbels et al., 2021), and limited variability in performance per participant (Model 3.3), allowed us to show that AV speech perception benefits created by speech-specific cues (See Fig 3.11) are already present at age 4. Surprisingly, we showed that AV gain is stable across a broad age range from preschool through adolescence in this simple task. We further found that both auditory and visual skills are important in determining the size of the AV speech perception benefit, but ultimately, the individual's auditory performance is most predictive.

We replicated that children's auditory, visual, and AV speech perception skills improve between 4 and 15 years of age (Fig 3.8), suggesting that speech recognition performance in noise improves well into adolescence (Leibold & Buss, 2019; McCreery & Stelmachowicz, 2011; McMurray et al., 2018; Wightman et al., 2006). While we did not test adults, we could reasonably assume that adult-like performance has been reached in the upper range of our age group (12-15 years; Ross et al., 2011) and therefore we expected larger multisensory enhancement for the older children in this sample (Fort et al., 2012; Jerger et al., 2009; Lalonde & Holt, 2016; Maidment et al., 2015; Ross et al., 2011). We found a range of AV speech perception benefits, but contrary to the expectation of current literature (Fort et al.,

2012; Jerger et al., 2009; Lalonde & Holt, 2016; Maidment et al., 2015; Ross et al., 2011), this variability was not correlated with age. The absence of an age effect could potentially be explained by gamification of the task making it motivating for all ages (Desjardins et al., 1997), and offering sufficient breaks such that the decline in attention was limited (Huang-Pollock et al., 2002). Importantly, our task required limited articulatory or motor skills and thereby did not penalize younger children when we measure AV gain (Coplan & Gleason, 1988). Additionally, our task only used age-appropriate vocabulary (Holt et al., 2011) to constrain language influence in this sample (Gijbels et al., 2021). Therefore, younger or older children in this task would not intrinsically vary in retrieval of linguistic cues (Holt et al., 2011), or linguistic context (Marslen-Wilson & Zwitserlood, 1989; Taitelbaum-Swead & Fostick, 2016).

We showed that temporal synchrony between auditory and visual signals was not sufficient to create a significant AV gain in our task. AV speech perception benefits only occur when visual phonetical differences are present (Fig 3.11B) and this is not different by age, in contrast to previous findings comparing between children and adults (Lalonde & Holt, 2016). This suggests that there is not an intrinsic age difference for AV speech perception benefits in noise. However, in typical conversational contexts where factors such as attention, memory and language development would play an important role, developmental differences in AV speech perception in noise could still occur. Nonetheless, our findings put forward that the perceptual mechanism of AV speech perception benefits itself are stable from kindergarten to adolescence.

While the above findings provide important constraints on how we interpret age as a factor in AV speech perception benefits, they do not address the factors that account for the individual differences seen across ages. As suggested by some AV speech perception models (Ma et al., 2009; Schwartz, 2010), the reliability of the incoming auditory and visual streams might have an important role in the size of an

individual's AV speech perception benefit. This work shows that, at least for children, it is not necessarily the actual performance on the individual modalities (auditory or visual) that determines the size of the AV gain, but rather how this performance is relatively related to the range of performance for that noise level. This means, for example, that an auditory performance of 75% correct only leads to a benefit from the visual cues when 75% is a medium/low score for the individual (e.g., for SNR -8 dB), but not when it is already at the high end of performance (as for SNR -11 dB, Fig 3.9A and B). Therefore, contrary to what some of the AV models might suggest, the reliability of the speech signal manipulated by the SNR is not what determines the size of the benefit; rather, it is how children perform on their auditory speech perception task in noise that determines if they would leverage the visual cues (Fig 3.9B). We posit that children who are relatively good at auditory speech perception in noise may not see the need for visual cues; in contrast, those who perform poorly may try to leverage visual cues at all noise levels. It is intuitively logical that when you learn from your experience that you could trust your auditory input, you could rely less on visual cues even if it objectively could still provide you benefit. This interpretation is also in agreement with the causal inference model (Cao et al., 2019) which suggests that previous experiences influence our bias towards a modality in the decision of integration or segregation of AV speech perception, and this interpretation has implications for children with hearing loss who show greater AV benefit (e.g., Lalonde & McCreery 2020).

While speechreading skills significantly contribute to the size of the AV speech perception benefit, at least for children, the weight assigned to how differences in visual speech perception of the AV speech signal explain the AV benefit, are smaller than the weight assigned to the auditory cues (Model 3.6). These findings are supported by previous work with incongruent stimuli (e.g., the McGurk illusion (McGurk & MacDonald, 1976) and flash-beep illusion (Nava & Pavani, 2013; Petrini et al., 2015)). To further interrogate this, we added a limited amount (20/160 trials) of

incongruent AV speech stimuli to our task. Based on the causal inference model (Körding et al., 2007) we expected children to pick the most informative speech stream when presented with incongruent AV stimuli. Since children still have limited speechreading skills and we used moderate noise levels (-8 dB SNR), the most informative stream was expected to be the auditory stream. Therefore, the child's performance should be similar to their auditory performance. Interestingly, we found that when young children were presented with these incongruent stimuli, their performance was similar to the auditory speech stream, but older children performed significantly worse on incongruent AV speech than what could be expected based on their auditory performance (Fig 3.10). This could indicate either that children of all ages respond equally to incongruent stimuli, suggesting that the causal inference problem (Cao et al., 2019; Körding et al., 2007) is considered prior to the task-related perceptual age benefit (Fig 3.8), or that older children found it harder to ignore incongruent stimuli and therefore their performance is even lower than what one would expect based on the auditory performance. Taking into account the causal inference model and the stimuli we used, the latter explanation seems more parsimonious. We presented the incongruent stimuli randomly between the congruent trials. Since the stimuli were single words, the children had no other priors to assess that the auditory and visual stream should not be integrated, as the previous trials were more likely to be congruent. Older children might employ prior information more to make decisions about the AV speech stimuli (i.e., to recalibrate) and therefore be more influenced by the incongruency of both streams. This is in line with the suggestion from Rohlf et al. (2020) that multisensory integration develops prior to crossmodal recalibration. By studying multisensory integration via the ventriloquist effect, they showed evidence for dissociation of the processes involved in multisensory integration and recalibration based on priors as described by the causal inference model. Translating this to our study suggests that although AV speech perception benefits are already stable in kindergarten, crossmodal recalibration needed for incongruent stimuli is not. These findings require further

exploration of the causal inference problem, especially with children. Future studies could help separate different perceptual processes of integration mechanisms for both congruent and incongruent AV speech stimuli, and how these processes differ across development.

3.2.5. Conclusion

We conclude that children between 4 and 15 years old show no differences in AV speech perception benefits when factors such as language and cognitive load are minimized. Individual variability of AV speech perception benefits can partially be explained by how children weigh their speech-in-noise performance and is not necessarily related to the quality of the auditory speech signal. Finally, younger and older children respond differently to congruent and incongruent AV speech and therefore caution is warranted in generalizing between studies using congruent and incongruent AV speech signals.

Chapter 4

Children with Developmental Dyslexia have Equivalent Audiovisual Speech Perception Performance but their Perceptual Weights differ.

This chapter is published as Gijbels et al. (2023) with co-authors, Adrian KC Lee, and Jason D. Yeatman in the journal Developmental Science. All authors and the copyright agreement of the journal agreed to include this work as a whole in the thesis.

Abstract

As reading is inherently a multisensory, audiovisual (AV) process where visual symbols (i.e., letters) are connected to speech sounds, the question has been raised whether individuals with reading difficulties, like children with Developmental Dyslexia (DD), have broader impairments in multisensory processing. This question has been posed before, yet it remains unanswered due to a) the complexity and contentious etiology of DD along with b) lack of consensus on developmentally appropriate AV processing tasks. We created an ecologically valid task for measuring multisensory AV processing by leveraging the natural phenomenon that speech perception improves when listeners are provided visual information from mouth movements (particularly when the auditory signal is degraded). We designed this AV processing task with low cognitive and linguistic demands such that children with and without DD would have equal unimodal (auditory and visual) performance. We then collected data in a group of 135 children (age 6.5-15) with an AV speech perception task to answer the following questions: (1) How do AV speech perception benefits manifest in children, with and without DD? (2) Do children all use the same perceptual weights to create AV speech perception benefits, and (3)

what is the role of phonological processing in AV speech perception? We showed that children with and without DD have equal AV speech perception benefits on this task, but that children with DD relied less on auditory processing in more difficult listening situations to create these benefits and weighed both incoming information streams differently. Lastly, any reported differences in speech perception in children with DD might be better explained by differences in phonological processing than differences in reading skills.

4.1. Introduction

Developmental Dyslexia (DD) is the most prevalent specific learning disorder affecting 5–17% of the English-speaking population (Shaywitz, 1998, Snowling, 2000). This neurodevelopmental disorder is characterized by persistent difficulties in reading, spelling, and writing. Individuals with DD have difficulty with the multisensory process of connecting the letters they see with the sounds they represent (5th ed.; DSM–5; American Psychiatric Association, 2013). The mechanisms behind reading difficulties in individuals with DD are a continuous topic of debate, ranging from disruptions in various cognitive processes varying from temporal processing (Zurif & Carson, 1970), general processing speed (Wolf et al., 2000), attention difficulties (Pelham & Ross, 1977), inhibitory control (Hasher et al., 1999), and phonological processing (Carroll & Snowling, 2004; Snowling, 2000; Ziegler & Goswami, 2005)), to perceptual processes ranging from visual difficulties (Eden et al., 1995) to auditory difficulties (Hämäläinen et al., 2013), or a combination of both (Birch & Belmont, 1964; Fox, 1994; Mittag et al., 2012).

Reading is inherently multisensory as visual orthographic representations (i.e., letters) must be connected to auditory, phonological representations (i.e., sounds), prompting some researchers to propose that dyslexia should be considered a multisensory audiovisual (AV) deficit (Blau et al., 2010; Froyen et al., 2008; Mittag et al., 2012). Reading being multisensory, in combination with phonological processing

deficits being reported in the majority of individuals with DD (Wagner and Torgesen, 1987; Snowling, 1998) raises the question whether multisensory processes, like AV speech perception, are also impacted and potentially causally related to reading development (See Hahn et al., 2014 for review).

Audiovisual speech perception benefits arise from leveraging visual information from mouth movements to improve the perception of the speech signal. AV processing is especially beneficial in noisy situations, like restaurants, playgrounds or classrooms (e.g., Hollich et al., 2005). In noisy environments, the access to both auditory and visual streams allows us to use information from the visual cues to fill up the gaps that background noise creates in the auditory stream (Grant et al., 1998; Sumby & Pollack, 1954). Whether individuals with DD show similar AV speech perception benefits as their typical developing peers is still up for debate. Some studies report significant group differences in AV speech perception (benefit) (e.g., Ramirez & Mann, 2005; van Laarhoven et al., 2018; Rüsseler et al., 2015), while others find no differences (e.g., Baart et al., 2012; Francisco et al., 2017; Megnin-Viggars & Goswami, 2013). It is often reported that children with DD perform worse on auditory tasks, especially when they are presented in background noise (Boets et al., 2011; Bradlow et al., 2003; Cunningham et al., 2001; King et al., 2002; Ziegler et al., 2009), and that children with DD have lower speechreading performance, meaning that they glean less of the information from mouth movements when no sound is presented (de Gelder & Vroomen, 1998; Ramirez & Mann, 2005). This raises the question of whether we can speak of multisensory deficits as children with DD already show worse unimodal performance for both the auditory and visual streams. Previous work has tried to answer this question by creating tasks resulting in equal unimodal performance between the two groups, but results remained inconclusive (Groen & Jesse, 2013; van Laarhoven et al., 2018).

These persisting inconclusive results might be related to the specific stimuli and research methods used in previous studies. For example, there could be multiple

reasons for not finding differences on any specific task even if there is an AV speech processing impairment (Francisco et al., 2017): the measure might not be sensitive enough to distinguish between groups when the listening conditions are too easy (e.g., in quiet) (Rüsseler et al., 2015; van Laarhoven et al., 2018); the impairment might only occur for meaningful words and sentences but not for the often-used nonsense syllables; or various compensatory mechanisms might allow for normal task performance despite impairments in certain underlying processes (Gelbar et al., 2018). Reported differences in task performance could also arise from task demands that differentially impact individuals with DD. For example, creating a task where instructions must be read might create lower task understanding for individuals with DD, or tasks with high vocabulary and/or cognitive demands might measure language (e.g., Chen et al., 2017; van Viersen et al., 2017) or attention (e.g., Pelham & Ross, 1977; Facoetti et al., 2003) rather than AV speech perception per se. The current work addresses these issues by utilizing a straightforward task with easy visual representation, auditory instruction, simple language, and limited task demands.

Another way of approaching this topic is to look at how individuals weight the auditory and visual input streams during AV speech perception. These weights in the multisensory speech stream are often quantified by using incongruent stimuli like the McGurk effect (McGurk & MacDonald, 1976). The McGurk effect creates an illusion by presenting different/incongruent speech signals in the visual (e.g., /ga/) and auditory (e.g., /ba/) modality, which can lead to the perception of an entirely different speech signal (e.g. /da/). Although there are limitations to generalizing findings from incongruent stimuli to congruent speech (Van Engen et al., 2017), it is worth mentioning that when children with DD are exposed to incongruent stimuli, they report the visual stimulus more than the control group (Hayes et al., 2003; Mohammed et al., 2006). It feels counterintuitive that speechreading performance is lower for children with DD (de Gelder & Vroomen, 1998; Ramirez & Mann, 2005), yet these same individuals rely more on the visual stimulus when making decisions

about uncertain incongruent stimuli. This contradiction could be explained by adding phonological processing to the equation (Francisco et al., 2017). Many individuals with DD struggle with phonological processing (Snowling, 2000; van Bergen et al., 2014). Phonological processing is fundamental for mapping letters (graphemes) and sounds (phonemes) and is an excellent predictor for reading success (Bast & Reitsma, 1998). When looking at the relationship between phonological processing and AV speech perception performance, Francisco et al. (2017) argued that phonological processing is inversely related to speechreading performance. A greater reliance on visual speech strives to compensate for phonological processing deficits (that may originate in the auditory stream).

The current work uses a task with low cognitive and linguistic demands to explore how AV speech perception benefits develop in children with and without developmental dyslexia. By specifically exploring the impact of auditory and visual performance for congruent AV speech we try to find a better explanation for the variable AV speech perception outcomes. And lastly, we are interested to see what the role of phonological processing in the purported relationship between reading ability and AV speech perception is.

4.2. Results and Discussion

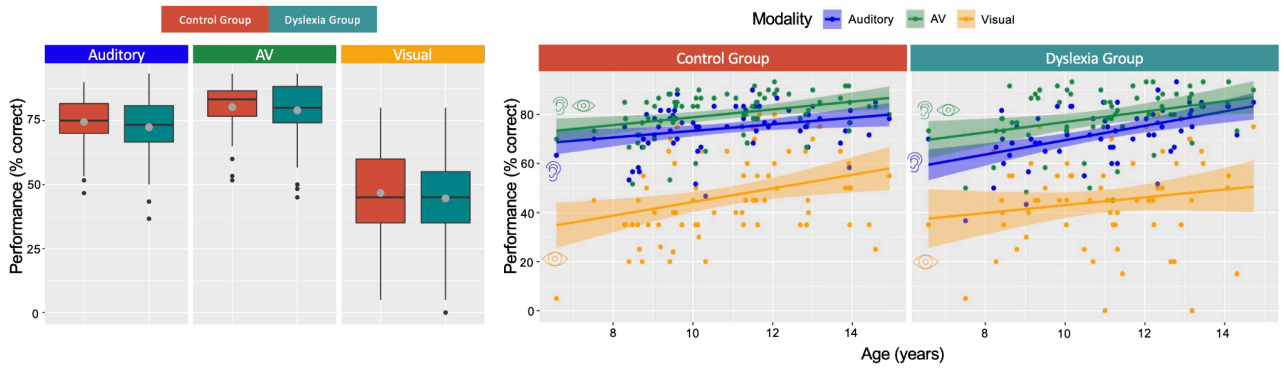
The first goal of this work is to address how AV speech perception benefits develop in children with and without developmental dyslexia (DD). Earlier work has shown that individuals with DD often perform worse on unimodal auditory or visual speech perception tasks compared to control participants (Hayes et al., 2003; Gelder & Vroomen, 1998; Ramirez & Mann, 2005; Boets et al., 2011; Bradlow et al., 2003; Cunningham et al., 2001; King et al., 2002; Ziegler et al., 2009); but also see (Groen & Jesse, 2013; Hayes et al., 2003; van Laarhoven et al., 2018). To be able to look at potential differences in AV speech perception benefits, we aimed to create a task that would result in similar auditory and visual performance for both DD and control

participants (similar to Groen & Jesse, 2013; van Laarhoven et al., 2018). We designed a task with low cognitive and linguistic demands (Gijbels et al., 2021), and manipulated the SNR of the speech signal to make the task sensitive to individual differences (Rüsseler et al., 2015; van Laarhoven et al., 2018). A group of 135 English-speaking children between 6.5 and 15 years old participated in this online study. Results from our linear mixed model (*Model 4.1:* $Performance_{(\%correct)} \sim Group_{(control,DD)} * Modality_{(A,AV,V)} * Age_{(years)} + (1|participant)$, See Methods for details) showed, as expected (Fig 4.1A & B), that performance for children across this age range is significantly different by Modality ($F_{2,262} = 15.26$, $p < .001$): performance was highest in the AV trials, followed by the auditory trials and lowest in the visual trials (similar to Ross et al., 2011; Gijbels et al., 2021). There was also a main effect of Age ($F_{1,131} = 21.14$, $p < .001$, Fig 14B.), indicating that auditory, visual and AV speech perception performance is better for older children (similar to Wightman et al., 2006; Leibold & Buss, 2019; McMurray et al., 2018; McCreery & Stelmachowicz, 2011; Ross et al., 2011). To our primary research question, there was no effect of Group ($F_{1,131} = 0.30$, $p = .59$, Fig 4.1A), meaning performance was not significantly different between the DD and control group. Moreover, no group by modality or group by age interactions were significant. The absence of an interaction effect indicated that we succeeded in creating a task with similar auditory and visual speech perception scores for both study groups.

As suggested by Rüsseler et al. (2015), differences in auditory performance between children with or without DD might be specific for scenarios where the signal-to-noise ratio (SNR) is low. We examined this by analyzing auditory and AV performance by SNR (*Model 4.1b:* $Performance_{(\%correct)} \sim Group_{(control,,DD)} * Modality_{(A,AV)} * Age_{(years)} * SNR_{(-5,-8,-11\ dB)} + (1|participant)$, See Methods for details), and found that auditory (and AV) performance was affected by SNR ($F_{2,655} = 9.37$, $p < .001$). However there was no group by SNR interaction ($F_{2,655} = 1.95$, $p = .14$). These results indicate that performance was impacted by noise, but even for more challenging listening

situations (i.e., SNR = -11dB), children with and without DD showed similar auditory speech recognition scores, for this task (Fig 4.2, left top).

FIGURE 4.1. Task performance between individuals with and without DD.



(A) (B)

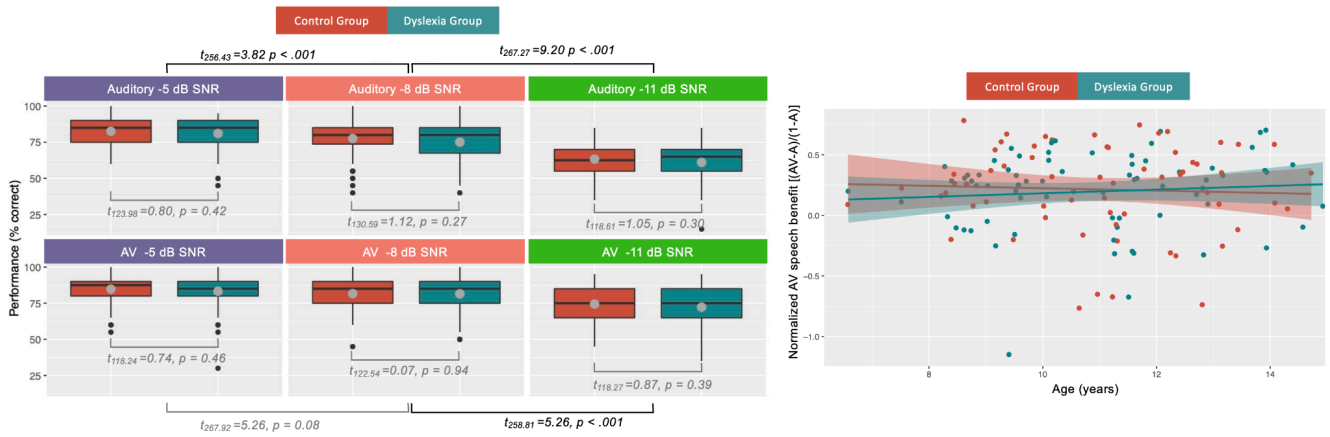
A. Task Performance (y-axis; in percentage correct) per group per modality. Performance is different per modality, but not per group.

B. Task Performance (y-axis; in percentage correct), by Age (x-axis; in years) for both children with DD and the control group, in each modality (Auditory, Visual and AV). Dots are scores per individual and modality and 95% confidence intervals are provided. Results show an effect of modality and age, in performance, but no significant difference between groups and no interaction effects.

The similar baseline performance in both individual modalities (auditory and visual), allowed us to inspect the AV speech perception benefit (*Model 3.2: Normalized AV Gain* $_{([AV-A]/[1-A])}$~ *Group*$_{(control,DD)}$* *Age*$_{(years),}$ See Methods for details). We were interested to see whether the size of the benefit - *the improvement of speech perception scores in the AV trials compared to the auditory trials* - is different for children with and without DD, and whether there was an effect of age. As no interaction effect was reported in Model 4.1, it was not surprising that the AV speech perception benefit was not different across ages or between the two groups (Fig 4.2, right). The summary statistics of Model 4.2 showed a significant AV speech perception benefit in both groups ($t_{control\ (71)} = 5.06$, $p < .001$; $t_{DD\ (62)} = 4.55$, $p < .001$), however an ANOVA of Model 4.2 clarified this effect is not different by Group ($F_{1,131} = 0.09$, $p = .76$), or by Age ($F_{1,131} = 0.06$, $p = .81$). The absence of an effect of Group, in combination with moderate proof (van Doorn et al., 2021) for the null hypothesis (Bayes Factor = 0.19), indicated that when children with or without DD are presented with a task low in

cognitive and linguistic demands, and they have similar baseline performance in either the auditory or visual modality, there is no difference in AV speech perception benefits (similar to Baart et al., 2012; Francisco et al., 2017; Megnin-Viggars & Goswami, 2013).

FIGURE 4.2. Task performance and AV benefit per group.



Left: Box plots of group performance (Dyslexia Group or control Group, in % correct), by modality (Auditory or AV) and SNR (-5 dB, -8 dB or -11dB), showing there is an effect of SNR, but performance within each SNR is not different between groups. Black lines are medians, grey dots are means, single black dots are outliers.

Right: Normalized AV speech perception benefit by Group (Dyslexia Group or control Group) and Age (in years), showing there is no difference in AV benefit between the two groups. Linear regression lines are plotted with 95% confidence intervals and individual scores are represented as dots.

Therefore, we conclude that both groups showed a similar development of auditory, visual and AV speech perception (Model 4.1), that both groups were similarly impacted by different SNRs (Model 4.1b), and that children with DD created AV speech perception benefits in degraded auditory signals to a similar extent as the control group (Model 4.2). Supplemental materials of the published manuscript (https://doi.org/10.1111/desc.13431) further report that both groups of children (DD and control) showed similar error patterns and that they both equally needed visual distinction provided by speech-specific cues, to have an AV benefit beyond the temporal congruence of the auditory and visual signals.

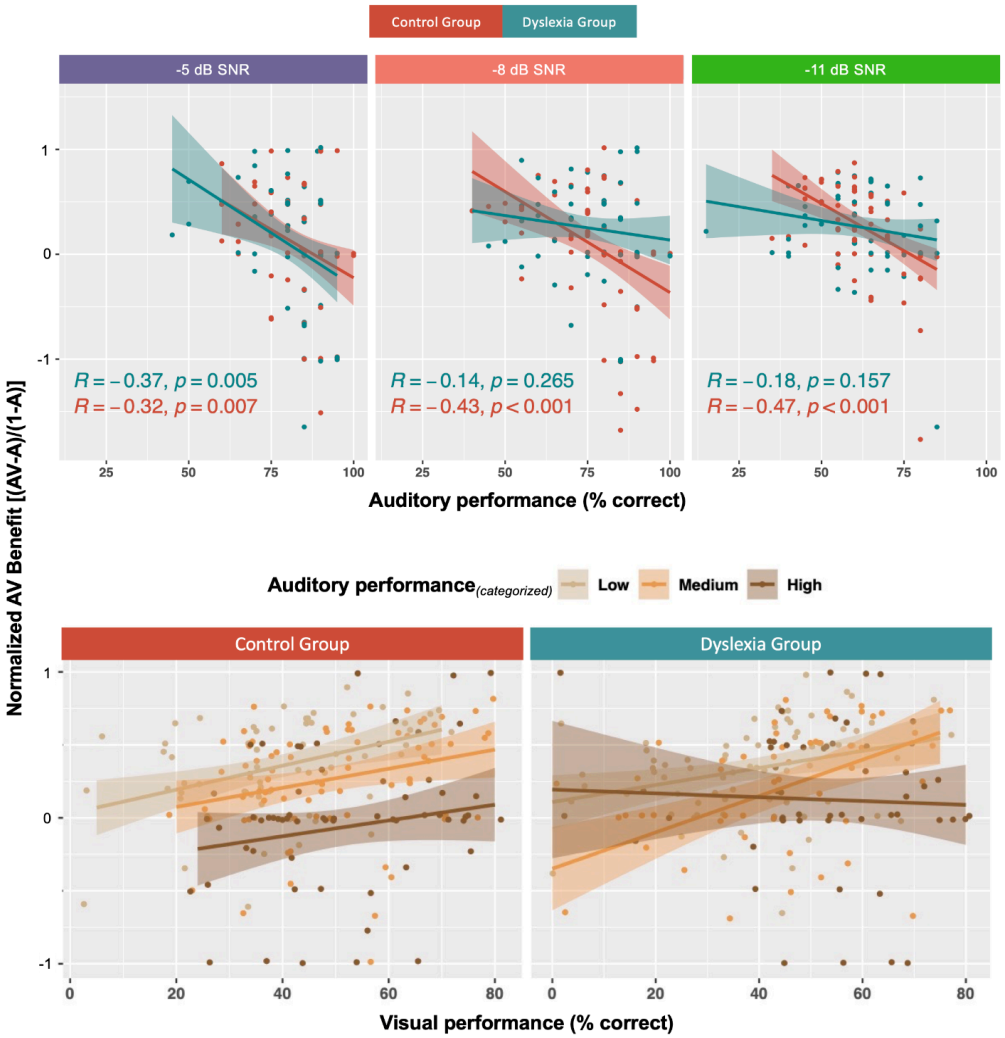
Next we explored the effect of auditory and visual speech recognition performance in relation to AV speech benefits to find a better explanation for the variable AV speech perception outcomes reported in earlier studies (e.g., Ramirez & Mann, 2005; van Laarhoven et al., 2018 vs. Baart et al., 2012; Francisco et al., 2017; Megnin-Viggars & Goswami, 2013). Studies with AV illusions show that children, in general, favor the auditory stream over the visual stream (e.g., the McGurk illusion (McGurk & MacDonald 1976; Hockley & Polcka 1994), flash-beep illusion (Petrini et al., 2015; Nava & Pavani, 2013)). Interestingly, children with DD often struggle with phonological processing, which is a high-level auditory skill, and therefore DD participants might not favor the auditory stream as much (Francisco et al., 2017). Given the limited generalizability of AV illusions to daily AV speech perception (Alsius et al., 2018; Van Engen et al., 2017), we specifically assessed how auditory and visual speech perception performance predict congruent AV speech benefits by group (*Model 4.3: Normalized AV Gain* $_{([AV-A]/[1-A])}$ *per SNR ~ Group*$_{(control,,DD)}$ * *SNR* $_{(-5,-8,-11\ dB)}$**Auditory performance*$_{(scaled\_scores)}$ * *Visual Performance*$_{(scaled\_scores)}$ + *(1|participant)*, see Methods for details & Fig 4.3). Overall, we found a main effect of auditory performance on the AV speech perception benefit ($F_{1,328}$= 89.00, $p < .001$). More specifically, for the control group the auditory performance was negatively correlated ($R_{control,-5dB} = -0.32$, $p < .01$; $R_{control,-8dB} = -0.43$, $p < .001$; $R_{control,-11dB} = -0.47$, $p < .001$) with the AV speech perception benefit, regardless of the SNR. This indicated that when a child has a high auditory performance within a certain SNR, and therefore could be confident about their auditory speech perception, they did not use the additional visual cues in the AV speech stream and did not show large AV speech perception benefits. However, for the children with DD this was only true for the highest SNR ($R_{DD,-5dB} = -0.37$, $p < .01$) (Fig 4.3, top). This means that although both groups overall showed similar AV speech perception benefits (Fig 4.2, right), they might rely on different cues to accrue this benefit. Typically developing children with high auditory performance did not use visual cues for AV speech perception benefits, but children with developmental dyslexia only showed similar behavior for

relatively good signal-to-noise ratios. Once the AV speech signal became more degraded (i.e., lower SNRs) DD participants were less affected by auditory performance in weighing visual cues. This showed as a three-way interaction (*Auditory performance * Group * SNR*) in our linear mixed effects model ($F_{2,308}$ =3.41 , p = .03).

Visual performance (i.e., speechreading skills) itself is also predictive of AV speech perception benefits (Macleod & Summerfield, 1987). When an individual is good at speechreading using visual cues, they might be able to benefit more from AV speech compared to an individual with poor speechreading skills. For our sample, this was true for both children with and without DD, and showed as a main effect (Model 4.3) of visual performance ($F_{1,198}$ = 21.06 , p < .001). However, we were particularly interested in how children weighted both auditory and visual cues to create AV speech perception benefits. As we described earlier, the auditory performance within each SNR predicted the size of the AV benefit. Thus, we grouped participants based on auditory performance into low, medium, and high performing groups (within each SNR) and then looked at how speechreading performance predicted AV speech perception benefits (Fig. 4.3, bottom). For the control group, we noted as expected, the better the speechreading performance, the larger the AV benefit, and this held up for every group of auditory performers (low, medium, and high). Additionally, we saw that children with good speechreading performance, and high auditory performance, still had less AV benefit, than children with low speechreading performance and low auditory performance. So, for the control group, both auditory and visual performance was important in creating AV speech perception benefits. However, these children heavily relied on their auditory performance to determine whether visual speech cues will be used when presented with AV speech signals. For children with DD, this relationship was less clear. Auditory performance was only predictive of AV speech perception benefits for better SNRs (-5 dB), and although speechreading performance was predictive of AV benefit, this was not dependent on

auditory performance. This was another indication that children with DD rely less on their auditory performance in creating AV speech perception benefits.

FIGURE 4.3. Normalized AV benefit by auditory and visual performance.



Top: Normalized AV benefit [(AV-A)/(1-A)] (y-axis) by auditory performance (% correct, x-axis), per SNR (-5 dB, -8 dB, -11 dB) for both the groups (control and DD), showing that better auditory performance is related to smaller AV speech benefits for the control group in SNR, but only for the best SNR in the DD group.

Bottom: Normalized AV benefit [(AV-A)/(1-A)] (y-axis) by visual performance (% correct, x-axis), for both the groups (control and DD), split up by their auditory performance. Within every SNR, the participants were categorized as low, medium or high auditory performers. Higher speech reading performance led to higher AV speech perception benefits, but only in the control group this is related to the auditory performance.

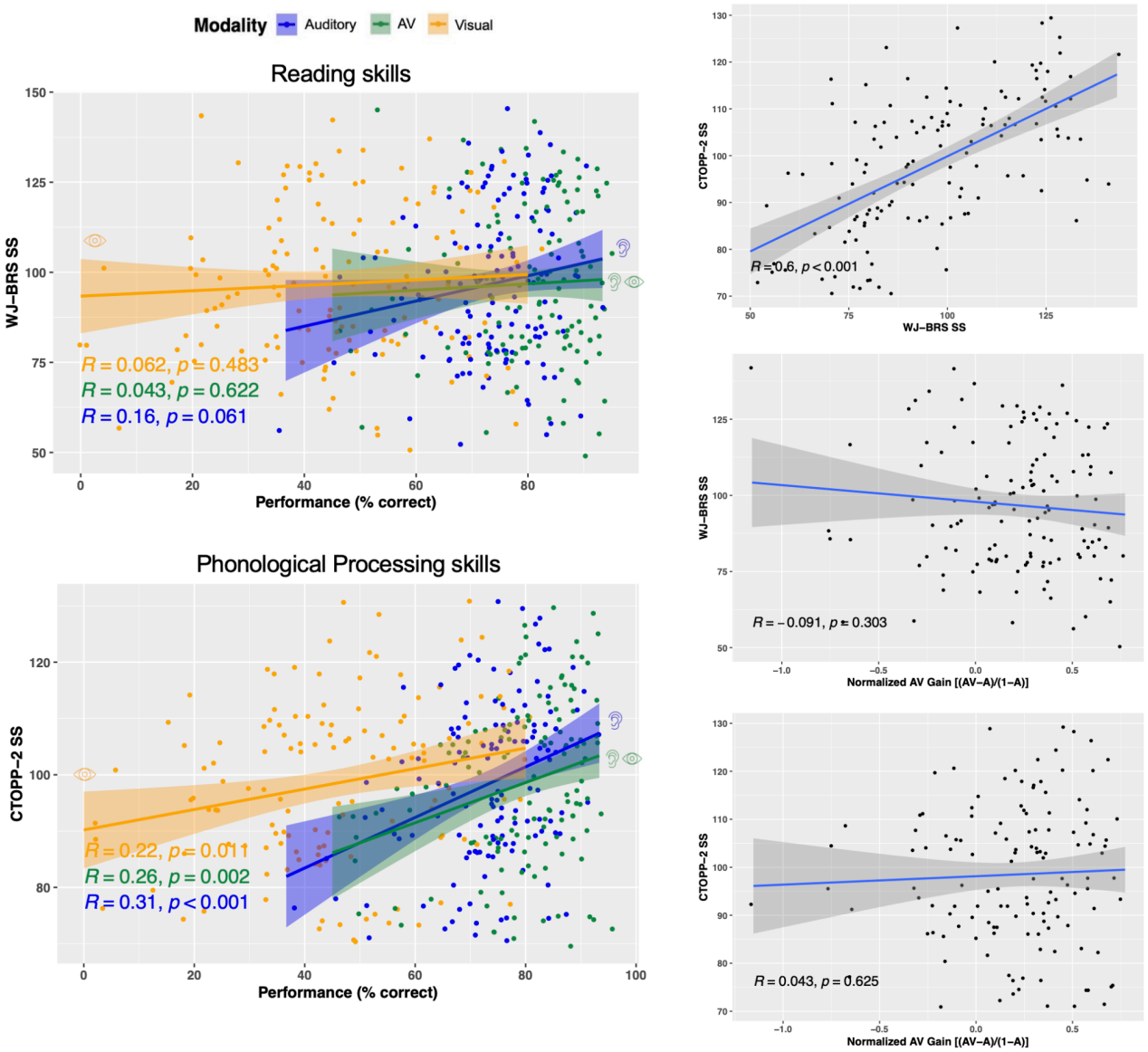
The differences described above were in line with findings from AV illusions in individuals with DD (Hayes et al., 2003; Pekkola et al., 2006), and as suggested by

Francisco et al. (2017), might therefore be explained by the individual's phonological processing skills. Francisco et al. (2017) argued that phonological processing is inversely related to speechreading performance. A greater reliance on visual speech strives to compensate for phonological processing deficits (that are mainly originated in the auditory stream).

For the children in this cohort, we collected measures of both reading performance (WJ-BRS, see Methods) and phonological awareness skills (CTOPP-2; see Methods). In general, phonological processing is predictive of reading performance (Scarborough, 1998; Snowling, 1998; Vellutino et al 2004; Wagner & Torgesen 1987); as expected, we saw a strong correlation between these skills in our group of participants (Fig 4.4, right top, R = .60). Because standardized scores are provided for both reading and phonological processing skills, we describe performance of our participants as a continuous variable, rather than breaking them up in children with and without DD. Looking at the relationship between AV speech perception benefits and reading scores or phonological awareness scores, we conclude that neither reading, nor phonological processing is predictive for AV speech perception benefits (Fig. 4.4, right middle and bottom). When we look at the task performance itself (Auditory, Visual or AV trials), a significant positive relationship is apparent between phonological processing skills and task performance in all modalities. However, this effect is absent for reading performance (Fig. 4.4, left). These results indicate that it is not necessarily individual differences in reading skills that are an impacting factor in auditory, visual, or audiovisual speech perception performance, but rather individual differences in phonological processing skills. Interestingly the AV speech perception benefit is not related to either reading or phonological processing skills, and as we found earlier (See Fig. 4.2, right), children with and without DD had similar AV speech perception benefits. This again suggested that children perceptually weigh AV speech cues differently contingent on their auditory (phonological) processing skills, but that this is not an influencing factor to the size

of their AV speech perception benefits. These results could furthermore explain why previous literature regarding AV speech perception and individuals with DD is so variable; (Hayes et al., 2003; Gelder & Vroomen, 1998; Ramirez & Mann, 2005; Boets et al., 2011; Bradlow et al., 2003; Cunningham et al., 2001; King et al., 2002; Ziegler et al., 2009); vs. (Groen & Jesse, 2013; Hayes et al., 2003; van Laarhoven et al., 2018). Although there is, in general, a moderate to strong correlation between reading and phonological processing (Scarborough, 1998; Snowling, 1998; Vellutino et al 2004; Wagner & Torgesen 1987) not all individuals with DD have poor phonological processing skills (e.g., Vidyasagar & Pammer, 2010; Ramus & Szenkovits, 2008), and this outcome might be subject/ group dependent. For example, when the group of participants with DD in a certain study/cohort have larger/more phonological processing deficits, they might show more speech perception deficits, than a group with less phonological processing deficits.

FIGURE 4.4. Audiovisual speech perception in relationship to phonological processing and reading skills.



Left: Standardized reading performance (WJ-BRS) (top) and standardized phonological awareness scores (CTOPP-2) (bottom) (y-axis) by performance (% correct, x-axis), per modality (Auditory, Visual, AV) for all participants, showing that phonological awareness scores significantly correlate with speech perception scores (A > AV > V), whereas reading performance does not.

Right:top: Correlational plot of reading performance (WJ-BRS), and phonological awareness scores (CTOPP-2) for all participants, showing a strong correlation (r = .60), between both tasks. Right: middle & bottom: Standardized reading performance (WJ-BRS) (middle) and standardized phonological awareness scores (CTOPP-2) (bottom) (y-axis), by normalized AV Gain [(AV-A)/(1-A)] x-axis) for all participants, showing no relationship between AV speech perception benefits and reading or phonological processing skills.

4.3. Conclusion

In summary, children with DD and the control group showed a similar development of auditory, visual, and AV speech perception. Both groups experienced AV speech perception benefits to a similar extent. Increasing the noise levels of the speech signal did not create group differences in the size of the AV speech perception benefit. However, both groups might weigh perceptual cues differently to accrue this benefit. Children with high auditory performance in the control group did not weigh visual cues heavily, even when there was still room for improvement in more challenging listening situations, whereas children with low auditory performance heavily weighed on visual cues at all tested SNR levels. Children with DD, however, only weighed on their auditory performance in favorable conditions. Once the noise increased, the AV speech benefit of this group was not dependent on the auditory performance anymore. Connecting these findings to reading and phonological impairments, as often discussed in individuals with DD, we found that phonological processing as an auditory skill was a better predictor of speech perception performance than reading performance. This work suggests no inherent AV speech perception deficit in children with DD, however as these children often struggle with auditory processing skills like phonological awareness it is informative for caregivers and teachers to be aware of the differences in reliance on the auditory and visual speech signal to maximize AV speech perception benefits in noisy situations like classrooms or playgrounds.

4.4. Methods

The methods described below to measure AV speech perception via this remote psychophysical assessment were validated in Gijbels and Lee (2023) and show notable similarities to the established and validated computerized speechreading tasks for children like 'Test of Child Speechreading (ToCS) (Kyle et al., 2009).

4.4.1. Materials and Methods

The methods described below to measure audiovisual speech perception via this remote psychophysical assessment were validated in Gijbels and Lee (2023).

4.4.1.1. Participants

A group of 135 English-speaking children between 6.5 and 15 years old participated in this remote online study. Two criteria were used to classify children with DD. Parents had to report an official diagnosis of DD and they had to score more than one standard deviation below the mean (normed score <86, mean = 100) on at least one standardized reading task (Accuracy: Woodcock-Johnson - Basic Reading Skills (WJ-BRS; Schrank & Wendling, 2018) or Speed: Test of Word Reading Efficiency – Second edition (TOWRE-2: Torgesen et al., 2012)). Sixty-three children qualified to be in the DD subgroup (Age: 6.58 – 14.72, μ = 10.98; F = 26, M = 37) with a mean WJ-BRS accuracy score of 78.86 and a mean TOWRE-2 speed reading score of 74.17. The control group consisted of 72 children, age-matched (Age: 6.58 – 14.92, μ = 10.87; F = 38, M = 34) with the DD group. Their mean WJ-BRS accuracy score was 113 and their mean TOWRE-2 speed reading score 104.5. In both groups parents reported no other developmental disorders (e.g., Autism Spectrum Disorder or Attention Deficit/Hyperactivity Disorder), typical speech and language, and no perceptual difficulties like hearing or vision problems via a parental questionnaire. Participants were recruited from a UW database University of Washington Reading & Dyslexia Research Program (http://ReadingAndDyslexia.com) or the UW Communication Participant Pool. Parents of all participants provided written and/or video-recorded informed consent under a protocol that was approved by the University of Washington's Institutional Review Board. All children had English as their first language.

Supplementary Table A can be found in the published manuscript (https://doi.org/10.1111/desc.13431) and provides an overview of the distribution of

race/ethnicity and Hispanic origin in the current study compared to the census data of the US and Washington State from July 2020.

4.4.1.2. Stimuli and Materials

All children completed four tasks remotely. Two reading tasks (WJ-BRS and TOWRE-2) and one phonological awareness task (Comprehension Test of Phonological Processing, Second Edition; CTOPP-2; Wagner et al., 1999) were assessed via a video-conferencing call. All original materials were presented in accordance with the test manual but were delivered as a PowerPoint presentation over the videoconferencing platform. The last task, the AV speech perception task, was assessed remotely via a web-browser. The set-up of this AV speech-in-noise recognition task was based on the task used in Gijbels et al. (2021) and both the assessment of standardized behavioral tasks via a videoconferencing call as the web-browser based AV speech perception task were validated in Gijbels & Lee (2023).

The AV speech perception task was designed in the free online study builder, Lab.js (Henninger et al., 2021), and hosted via an online hosting platform created by Psychopy to independently run online studies. The task consists of 150 trials, with 20 different target stimuli. All words (target and foils) used for this experiment were consonant-vowel-consonant (CVC) words. The target words originated from the CHILDES database (MacWhinney & Snow, 1985) and should be acquired at ages 3 to 5 years old (Holt et al., 2011). These stimuli were recorded by a female native English speaker (Holt et al., 2011) and previously used by Lalonde and Holt (2016) and Gijbels et al. (2021). The stimuli were presented in one of three modalities (auditory, visual, or AV). In the AV modality a target word was presented in noise with a synchronous matching video of a woman speaking the target word. During the visual-only trials the video was shown with no auditory signal other than noise. The stimuli in the auditory modality consisted of the target word in noise accompanied

by a picture of the woman's face, in a neutral position. This still image was a neutral frame selected from one of the target video recordings. How the stimuli and the weighted long-term average speech spectrum noise were generated and applied is described in Gijbels et al. (2021). However, for this study three signal-to-noise ratio (SNR) levels were selected: -5 dB, -8 dB, or -11 dB. The selection of these levels allowed us to compare over multiple SNRs and to give children (between 6.5-15 years old) success experiences, as an SNR of -5 dB might be easy for the oldest, an SNR of -11 dB might be difficult for the youngest children (Massaro et al., 1986; Elliott, 1979). Picture selection and response layout in this one-interval four-alternative forced choice (1I-4AFC) paradigm were similar to Gijbels et al. (2021). The foil words in our 1I-4AFC task were picked to allow for an error analysis. Specifically, besides the target word picture (e.g., "run"), we presented an image that was a minimal pair of the goal word (e.g., "sun"), another one with the vowel in common (e.g., "gum"), and one completely random word (e.g., "pink"). Within the target-minimal pair relationship we also scored visual similarity, ranging from 1 (most similar) to 4 (least similar), based on confusion matrices of previous research (Owens & Blazek, 1985).

4.4.1.3. Procedures

All three standardized tasks (WJ-BRS, TOWRE-2, CTOPP-2) were administered by a trained speech-language-pathologist, as similar as possible to in-person testing. The participant (with caregiver close by) sat at a table in front of a computer (no tablet or phone) with audio and video turned on. Assessment of these tasks would be completed in one 30–45-minute session.

For the AV task, caregivers were directed to load the task in their web browser. Children were asked to be seated at a table, with their caregiver nearby in a quiet environment, facing the computer screen. Stimuli were introduced through the computer's speakers and answers were recorded via mouse clicks. At the beginning

of the task, participants and their caregivers were asked to set the speakers to a comfortable loudness level, based on a 7000 ms English speech fragment. All participants were trained with all 30 clipart pictures (target words and foils) and the participants were familiarized with the AV speech task by exposing them to 8 trials with feedback. The full AV speech perception task could be completed in 20-25 minutes. Both the training and the task itself were narrated by a cartoon character. This character would give the instructions verbally, while they were also presented in writing. All 150 trials were divided into 8 blocks and participants were offered a break after every block. Neither the trials, nor the breaks were limited in time. For every participant all types of trials were randomized before the start of the task. The task consisted of 60 AV trials (20 for each SNR level), 60 auditory trials (20 for each SNR level), 20 visual trials. The remaining 10 trials were catch-trials, in one of three modalities (auditory, visual or AV). During these catch trials, an image of the cartoon character would pop on top of the video/picture and serve as a control for visual attention to the screen at all times. Every trial started with a fixation target (500 ms) followed by a 2-second-long noise with the word stimulus appearing 500 ms after the beginning of the noise. Then, a screen with four pictures (arranged in a 2x2 box) appeared. The participant selected a response by clicking one of four (positionally randomized) images. No feedback was provided throughout the task.

4.4.2. Analysis methods

Statistical analyses were performed using the lme4, lmerTest, stats, and psych packages in R and RStudio (version 1.3.1093).

We used a linear mixed effects model; *Model 4.1:* $Performance_{(\%correct)} \sim Group_{(control,DD)} * Modality_{(A,AV,V)} * Age_{(years)} + (1|participant)$, to test whether we created a task sufficiently low in cognitive and linguistic demands so that there would not be a difference in auditory or visual performance between the control and DD group. The dependent variable was task performance, in all three modalities for all SNR levels. We

examined the fixed effects of modality, age, and group. Age was included as an interaction term since we expected older children to score higher in all modalities (Wightman et al., 2006; Leibold & Buss, 2019; McMurray et al., 2018; McCreery & Stelmachowicz, 2011; Ross et al., 2011). Group was included as an interaction term as it was our main focus of interest, and potentially expected different performance in different modalities by group (Hayes et al., 2003; Gelder & Vroomen, 1998; Ramirez & Mann, 2005; Boets et al., 2011; Bradlow et al., 2003; Cunningham et al., 2001; King et al., 2002; Ziegler et al., 2009).

For model 4.1b $(Performance_{(\%correct)} \sim Group_{(control,,DD)} * Modality_{(A,AV)} * Age_{(years)} * SNR_{(-5,-8,-11\ dB)} + (1|participant))$, we only used the auditory and AV modality and included the different SNRs (i.e., −5 dB, −8 dB, and -11 dB) as a main effect, since we expected performance per modality to be better for higher SNRs (Ross et al., 2011; Sumby & Pollack, 1954; Erber 1975; Grant et al., 1998). We did include SNR as an interaction term, since differences in auditory performance between children with or without DD might be specific for scenarios where the SNR is low (Rüsseler et al., 2015). The auditory modality, -8dB SNR and control group served as references in each category.

Model 4.2 $(Normalized\ AV\ Gain_{([AV-A]/[1-A])} \sim Group_{(control,DD)} * Age_{(years)})$ was created to inspect whether the AV speech perception benefit was different by group and/or age. The control group served as a reference group. We did not use the simple difference score between AV and auditory modalities to calculate the AV speech perception benefit. This would create a bias that high auditory scores necessarily lead to low benefit scores and therefore will be avoided by using a normalized AV gain [(AV−A)/(1−A)] (Alsius et al., 2016; Grant & Seitz 1998; and Sumby & Pollack 1954). We calculated the Bayes factor of normalized AV gain by group and age. This tested whether our data supports the null hypothesis, or whether it is too weak to yield evidence against it. A Bayes Factor of < .33 yields evidence for the null hypothesis

(H_0) being favored over the alternative hypothesis (H_1), whereas .33 - 3 shows data that is too weak to show favor for H_0 or H_1 (Biel & Friedrich, 2018).

Model 4.3 *(Normalized AV Gain $_{([AV-A]/[1-A])}$ per SNR ~ Group$_{(control,,DD)}$ * SNR $_{(-5,-8,-11\ dB)}$*Auditory performance$_{(scaled_scores)}$ * Visual Performance$_{(scaled_scores)}$ + (1|participant))*, is a linear mixed effects model to predict normalized AV gain per SNR, based on individual auditory performance, SNR, Group and visual performance. The control group and -8dB SNR served as references in each category. The model included a random intercept for participants, which estimated a variance component for the fixed factors such that the model fits an intercept for each participant.

We used Pearson correlation coefficients to look at speech perception performance in relationship to reading and phonological processing skills, and we treated our participants as one continuous group. By using the standardized reading scores of the WJ-BRS test and the standardized phonological processing scores of the CTOPP-2 we can look at continuous performance rather than breaking up our participants in two groups.

Chapter 5

Characterizing the Role of Prelinguistic and Linguistic Information in Integration of AV Speech in Adults

5.1. Introduction

Audiovisual speech perception is complex and pivotal in comprehending speech. Research consistently shows that incorporating visual information from the talker's face and mouth enhances auditory speech perception, especially in challenging listening environments (e.g., Binnie et al., 1974; Erber, 1969; MacLeod & Summerfield, 1987; Sumby & Pollack, 1954). This improvement is evident in accuracy, processing speed, comprehension, and reduced listening effort (e.g., Tye-Murray et al., 2007a; 2008, Reisberg et al., 1987, Sommers & Phelps, 2016). Audiovisual speech integration is characteristic to an individual based on (1) changes in response to experience with auditory and visual sensory processing across the lifespan (Baum & Stevenson, 2017; Lalonde & Werner, 2021; Tye-Murray et al., 2016), and (2) individual variability in AV speech integration within a phase of life (Dey & Sommers, 2015; Grant et al., 1998; Nath et al., 2011). Yet, the precise mechanisms behind the large individual differences of AV speech perception across development, in developmental disorders, or even in adults have not been fully understood.

Chapter 3 and 4 therefore focused on how task- and person-dependent factors influence AV speech perception and AV speech perception benefits throughout development. In Chapter 3 we established that although speech perception (in all modalities) improved with increasing age (Fig 3.8), the size of the AV speech perception benefit was better explained by auditory speech-in-noise performance

than age (Fig 3.9). However, Chapter 4 indicated that auditory speech-in-noise perception as a predictor for AV speech perception benefits in noisy backgrounds only remained valid for children with no phonological difficulties (Fig 4.3). As a result, this triggered our interest to further examine how, and how much, linguistic information influences AV speech perception performance.

Furthermore, we reported in Chapter 3 that there are fundamentally different mechanisms of development involved in AV speech processing when congruent and incongruent stimuli are used (Fig 3.10). This motivated us to further delineate individual differences in solving the causal inference problem (Körding et al., 2007). This causal inference problem, viz., the process to decide if and how to integrate information from the auditory and visual stream, is based on coherence of both information streams (Cao et al., 2019), and synchronous temporal information might be one of the most important cues in this process (Bizley et al., 2016; Lee et al., 2019). Therefore, we studied the intricate relationship between prelinguistic and linguistic information of the AV integration process by manipulating temporal synchrony and linguistic complexity.

5.1.1. Audiovisual Temporal integration

Temporal cues play an informative role in AV speech identification (Munhall & Vatikiotis-Bateson, 2004; Zion Golumbic et al., 2013). Furthermore, the synchronization of AV signals enhances the likelihood of object formation, as evidenced by studies examining single neurons (e.g., Meredith et al., 1987), neural populations (e.g., Senkowski et al., 2007; Macaluso et al., 2004), behavioral experiments (Hillock et al., 2011; van Wassenhove et al., 2007; Wallace et al., 2004), and observations among neurodiverse groups (such as dyslexia; de Boer-Schellekens et al., 2013).

Temporal information is strongly correlated in the auditory and visual modality, yet, especially for conversational AV speech stimuli, there is often no absolute temporal

coherence. Visual information precedes auditory information in conversational contexts as the timing of the mouth movements precedes the onset of the voice by tens to a few hundred milliseconds (Chandrasekaran et al., 2009; van Wassenhove et al., 2005), and visual information reaches the listener sooner than auditory information due to differences in the speed of acoustical versus electromagnetic waves. Fortunately, absolute synchrony between auditory and visual signals is not necessary. Rather, the perception of synchrony is crucial for integrating AV speech. Our perceptual system can accommodate a range of signal asynchronies (viz., up to a few hundred milliseconds) and perceive them as simultaneous (Dixon & Spitz, 1980). Within this window of simultaneous perception, the likelihood of combining AV information into an object increases (Stevenson et al., 2012). Consequently, the Temporal Binding Window (TBW) is a theoretical construct that has been used as a measure to describe a range of Stimulus Onset Asynchronies (SOAs) between the auditory and visual signals, that are perceived as simultaneous. This TBW is defined as "*the epoch of time within which stimuli from different modalities is likely to be integrated and perceptually bound"* (Wallace & Stevenson, 2014), and suggests that AV speech perception involves rather loose temporal associations (Munhall & Vatikiotis-Bateson, 2004).

When assessing the TBW, participants are exposed to stimuli in two different modalities that ideally share a natural connection (e.g., video of a hammer hitting with the sound of a hammer, or a speech signal with a matching video of the speaker). These two modalities are presented synchronously (SOA = 0ms), or asynchronously (often up to SOA = 500ms), and participants must judge the synchrony of these presented stimuli. Four common procedures are used to measure the TBW, which vary in terms of the judgment type (simultaneity, temporal order, or perceptual fusion) and type of stimulus presentation (1-interval versus 2-interval-2-alternative forced choice). The asynchrony is either audio-leading or visual-leading, resulting in 1 sigmoid or 2 sigmoid functions, depending on whether

or not your stimuli were either or both audio- / video-leading (see Stevenson & Wallace, 2013 for details). From this function two key values are retrieved: the SOA corresponding to the highest simultaneity perception (or Point of Subjective Simultaneity; PSS) and a window of simultaneity perception known as the temporal binding window (TBW).

However, insights from the PSS may not be informative, as they can vary depending on stimulus presentation (Vatakis et al., 2008; Vroomen & Stekelenburg, 2011) and task demands (Van Eijk et al., 2008; Vroomen & Stekelenburg, 2011; Shore et al., 2001). Furthermore, the reliability of the PSS within an individual remains unclear (Recio et al., 2019 vs. Stone et al., 2001).

The TBW is modeled by two sigmoid functions (centering around the PSS), as visual- or auditory-leading SOAs are presented. For AV stimuli, and especially for AV speech stimuli, the two functions are often asymmetrical (Conrey & Pisoni, 2006; Grant et al., 2004; Kawase et al., 2016). The perceptual system is less tolerant for AV asynchrony when the auditory signal is leading, and it may be fine-tuned to the natural statistics of AV speech, as visual information reaches the perceptual system before acoustic speech (Maier et al., 2011). Cecere and colleagues (2017) argue that there are consistently different neural activation maps for auditory-leading or visual-leading asynchronies, indicating that the leading sense activates at least partially separate networks in the brain. Auditory-leading AV maps tend to be more right-lateralized (i.e., suggested to direct visual attention), whereas visual-leading maps are more symmetrical organized in the brain (i.e., suggested to enhance speech perception).

The TBW measure has been successful in showing group level differences, across development (Stevenson et al., 2018), and neurodiverse populations (Wallace & Stevenson, 2014). More specifically, the width of the TBW, often defined as the width between the audio-leading and visual-leading 50% or 70% point of synchrony

perception, or from 0 ms SOA or the PSS to the 50% or 70% point (when only one half is measured), has been identified as crucial in this process. Furthermore, significant intersubject variability has been reported (Conrey and Pisoni, 2006). This large intersubject variability, ideally with little intrasubject variability, in combination with consistent developmental effects (See Stevenson et al., 2018 for review), positions the TBW as an intriguing measure for understanding AV integration mechanisms. However, there is very limited understanding regarding the within-subject stability of the TBW. To the best of our knowledge there is currently only one report of test-retest reliability of the TBW at the individual level (Flanagan, Zumbrunn et al., 2023). Their results are very promising as they show a very strong correlation (R = 0.92) between the two test points tested across different platforms — online and in the laboratory. However, they used an AV illusion to measure the TBW, and the decisions made regarding temporal synchrony perception might be different for congruent and incongruent stimuli, like AV illusions (van Wassenhove et al., 2007).

Nonetheless, the concept that perceived temporal synchrony plays a crucial and distinct role in the object formation of AV speech integration gains credibility, given that speech intelligibility remains largely unaffected within the TBW (Grant & Greenberg, 2001; Grant et al., 2004; Gordon-Salant et al., 2017; Munhall et al., 1996; Pandey et al., 1986). Beyond the TBW, speech intelligibility decreases. The fact that AV speech performance still surpasses auditory-only performance, even for significant asynchronies in the order of 400 ms, might represent late-stage integration mechanisms (Grant & Greenberg, 2001). Furthermore, temporally synchronous AV stimuli might lead to sensory enhancement of perceived loudness (Odgaard et al., 2004). Odgaard and colleagues (2004) showed a very robust effect of enhancement in loudness judgments when a white noise signal was accompanied by a concurrently presented light. In contrast to noise-induced enhancement in ratings of brightness (a more widely studied topic: Stein et al., 1996; Odgaard et al., 2003),

this loudness effect was robust to different task manipulations. Therefore Odgaard and colleagues (2004) suggested that light-induced enhancement of loudness may reflect an early-stage sensory interaction. Indeed, as loudness perception represents an orthogonal feature that creates object formation for the AV signal, this AV loudness enhancement could represent an effect of AV binding (Bizley et al., 2016). This effect might especially be apparent for temporally congruent stimuli. Although Gillmeister and Eimer (2007) studied auditory and somatosensory perception, they varied the SOAs from -200 ms to 200 ms, and showed that the loudness enhancement effect in multisensory stimuli was significantly larger for synchronous than asynchronous stimuli. Thus the detection benefit of auditory signals that synchronous AV signals provide over asynchronous AV signals might go hand in hand with enhanced loudness perception (Lovelace et al., 2003).

At the individual level, there is a suggested negative correlation between temporal asynchrony perception (i.e., the TBW width) and auditory and AV speech perception performance (Baskent & Bazo, 2011; Conrey & Pisoni, 2006), AV speech perception benefits (Grant & Seitz, 1998), and AV illusions (Stevenson et al., 2012). This implies that individuals who depend more on AV cues (experiencing greater AV benefits and more AV illusions) are more adept at detecting AV asynchrony.

Understanding these effects of temporal synchrony perception on both a group and individual level sheds light on the dynamics of AV integration mechanisms. This points to the importance, especially for AV speech signals, in understanding how synchrony perception relates to key contributors of AV speech, like linguistic information.

5.1.2. Linguistic information in AV integration

The processing of linguistic information, influenced by the complexity of the stimulus and the linguistic knowledge of the listener, is likely a late-stage integration process, distinct from the early stages of object formation in AV stimuli (Bizley et al.,

2016; Fiscella et al., 2022). Linguistic information thus contributes to assigning weights to each modality and providing contextual information (as priors) (Ma et al., 2009).

We define linguistic information as all aspects related to 'language.' At its core, both auditory and visual speech signals carry speech-specific cues, such as phoneme-viseme relationships. Phonemes represent the smallest distinctive auditory units of speech, while visemes represent the smallest distinctive visual units of speech (Fisher, 1968). Visemes portray the facial and oral positions during phoneme production. Visemes can, however, represent multiple phonemes due to similar lip appearances (Bear & Harvey, 2017). Nonetheless, each language possesses a unique set of visemes corresponding to its specific phonemes, making phoneme-viseme connections distinctive to the language. Phoneme-viseme connections enhance speech perception linguistically by aiding in talker identification and improving speech quality (Tye-Murray et al., 2007a; Eskelund et al., 2011), as well as the speed of speech comprehension (Reisberg et al., 1987).

Higher order layers of linguistic information contributing to AV speech integration involve (1) the lexical content within the speech signal and (2) the complexity of the speech stream (ranging from individual words to sentences and extended passages). The extent to which this linguistic information influences speech perception depends on various factors, including the individual's lexicon, semantic and syntactic knowledge, utilization of contextual cues at both word and sentence levels to compensate for misunderstandings or limited auditory and visual input, as well as memory processes and strategies for accessing lexical information based on incomplete signal data (Grant et al., 1998; Ma et al., 2009).

Temporal information is expected to mainly play a role at the signal detection level (Lalonde & Werner, 2019; 2021), and linguistic information contributes more to signal recognition (Lalonde & Werner, 2019). While there is a large body of literature

representing either AV temporal synchrony perception or AV speech intelligibility, the correlation between prelinguistic cues important for signal detection (like temporal synchrony perception) and linguistic cues important for intelligibility is relatively understudied.

5.1.3. Temporal and linguistic information in AV integration

Grant and Seitz (1998) explored the relationship between AV speech perception benefits and temporal asynchrony perception, examining both sentences and syllables. They observed that individuals experiencing greater negative effects of AV asynchrony on AV speech intelligibility demonstrated significantly larger AV benefits for synchronous stimuli at the sentence level (R = -0.45), but not at the syllable level (R = -0.05). This discrepancy likely stems from increasing interruption of word for temporally asynchronous AV sentences, but not for syllables. Additionally, they noted a larger overall AV speech perception benefit for syllables compared to sentences, with sentences exhibiting greater intersubject variability. These findings contrast with Fort and colleagues' (2010) observation that visual cues in AV speech aid in accessing lexical information, particularly for low-frequency words where lexical access is challenging. Therefore, it would be expected to show larger AV benefits for sentences in Grant and Seitz's study, as listeners could leverage both morpho-syntactic structure, lexical information, and semantic context within sentences, but not syllables. The increased intersubject variability in AV sentence performance, however, may be attributed to individual variations in linguistic proficiency.

Grant and Seitz (1998) did not report how temporal synchrony perception differed by linguistic complexity. Vatakis and Spence (2006) reported that listeners are better at temporally discriminating AV stimuli of lower complexity (i.e., syllables vs. sentences), independent of the stimulus duration. They argued that increased stimulus complexity promotes improved synchrony perception (i.e., for wider SOAs)

as it would become more likely for these stimuli to come from a single AV event (Bertelson & de Gelder, 2004). However, Lee and Noppeney (2014) reported significant effects of stimulus duration, noting that longer stimuli yielded narrower TBWs (for both music and speech stimuli). Listeners would be able to accumulate more information over time for long stimuli and therefore obtain more precise temporal estimates. These seemingly contrasting findings could be reconciled in that although for Vatakis and Spence (2006) stimulus complexity also correlated with stimulus duration, the stimuli (shorter vs slightly longer sentences) of Lee and Noppeney (2014) did not necessarily vary in linguistic information that could be retrieved from the stimuli. As a result, adding temporal information from syllable to sentence level could improve temporal accuracy in a judgment task, but having linguistic context (lexical, syntactic and semantic), could elongate the TBW more, as more complex stimuli are more likely to come from the same event.

To further elucidate these seemingly contradictory AV speech processing findings, it would be beneficial to understand the implications when both semantic and temporal information are manipulated to the extent that they both match or mismatch. Ten Oever and colleagues (2013) and Vatakis and Spence (2007) varied semantic reliability by providing a match/mismatch in voice (Vatakis & Spence, 2007) or phoneme-viseme connections (ten Oever et al., 2013). Their findings, concerning syllables (ten Oever et al., 2013; Vatakis & Spence, 2007) and words in sentences (Vatakis & Spence, 2007), revealed that semantically congruent stimuli led to wider TBWs, suggesting integration at more temporally disparate offsets.

To better understand these findings it is helpful to touch on how AV integration is represented in the brain. AV integration is most likely a multistage process (Peelle & Sommers, 2015), where early integration mechanisms in the primary cortices (auditory and visual) implement AV speech to increase sensitivity to acoustic information (i.e., the amplitude envelope, influencing attention and perceptual sensitivity; e.g., Calvert et al., 1999) and late-stage integration mechanisms, in more

higher order cortices like the Superior Temporal Sulcus (e.g., Möttönen et al., 2004), and between temporal and frontal regions (Giordano et al., 2017), incorporate content information (i.e., phoneme-viseme connections, constraining lexical selection). To explain the findings of ten Oever and colleagues (2013), Stevenson and colleagues (2014) suggested that the idea of multistage integration might not be sufficient. They argued that the interaction between timing and linguistic congruency, as illustrated by ten Oever and colleagues (2013), really points to parallel accumulation of evidence. The temporal relationship of AV speech signals provides important evidence of the likelihood that the multisensory information originated from the same talker and should be integrated. In addition, linguistic information (e.g., phoneme-viseme congruency) also provides evidence as to whether or not integration should take place. And both of these decision processes result in a single decision criterion. Thus, when stimuli are semantically congruent (e.g., same talker, phoneme-viseme match) less temporal alignment is needed to cross a decision boundary that would result in integration of the AV signal (i.e., wider TBW). Cappelloni and colleagues (2023) similarly reported that incongruent semantic information, like talker identity, influenced the weight assigned to temporal synchrony. However, they used it in the context of competing stimuli and concluded that if talker identity already helps in the decision not to integrate information streams, temporal coherence across auditory and visual streams does not offer any additional benefit. This questions the theoretical basis of using TBW tasks for incongruent stimuli. Specifically, if other more prominent features lead the integration decision, rather than temporal synchrony perception as measured with the TBW, the TBW task might capture a decision based solely on synchrony judgment as opposed to a general measure of AV integration.

More generally, this opens up the idea of multistage integration. If very prominent characteristics of the AV signal indicate not to integrate, then no assumption of a common source will be made and the need for an extensive causal inference and

bayesian weighting process is limited. However, when there is some indication that AV stimuli originate from the same source (e.g., semantic congruency but temporal asynchrony), multistage integration, as discussed through the causal inference model, (Körding et al., 2007; Ma et al., 2009; Peelle & Sommers, 2015) might take place, and result in integration of AV speech. Furthermore, if any AV integration is reported even when there is clear semantic stimulus incongruency, the task might force the listener to match auditory and visual information based on prior knowledge.

But, could the incongruency of phoneme-viseme connections, as used by ten Oever et al. (2013) or by van Wassenhove et al., (2007; with the McGurk illusion) – leading to more narrow TBWs – be sufficient to make the AV integration decision and thus leading to temporal coherence not providing any additional benefit? The overall lower report of temporal synchrony perception (regardless of the width of the TBW) for incongruent phoneme-viseme information (van Wassenhove et al., 2007), in combination with the significant effect of priors (Alsius et al., 2018; Ma & Schnupp, 2023), large individual variability, and the importance of a closed-set task (Alsius et al., 2018) to even report any integration of incongruent phoneme-viseme information, suggests that integration of these stimuli might be dictated by the task.

For congruent speech, we need to identify the linguistic contribution of different types of speech stimuli in AV speech perception, by studying the phoneme-viseme connections in different contexts. Higher order linguistic cues – lexical information, word frequency, word density, and morphosyntactic structure – play a key role in understanding how listeners process isolated words, sentences and paragraphs. At the more basic levels (i.e., pseudowords (or nonsense words), syllables, or even consonants) listeners only have low-level linguistic cues like phoneme-viseme connections to consider. If we want to measure the specific effect of temporal information on phoneme-viseme connections, these language specific connections might need to be broken (Fisher, 1968). Using an unfamiliar language could provide

this disconnect, however it might be hard to construct an experiment practically since languages like English and Spanish, for example, have 14 overlapping phonemes for the consonants (Whitley, 2002). An alternative is the use of AV congruent stimuli like reversed speech (i.e., backwards played speech signal). Audiovisual reversed speech has temporally aligned onset and offset auditory and visual information and identical modulation of the speech envelope as forward played speech (Kaufeld et al., 2020), however "t*emporal reversal of natural speech distorts temporally based segmental linguistic attributes, affecting consonants, diphthongs, formant transitions, syllable shape, and the relative duration of segments at word initial or word final position. It is impossible to perceive the lexical content of an utterance played backward. The speech seems to be composed of an unknown foreign language, yet the natural vocal timbre is preserved* (Sheffert et al., 2002)." Thus, it breaks phoneme-viseme connections as viseme recognition relies on the sequential arrangement of facial and mouth movements associated with speech sounds (Michon et al., 2020), and eliminates any linguistic information (Sheffert et al., 2002).

At a syllable level, Vatakis and Spence (2006) reported no significant difference in temporal synchrony perception between normal and reversed AV speech. Thus, the absence of lexical information for temporally congruent stimuli did not narrow the TBW. Interestingly, at a sentence level, Maier and colleagues (2011) showed that the proportion of synchronous responses for reversed AV speech never reached a PSS close to 100%, unlike normal sentences. Whether this is caused by the subtle changes in the temporally based segmental linguistic attributes (Shefftert et al., 2002) or the unfamiliarity of reversed AV speech stimuli remains unclear.

Therefore, we believe that there is a need to better understand the influences of different aspects of AV speech processing, both at the prelinguistic and linguistic levels, as well as the interaction between these two factors, to further understand how AV speech is processed at each stage of processing.

5.1.4. Research questions

By manipulating both prelinguistic perceptual information (temporal synchrony, temporal speech envelope) and linguistic information (phoneme-viseme connections, lexical, and morpho-syntactic context), we executed four related experiments using within-subject design to answer four questions:

1. What is the role of linguistic information on AV temporal synchrony perception? (Section 5.2)

2. Is temporal synchrony perception reliable across sessions and linguistic levels, within each individual for AV speech stimuli? (Section 5.3)

3. Can temporal synchrony and linguistic complexity influence other fundamental features of AV perception (e.g., loudness of the AV signal)? (Section 5.4)

4. Do linguistic complexity and temporal synchrony collectively influence AV speech intelligibility? (Section 5.5)

Based on the literature, as described in Section 5.1.3, we hypothesized that the TBW widens with increasing linguistic complexity (Vatakis & Spence, 2006): increased stimulus complexity would promote increased synchrony perception (i.e., for wider SOAs) as it would become more likely for these stimuli to come from a single AV event (Bertelson & de Gelder, 2004). If added linguistic information triggers an increase in the TBW, we would see wider TBWs for words than pseudowords (i.e., nonsense words that have the same amount of phoneme-viseme information available and the same duration, but not the same lexical information). If added exposure leads to accumulation of information in the integration decision (Lee & Noppeney, 2014) we would see no differences between words and pseudowords, but narrower TBWs (i.e., more precise temporal estimates) would be observed for meaningless sentences (i.e., random combination of words) than for words. The more interesting question is how this seemingly contradicting information is

integrated when both linguistic complexity and exposure change in a step-by-step manner.

To our second question, we hypothesized that the TBW is a stable measure and therefore has good test-retest reliability. To the best of our knowledge, there are currently no peer reviewed reports of the stability of the TBW on a group level, or at the individual level. This is needed to answer whether the TBW is characteristic for an individual (considerable intersubject variability, but low intrasubject variability) and whether this measure is different for different levels of linguistic complexity. We hypothesized that the TBW width is stable on a group level and shows as a characteristic specific to an individual, based on the large effects of the TBW in neurodiverse populations (Wallace and Stevenson, 2014).

Third, based on the findings of Odgaard et al. (2004) and Gillmeister and Eimer (2007), we predicted that synchronous AV speech will result in significantly larger loudness perception than asynchronous AV speech. As there is so little literature about this effect, we had no hypothesis of the size of this effect. However, as it is indicated as an early-stage integration mechanism (Odgaard et al., 2004), we did hypothesize a correlation to the TBW width on an individual level. Furthermore, as this is an early-stage integration effect, there should be no influence of linguistic complexity (Fiscella et al., 2022), therefore we expect similar loudness changes across linguistic levels.

Finally, we hypothesized that AV speech intelligibility would improve with increasing linguistic complexity for synchronous stimuli. For synchronous stimuli and asynchronous stimuli up to 150 - 200 ms (Grant et al., 2004) we expect AV speech intelligibility to be rather stable, but decline outside of this asynchrony range (which is similar to the reported TBW width; Grant et al., 2004). We expect it to still be enhanced compared to auditory-only performance outside of the TBW (Buchwald et al., 2009; Campbell & Dodd, 1980; Grant & Greenberg, 2001). However, once outside

of the TBW, the duration of the stimulus (i.e., sentences vs. words) might create additional interference. For stimuli on a word level one could hold on to both the auditory and the visual representation and later integrate information from these two. For sentence stimuli, there will be a point where the audio and video overlap, but are mismatched, it is then up to our cognitive skills like attention and memory to define how much information can be retained and integrated. This interference is expected to be larger for anomalous sentences compared to meaningful sentences as it is harder to retain unpredictable information, and it can reach a point where our AV performance is worse than auditory-only performance.

On an individual level we did not necessarily expect to find strong correlations between the TBW task or the loudness rating task and the speech intelligibility score or the AV speech perception benefit as the TBW and the loudness effect are supposed to represent early-stage integration, whereas AV speech intelligibility and AV speech enhancement can also be influenced by late-stage integration. However, there are some indications that the TBW is related to AV speech perception measures (Grant & Seitz, 1998; Stevenson et al., 2012).

5.1.5. Methods - General paradigm and experimental platform

To answer these four research questions, four different experiments were carried out. All of these experiments were implemented remotely and to participate in experiment 2,3, or 4, the participant must have completed experiment 1.

The specific methods will be described per research question. However, as there were some general commonalities between all experiments, we start with a general method section below.

5.1.5.1. Participants

The selection criteria to participate in these experiments were (1) being a younger adult (age 18 - 40-years-old), (2) living in the U.S. (for the purpose of participant

payment processing), (3) being a self-identified native English speaker, (4) having self-reported normal hearing thresholds, and (5) self-reported normal or corrected to normal vision.

Participants were scanned for these criteria based on self-report questionnaires. To further verify self-reported normal hearing thresholds, a short sentence recognition task in quiet had to be completed at the start of the first experiment. The participants had to obtain a perfect score on this sentence recognition task in quiet to be able to qualify as a participant. This task consisted of 6 trials (randomly selected from 10 sentences) recorded by a female English speaker (Holt et al., 2011). For each trial the participant was presented with an auditory stimulus and had to type the sentence in a text box. No feedback was provided.

Questions about age, race, and Hispanic origin were formulated similar to questions from the U.S. Census 2020.

In total, 46 participants qualified for these experiments, with a mean age of 29.8 years-old (min = 21, max = 40, Q2 = 29.5, sd = 6.07). More demographic details are reported in Table 5.1.

5.1.5.2. Recruitment and consent

Participants were recruited at the University of Washington or from other oral, or online advertising. A flyer was distributed to undergraduate and graduate students in the Speech and Hearing and Engineering departments at the University of Washington, and advertising on a social media platform was also used to recruit from a wider age and demographic range. A website was developed to provide potential participants with more information, to give them an opportunity to ask questions and to lay out a step-by-step plan to run the experiment remotely.

All our participants provided remote informed consent for each task, under a protocol that was approved by the University of Washington's Institutional Review Board. The task did not progress if consent was not given.

All participants were financially compensated for each task at a rate of $15/hrs in the form of an online gift card.

TABLE 5.1. Participant information.

N	Intensity Computer (%)	Timezone (Standard Time)	Age	Gender	Race	Hispanic, Latino, or Spanish origin?	Task 1: TBW	Task 2: Reliability	Task 3: Loudness	Task 4: Intelligibility
1	75%	Pacific	33	F	White	no	X	X	X	X
2	75%	Central	24	F	White	no	X	X	X	X
3	75%	Central	23	F	White	no	X	X	X	X
4	74%	Central	27	F	White	no	X	X	X	X
5	100%	Eastern	21	M	White	no	X	X	X	X
6	75%	Pacific	23	F	Chinese	no	X	X	X	X
7	100%	Pacific	23	M	Filipino	no	X	X	X	X
8	75%	Central	38	F	White	no	X	X	X	X
9	34%	Eastern	34	M	White	no	X	X	X	X
10	94%	Eastern	34	F	White	no	X	X	X	X
11	82%	Central	40	F	White	no	X	X	X	X
12	50%	Eastern	31	F	White	no	X	X	X	X
13	60%	Pacific	26	F	White	no	X	X	X	
14	75%	Pacific	32	F	White	yes (1)	X	X	X	
15	70%	Pacific	21	F	Black or African American	no	X	X	X	
16	75%	Pacific	31	M	White	no	X		X	X
17	75%	Pacific	34	F	White	yes (2)	X		X	X
18	60%	Pacific	33	F	Chinese	no	X		X	X
19	25%	Eastern	38	F	Korean	no	X		X	X
20	50%	Pacific	22	F	White	no	X		X	X
21	50%	Eastern	21	F	Chinese	no	X		X	X
22	76%	Pacific	21	F	Black or African American	no	X		X	X
23	75%	Pacific	25	F	Other Asian	no	X	X		
24	74%	Central	24	M	White	no	X	X		
25	67%	Pacific	34	F	White	no	X	X		
26	60%	Pacific	22	F	Chinese	no	X		X	
27	75%	Pacific	29	F	Chinese	no	X		X	
28	40%	Pacific	27	F	White	yes (1)	X		X	
29	28%	Central	26	F	White	no	X		X	

30	44%	Mountain	30	F	White	no	X		X
31	65%	Pacific	27	M	White	no	X		
32	50%	Pacific	29	F	White	yes (1)	X		
33	80%	Pacific	35	F	White	no	X		
34	100%	Eastern	34	M	White	no	X		
35	75%	Pacific	24	F	White	no	X		
36	50%	Pacific	34	F	White	no	X		
37	50%	Pacific	34	F	Chinese	no	X		
38	65%	Pacific	29	M	Black or African American	no	X		
39	80%	Pacific	38	F	White	no	X		
40	40%	Eastern	35	F	Other Asian	no	X		
41	70%	Eastern	40	F	White	no	X		
42	55%	Pacific	23	F	White	no	X		
43	80%	Eastern	38	F	White	no	X		
44	30%	Eastern	25	F	White	no	X		
45	75%	Central	37	F	White	yes (2)	X		
46	75%	Eastern	40	F	White	no	X		

Count (N), self-reported intensity setting of computer audio, geographical location (Timezone), age (min = 21, max = 40), Gender, Race, Hispanic origin, and task completion. yes(1) = Mexican, Mexican Am., Chicano, yes (2) = Other

5.1.5.3. Temporal asynchrony

For all experiments the SOA was manipulated between 0 and 500 ms. The video was presented simultaneously with or preceding to the audio. Only visual-leading asynchronies were introduced as (1) auditory-leading stimuli do not occur in naturalistic speech environments (Chandrasekaran et al., 2009), (2) auditory-leading stimuli are not expected to correlate with AV speech perception performance (Grant & Seitz, 1998; Stevenson et al., 2012) and (3) to keep data collection (per task) within a timeframe that is reasonable to complete within one session.

The selection of SOAs between 0 and 500 ms was based on earlier work with speech stimuli from Stevenson and colleagues (2012), showing that this range should encompass the complete synchrony-asynchrony window of perception for speech stimuli. We decided on a random order presentation of fixed-interval SOA's rather

than an adaptive procedure as Massaro and colleagues (1996) reported that the stimuli preceding the target stimuli in an adaptive staircase influence the decision of synchrony perception and therefore the TBW.

All SOAs were generated offline using ffmpeg software (*Python 3.7*) to ensure the intended temporal relationship between the auditory file and its corresponding silent video file.

5.1.5.4. Linguistic levels

The audio and video of all speech stimuli were recorded (i.e., the voice and the video match naturally). Therefore, there was no dubbing or artificial changes other than introducing temporal asynchrony by delaying the audio presentation.

Five different levels of linguistic complexity were established (Table 5.2) and used for these tasks. At the most basic level to establish audio-visual correspondence without lexical information, we chose to play words or pseudowords backwards, collectively referred to as reversed AV speech. Reversed speech maintains the general temporal envelope structure, with onset and offset, amplitude changes, and phase correspondence in both modalities. In contrast to all other stimuli used here, reversed speech breaks the American English language phoneme-viseme connections (Michon et al., 2020). With these reasons, it has been suggested that reversed speech signals might be experienced similarly to speech of an unfamiliar language (Sheffert et al., 2002; Vatakis & Spence, 2006). We argue that reversed speech has a more systematic way of breaking phoneme-viseme connections as we are more confident no phoneme-viseme connection can be recognized by the listener, as opposed to unfamiliar languages that often have some overlap in phoneme-viseme connections with the native language (Whitley, 2002). By using reversed words rather than reversed sentences we limited the impact of reversing the speech envelope structure beyond the phoneme-viseme connections.

We utilized pseudowords (e.g., kinit) as stimuli for the second linguistic level. These are nonsense words that have the same amount of phoneme-viseme information available and the same duration. However, unlike actual words, pseudowords lack lexical meaning. While adhering to the syllabic structures of the language, pseudowords do not convey any lexical or semantic content. As a result, factors such as word frequency and density are not applicable to pseudowords. Consequently, our third linguistic level consisted of English words.

TABLE 5.2. Different levels of linguistic complexity representing the speech stimulus types used in these experiments.

Stimulus Type	**Example**	**Phoneme-Viseme connection**	**Lexical knowledge**	**Morpho-syntactic knowledge**	**Predictable linguistic context**
Reversed words	*"KAVIET" or "NOIL"*				
Pseudowords	TEIVAK	X			
Words	LION	X	X		
Anomalous Sentences	The STINKY ROPES FLUSH the CHICKEN.	X	X	X	
Meaningful Sentences	The BOY GAVE the FOOTBALL a KICK.	X	X	X	X

Each level is presented with an example and the linguistic information they represent. All stimuli are semantically congruent speech stimuli.

The fourth and fifth linguistic levels comprised sentence structures. Transitioning from individual words to complete sentences introduces several elements: (1) extra temporal onset/offset information represented in the longer duration of a sentence, and added envelope information (Grant & Seitz, 1998), (2) a larger variety of coherent phoneme-viseme connections, (3) predictable linguistic information based on the morpho-syntactic structure in English and, (4) predictable context information. By incorporating both anomalous sentences (e.g., The stinky ropes flush the chicken) and meaningful sentences (e.g., The boy gave the football a kick), we

aim to make a distinction specifically based on context predictability within a sentence.

5.1.5.5. Stimuli

The target stimuli for this experiment were sourced stimuli that have been validated in previous work (Al-Salim et al., 2020; Kocins et al., 2022; Lalonde, 2019; Grieco-Calub et al., 2023; Holt et al., 2011; Shatzer et al., 2018; Stelmachowicz et al., 2000; Van Engen et al., 2014) and sentences from the SteVi Speech test corpus (STeVi Speech Test Video Corpus. (n.d.). Sensimetrics' Speech). For each linguistic level, three different female speakers were selected, with a total of 20 different stimuli per linguistic level (for Tasks 1-2). Fifteen different stimuli per linguistic level were used in task 3, and task 4 contained completely new speech items, from the same pool of recordings that have been previously unused in the other tasks.

The word, pseudoword, and reversed speech stimuli were selected to each have 1-, 2- and 3-syllable stimuli. We introduced word stimuli with different syllable lengths to limit the effect of absolute duration, envelope information and the variety in phoneme-viseme connections between word and sentence stimuli. The reversed speech was a selection of both word and pseudoword stimuli. Each sentence stimulus had a similar structure, starting with 'The' and containing four keywords.

5.1.5.6. Noise fragments

All stimuli were presented in noise, a four-talker-babble masker. This specific babble was used as a masker in Van Engen and colleagues (2014). Each masker fragment was a random snippet of 4000 ms, consisting of 4 female American-English talkers producing meaningful sentences (from Bradlow & Alexander, 2007). The 4-talker-babble was equalized for RMS amplitude with Audacity 2.4.2. The stimulus started playing randomly 500 to 1000 ms after the start of the babble masker, with a target-to-masker ratio (TMR) of -6 dB. All target and masker files were combined in advance and offline.

5.1.5.7. Processing of the speech signals

All materials were processed in a similar way:

1. iMovie was used to

    - Select the size of the video frame so all speakers were equally visible (head and shoulders filling 90% of height of the frame (1280 x 720 px)).
    - Cut the length of the videos to 2500 ms for both sentence and word stimuli.
    - Split the audio from the video fragment.
    - Place a black frame around the video, to make all videos more similar. The speaker shows in a cut-out circle. Before and after the 2.5s-fragment black frames were added to make all videos exactly 4000 ms long.

2. Audacity 2.4.2 was used to

    - Perform noise reduction. Although there was no significant noise on any of the original videos, we still completed this step for all videos.
    - Perform amplitude equalization (dB RMS) for both stimuli and masker.

3. Ffmpeg software (Python 3.7) was used to

    - Select random masker fragments of 4000 ms
    - Assemble target and masker with a TMR of -6 dB.
    - Combine audio and video (with the selected SOAs).

5.1.5.8. Target-masker-ratio

As integration of AV speech is most beneficial in moderate noisy environments (Ma et al., 2009) we ensured all of our tasks (i.e., temporal synchrony perception, loudness perception and speech intelligibility) were presented in background noise. To determine a moderate noise level that did not result in floor or ceiling performance for stimuli of different levels of linguistic complexity a pilot study was conducted.

Eight participants (F=6, Age 24-45, native English speakers) completed a brief (μ = 15 min) remote task to select an appropriate TMR for testing. None of these participants participated in the actual experiments. The participants had to complete 6 practice trials and 90 test trials. The participants were provided with audio fragments (words, pseudowords, or sentences) in four-talker-babble (4000 ms). The words and pseudowords had three or less syllables. The sentences would always start with the word 'The' and have four target words within each sentence. During the practice trials the target was presented at a TMR of +3 dB, during the test trials TMRs of 0 dB, -3 dB, and -6 dB were presented. The words, pseudowords, or sentences were presented in blocks; however, the order of the blocks was random. The TMRs were randomly mixed within a block. The target would start playing after 1000 ms, and the participant had to type what they heard to the best of their abilities. Every participant was asked to sit in a quiet room, run the experiment online via Google Chrome or Firefox and not wear headphones. The participant had the possibility to take 3 breaks (after 30 trials each).

TABLE 5.3. Results from pilot study to determine TMR.

TMR	**Mean percent correct (auditory-only)**	**Standard Deviation**
0 dB	65.92 %	18.57
-3 dB	48.62 %	24.17
-6 dB	21.20 %	14.57

Mean percent correct performance on an auditory only speech-in-noise identification task, with standard deviation, for 3 different levels of target-masker-ratio (-6, -3, and 0dB) combined over 3 different linguistic levels (words, pseudowords, sentences).

Results from this pilot task showed a good variety of performance between the 3 TMR levels. All participants mentioned it was a challenging task, but not impossible. The results (Table 5.3) favored a TMR choice of -6 dB for the experiments as we are looking for a TMR where participants would not bottom out in the auditory-only

condition, but would still have a wide range for improvement when AV speech is presented at different levels of linguistic complexity.

5.1.5.9. Data collection

All four tasks were completed remotely, but offline. Instructions were provided via email and a website. All tasks were completed in the comfort of the participants' home on their personal computer. Similar remote paradigms have successfully netted consistent and repeatable results even for children (Gijbels, Cao et al., 2021; Gijbels & Lee 2023), which are much harder to control for their general attentiveness.

Tasks for all phases were designed in the free online study builder, Lab.js (Henninger et al., 2021). One of the biggest concerns regarding online assessments of temporally sensitive measures like the TBW is the guarantee of exact stimulus presentations. By pre-assembling all stimuli and executing the experiment remotely but offline, we retained more control.

For all tasks the participants were asked to reside in a quiet environment, in front of a computer (no tablet or phone) with built-in speakers. The participant was asked to not wear headphones and complete all tasks offline via their Google Chrome or Firefox browser. At the beginning of the first task, participants were asked to set the speakers to a comfortable loudness level, based on a repeating English speech fragment. They were requested not to change the loudness setting of their computer for the entirety of the experiment and had to provide an estimate of this setting (0-100%), so the same intensity could have been used for all experiments.

The participant had to download the experimental folder to their personal computer and open an index.html file. At the end of the experiment a .csv file was generated by the experiment builder and the participant emailed this to the researcher. Except for some unzipping issues, most participants reported this process was straightforward.

5.1.5.10. Data analysis

Statistical analyses were performed using the lme4, nlme, stats, and psych packages in R (Bates et al., 2015) and RStudio (version 1.3.1093).

5.2. The role of linguistic information on AV temporal synchrony perception

We asked whether the width of the visual-leading TBW was specific to the level of linguistic complexity of the AV stimulus. We hypothesized (1) that the complexity of the linguistic structure increases the likelihood that AV signals come from the same event and therefore widens the TBW (Vatakis & Spence, 2006), and (2) that increased exposure (i.e., temporal and phoneme-viseme connections), which often goes hand in hand with increased linguistic complexity, leads to accumulation of information over time resulting in more precise temporal estimates, narrowing the TBW width (Maier et al., 2011). As these two mechanisms seem to pull in both directions (i.e., elongating or shortening the TBW), we expected for the five levels of linguistic complexity employed in this task (Table 5.2) that reversed words would have narrower TBWs than pseudowords, and pseudowords narrower than words. Similarly we expected anomalous sentences to have narrower TBWs than sentences. However, we made little predictions about the impact of stimulus duration on this effect when comparing the word type stimuli to the sentence type stimuli.

5.2.1. Methods

5.2.1.1 Participants

In total 46 younger (age 21 - 40), Native English speaking adults with self-reported normal hearing and normal or corrected-to-normal vision successfully completed the task. More demographic information is reported in Table 5.1.

5.2.1.2. Temporal asynchrony measure

To measure the TBW a One-Interval Two-Alternative Forced Choice Simultaneity Judgement Task (1I-2AFC-SJT) was employed. We asked the participant to judge whether the auditory and visual stimulus were presented at the same or different times. This subjective judgment task was chosen because it has been shown as the more stable measure among TBW tasks (van Eijck et al., 2008; Vroomen & Stekelenburg, 2011). Synchronous or visual-leading stimuli were presented. Eight SOAs (0 ms, 50 ms, 100 ms, 150 ms, 200 ms, 300 ms, 400 ms, 500 ms) were selected, within a range (0-500 ms) based on earlier work with speech stimuli from Stevenson and colleagues (2012). All AV stimuli were generated offline using ffmpeg software (Python 3.7) to ensure intended SOAs between the auditory file and its corresponding silent video file.

As described in Section 5.1.5, all stimuli were presented in 4000 ms 4-talker-babble fragments at a TMR of -6dB. All target stimuli were recorded by female talkers, based on a collection of validated stimuli sourced from studies that have already been published (Al-Salim et al., 2020; Kocins et al., 2022; Grieco-Calub et al., 2023; Holt et al., 2011; Shatzer et al., 2018; Stelmachowicz et al., 2000; Van Engen et al., 2014; SteVi n.s). All videos were presented in a black frame with a cut-out circle showing the head and the shoulders of the talker. Before and after the stimulus video a black frame was added to provide a visual representation for the total duration of the stimulus presentation. Each stimulus type had three different talkers to average out any other unforeseen effects linked to the stimuli.

Each of the 8 SOAs was presented 20 times (once for every stimulus), resulting in 160 trials per stimulus type. By using 5 different levels of linguistic complexity, this results in a total of 800 trials. Table 5.2 shows the different stimulus types with the different levels of information each provides.

Each trial consisted of the presentation of a fixation cross (500 ms), followed by an AV stimulus presentation (4000 ms) and a participant response where the participant indicated whether the audio and video were presented at same (pressing [s] on keyboard) or different times (pressing [d] on keyboard). The 800 trials were broken down in 10 blocks. The stimuli were blocked by linguistic level (5 blocks), and these were randomly broken in half (80 trials per block). All 10 blocks were presented in a random order. As a result, we were able to obtain sufficient data to estimate 5 sigmoid curves per participant (one per linguistic level).

5.2.1.3. Data analysis

Item analysis

For all words used in this task we defined number of syllables, phoneme length, word frequency per million and word density, based on https://clearpond.northwestern.edu/clearpond_database.cgi. We then analyzed whether any of these characteristics had a significant effect on the mean judgment score (same or different) by using a linear model (Model 5.1: Mean judgment score ~ word frequency + word density + (number of syllables * phoneme length)).

Talker analysis

We used a second linear model (Model 5.2: Mean judgment score ~ talker) to analyze the mean judgment score by talker. Sum contrast coding was used for the categorical variable (talker) so we could compare the mean of the mean judgment score for a given talker to the overall mean.

Response time analysis

Our data collection further allowed us to analyze response times. Although the presented noise fragments always have the same length, we might find different response times for short stimuli like words and pseudowords in contrast to sentences. We used a linear mixed effects model (Model 5.3: Response time

~linguistic level * SOA + (1|participant)), with participant as random effect, and used contrast sum coding for both the linguistic levels and the different SOAs.

TBW analysis

The [s]/[d] responses were transformed to 1 and 0. This allowed us to calculate the percentage perceived synchronous for all stimulus conditions (5 linguistic levels * 8 SOAs), with a higher score meaning more synchronous stimulus perception.

We used a generalized linear model (GLM) with a binomial distribution (logit link function) to obtain a sigmoid fit for each linguistic level. This allowed us to define (1) an upper and lower asymptote, (2) a slope, and (3) the TBW-width (defined as the SOA at the 50% synchronous point) for each linguistic level. We did this on a group level, and for every individual separately.

Mixed effects models are useful when there is variation in the effect of a factor (i.e., here TBW) across individuals, but some of the variation is systematic (i.e., suggested linguistic levels) and some is random (i.e., high between-subject variability of TBW). We therefore used linear mixed effects models to see whether linguistic complexity predicts any characteristic of the TBW (Model 5.4a: TBW width ~ linguistic level + (1|participant), Model 5.4b: Asymptote amplitude ~ linguistic level + (1|participant), Model 5.4c: Slope ~ linguistic level + (1|participant)). Meaningful sentences were used as a reference for the treatment coding of the variable linguistic level, and participant was included as a random effects factor.

To further distinguish the effect of stimulus duration, potentially resulting in extra temporal information and extra exposure to phoneme-viseme connections, we included 1-,2- and 3- syllable words, pseudowords and reversed words. We analyzed the effect of the number of syllables by linguistic level on the TBW width, by using a linear mixed effects model (Model 5.5: TBW width ~ number of syllables * linguistic level+ (1|participant)), with participant as random effect, and 1-syllable and word as reference values.

Response times were not included in these final models as there was no effect of adding response time to the models for any of the measures of interest (TBW width, asymptote amplitude, slope).

The difference between performance of the TBW width on an individual level were interpreted via Pearson correlation plots between the 50% TBW width of meaningful sentences (y-axis) and the 50% TBW width of all other linguistic categories (x-axis). We interpreted both the Pearson correlation coefficient and the slope. We compared the slopes between the different categories (e.g., slope 1: TBW width meaningful sentences ~ TBW width words versus slope 2: TBW width meaningful sentences ~ TBW width pseudowords) by using t-statistics. And used similar methods to compare the slope to 1 (e.g., slope 1: TBW width meaningful sentences ~ TBW width words versus slope 2 =1). This allowed us to interpret the relationship between different linguistic stimuli (reversed words, psuedowords, words, and anomalous sentences), referenced to meaningful sentences and the relationship between each of these linguistic levels and meaningful sentences.

5.2.2. Results

5.2.2.1. Item / talker analysis

Model 5.1 showed there was no effect of number of syllables, phoneme length, word frequency per million or word density on the temporal synchrony judgments for the words used in this task.

We had a total of eight different female talkers across the 5 linguistic categories, with 3 talkers in each category. The linear model (Model 5.2) showed no significant effect of talker.

5.2.2.2. Response time analysis

The mean response time was 3323 ms per trial. Our linear mixed effects model (Model 5.3) showed a significant effect of response time for each linguistic level. This means that every linguistic level was significantly different from the mean. Response times for both words and pseudowords were significantly faster than the mean. Response times for anomalous sentences, meaningful sentences, and reversed words were significantly slower than the mean.

When comparing response time per SOA, we found that response times at 50 ms were significantly faster than at any other SOA. For all linguistic levels, except reversed words, we noticed a trend of longer response times for longer SOAs, however this interaction was not significant.

5.2.2.3. TBW analysis

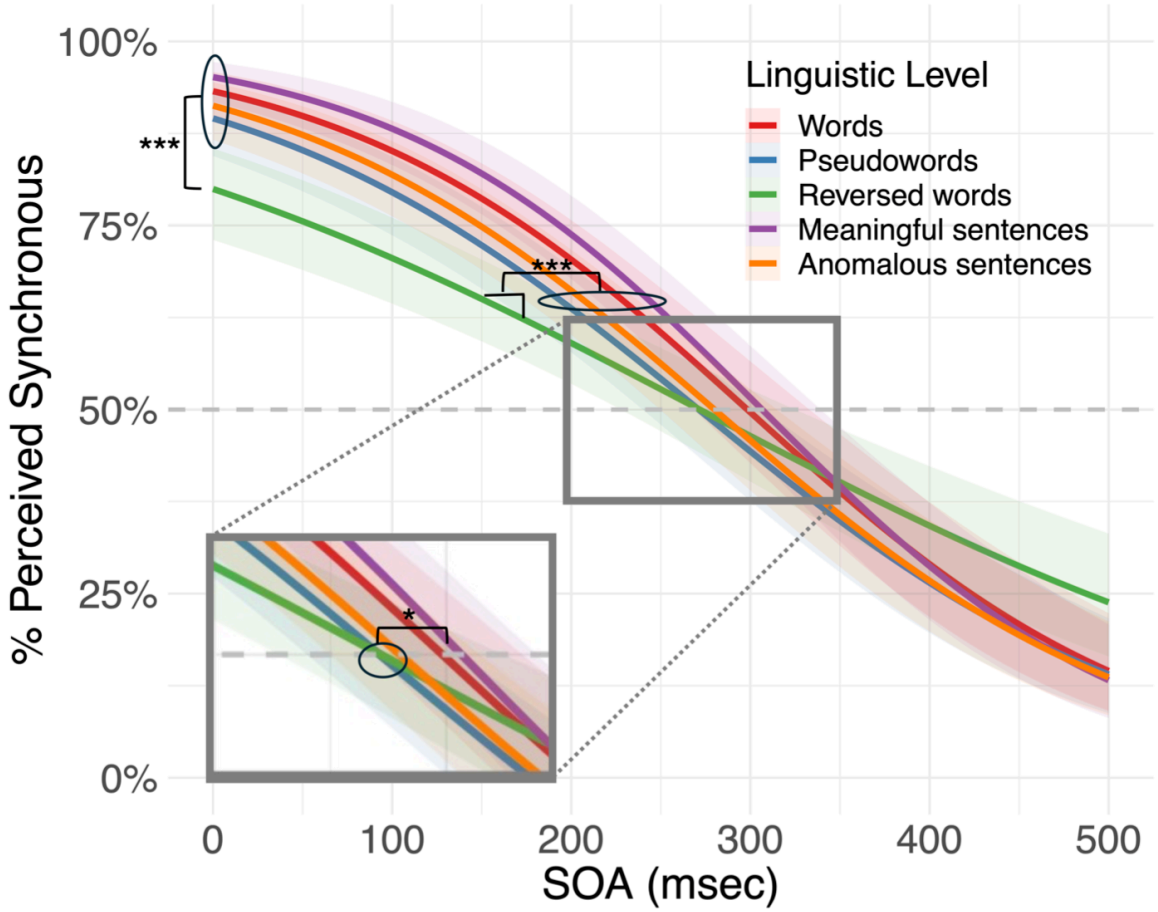
After applying logistic regression models to our dataset, we obtained five sigmoid functions (Fig 5.1). Upon examination of Fig 5.1 and Table 5.4, which detail the relevant values of the functions (such as slope, TBW width, and asymptotes) at a group level, we observed variations in TBW width across different linguistic levels: reversed words = pseudowords < anomalous sentences < words < meaningful sentences. A similar pattern emerged for the slope, ranging from pseudowords to meaningful sentences, with the steepest slope observed for meaningful sentences. Notably, the slope of reversed words was distinctly more shallow compared to the other slopes. Consistent with these slope observations, the amplitude between the asymptotes was notably smaller for reversed words. While minimal differences were observed across the lower asymptotes of other linguistic levels, the upper asymptote increased from pseudowords to meaningful sentences. To observe the spread of performance for TBW width, slope and asymptote amplitude by individuals, see Supplemental Figure A.

We concluded from the linear mixed effects model (Model 5.4a) that the TBW-width for meaningful sentences was significantly different from reversed words ($t = -2.89$, $p < 0.01$), pseudowords ($t = -3.00$, $p < 0.01$) and anomalous sentences ($t = -2.24$, $p < 0.05$), but not from words ($t = -0.62$, $p = 0.54$). Post-hoc analysis showed that the TBW-width of words was not significantly different from the TBW-width of anomalous sentences ($t = -1.62$, $p = 0.11$), but it was from reversed words ($t = -2.21$, $p < 0.05$) and pseudowords ($t = -2.38$, $p < 0.05$).

The slope of the meaningful sentences (Model 5.4b) was significantly steeper than any of the other linguistic levels (reversed words: $t = 12.84$, $p < 0.001$; pseudowords: $t = 5.10$, $p < 0.001$; anomalous sentences: $t = 4.12$, $p = 0.001$; words: $t = 2.59$, $p < 0.05$), and post-hoc analysis revealed a significantly more shallow slope in reversed words compared to any of the other slopes ($p < 0.001$), but the slope between words, pseudowords and anomalous sentences was not significantly different.

Analysis of the asymptote amplitude (Model 5.4c) showed a significantly larger amplitude for meaningful sentences compared to anomalous sentences ($t = -2.33$, $p < 0.05$), pseudowords ($t = -3.21$, $p = 0.001$) and reversed words ($t = -13.35$, $p < 0.001$), but not compared to words ($t = -1.66$, $p = 0.09$). Post-hoc analysis showed a significantly smaller amplitude for reversed words compared to all other linguistic levels ($p < 0.001$), and no significant differences in asymptote amplitude between words, pseudowords and anomalous sentences. However, when solely looking at the upper asymptote, there was a significant difference between words and pseudowords ($t = -2.78$, $p < 0.01$).

FIGURE 5.1. The Temporal binding window of the 5 different linguistic levels (words, pseudowords, reversed words, meaningful sentences, and anomalous sentences).



Percentage perceived synchronous is plotted in function of SOA, ranging from 0 to 500 msec, visual leading stimuli. The sigmoid functions are calculated based on a general linear model with a binomial distribution (logit link function), and the 95% confidence interval is plotted as a shaded band around the sigmoid function. The cross-section of the sigmoid functions and the horizontal gray dashed line represents the 50% synchrony perception point. The TBW-width is obtained by extrapolating the cross-section of this horizontal line with the sigmoid function to the x-axis, resulting in the SOA that corresponds to 50% synchrony perception. Significance: $p < 0.05$ (*), $p < 0.001$ (***)

We further analyzed the effect of added exposure caused by stimulus duration by breaking up our three word level stimuli (words, pseudowords and reversed words) by 1-,2- or 3-syllables. The mixed effects model (Model 5.5) showed no main effect of number of syllables on the TBW width. As expected there was a main effect of linguistic level on the TBW width, but also an interaction effect between linguistic level and number of syllables, showing that the TBW-width was significantly wider for 1-syllable stimuli, but only for reversed words ($p < 0.01$).

In sum, synchrony judgment of AV reversed words resulted in a significantly shallower slope and smaller asymptote amplitude than any of the other linguistic levels. Meaningful sentences had the steepest slope and largest asymptote amplitude; however, the asymptote amplitude was not significantly different from the amplitude of the word stimuli. The TBW-width was similar for pseudowords and reversed words, but for more complex stimuli there was a stepwise distinction with an increase from pseudowords, anomalous sentences, and words, to meaningful sentences.

TABLE 5.4. Summary values of logistic regression of temporal synchrony perception judgments by linguistic level.

Linguistic Level	TBW width (ms)	Slope	Lower Asymptote (%)	Upper Asymptote (%)	Asymptote amplitude (%)
Reversed words	271.7	-0.51	23.81	79.97	56.16
Pseudowords	271.2	-0.79	14.03	89.56	75.53
Words	298.2	-0.88	14.50	93.23	78.74
Anomalous sentences	279.9	-0.84	13.63	91.28	77.64
Meaningful Sentences	306.5	-0.97	13.24	95.16	81.92

For each linguistic level (reversed words, pseudowords, words, anomalous sentences and meaningful sentences) the TBW width (in ms), the slope of the function, the lower and upper asymptote and the difference between the asymptotes (amplitude) is reported.

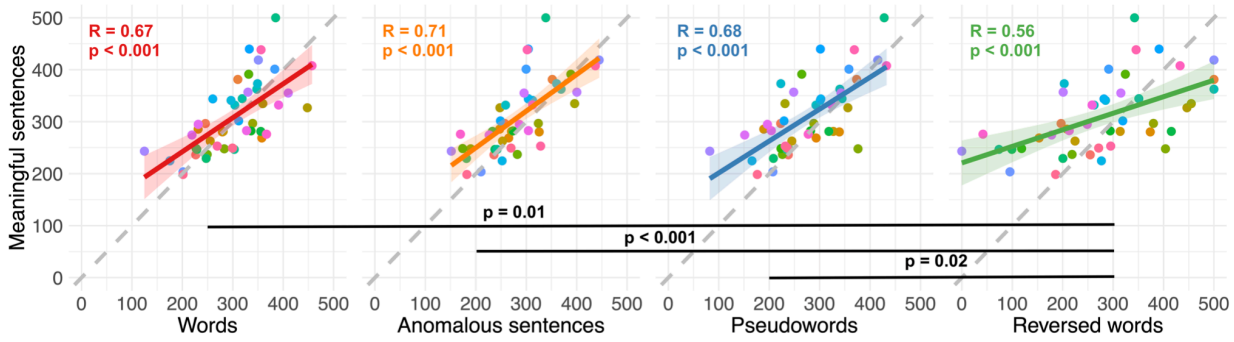
To further interpret differences in the TBW on an individual level we plotted the correlation between the 50% TBW width of meaningful sentences and the 50% TBW width of the other linguistic categories (i.e., words, anomalous sentences, pseudowords and reversed words – Fig 5.2). We interpreted both the Pearson correlation coefficients, and the slopes. We found that the TBW width of all linguistic levels was strongly correlated to the TBW width of the meaningful sentences ($R_{sent.anom} = 0.71$, $p < 0.001$; $R_{pseudowords} = 0.68$, $p < 0.001$; $R_{words} = 0.67$, $p < 0.001$), with

only the TBW width of the reversed words showing a moderate correlation ($R_{rev.words}$ = 0.56, $p < 0.001$). These strong correlations suggested that individuals who have wide TBWs for meaningful sentences, are more likely to also show wide TBWs for the other linguistic categories, and vice versa. The lower correlation for reversed words then means that this relation is less strict between meaningful sentences and reversed words.

Furthermore, as we reported that the 50% TBW width was wider for meaningful sentences than for all other categories we would expect the majority of the data points to be on the left side of the diagonal line in Fig 5.2. This can be observed for pseudowords and anomalous sentences. For words they were spread more evenly on both sides of the diagonal. This is not surprising as we did not find a significant difference between the TBW width in these two categories on a group level. On a group level, the TBW width of reversed words was significantly narrower than the TBW width of meaningful sentences ($t = -2.89$, $p < 0.01$), however; on an individual level a noteworthy trend showed. We observed that individuals with narrow TBWs for meaningful sentences were more likely to have even more narrow TBWs for reversed words (left of diagonal), but individuals who had wider TBWs for meaningful sentences were more likely to show even wider TBWs for reversed words (right of diagonal). This resulted in a significantly shallower slope ($\beta = 0.31$, $t = -5.55$, $p < 0.001$) than if the slope would be 1. A slope $\beta = 1$ would be interpreted as a directly proportional relationship between meaningful sentences and reversed words. When comparing this slope to the slopes of the other plots, thus comparing the linear relationship between meaningful sentences and reversed words with the linear relationship between meaningful sentences and other linguistic categories ($R_{mean.sent.*rev.words}$ vs. $R_{mean.sent.*pseudo.words}$; $R_{mean.sent.*rev.words}$ vs. $R_{mean.sent.*words}$: $R_{mean.sent.*rev.words}$ vs. $R_{mean.sent.*anom.sent.}$), we reported significant differences for all comparisons ($t_{pseudoword}$ = -2.41, $p = 0.02$; $t_{sent.anom}$ = -3.01, $p < 0.001$; t_{word} = -2.56, $p = 0.01$). None of the linear relationships between meaningful sentences and the other three categories were

significantly different from each other ($R_{mean.sent.*words}$ vs. $R_{mean.sent.*pseudo.words}$; $R_{mean.sent.*words}$ vs. $R_{mean.sent.*anom.sent}$; $R_{mean.sent.*anom.sent}$ vs. $R_{mean.sent.*pseudo.words}$). Thus, this means that the proportional relationship between meaningful sentences and reversed words was significantly different from the proportional relationship between meaningful sentences and the other linguistic categories, and that the proportional relationship between meaningful sentences and words, pseudowords, or anomalous sentences did not differ from each other. Accordingly, the linear relationship between meaningful sentences and all other linguistic categories showed the same trend, yet it was significantly more expressed for reversed words.

FIGURE 5.2. Correlational plots of 50% TBW width measures between meaningful sentences and all other linguistic categories.



The Pearson correlation coefficients between the 50% TBW width of meaningful sentences (y-axis) and the other categories (x-axis) are plotted with their p-value in the left top corner. The p-values at the bottom represent the significant differences between the slopes of the four plots. Individual participants are represented as colored dots. The gray dashed line represents a slope of 1. The TBW width is plotted in ms.

We also interpreted whether the slopes were significantly different from 1. A significant difference for all linguistic categories ($\beta_{pseudoword}$ = 0.61, t = -2.73, p = 0.008; $\beta_{sent.anom}$ = 0.70, t = -2.02, p = 0.049; β_{word} = 0.65, t = -2.33, p = 0.02), thus showed that the proportional relationship between meaningful sentences and all other categories was not directly related.

5.2.3. Discussion

In this first task we aimed to describe the role of linguistic information on AV temporal synchrony perception. On a group level our results showed a significant effect of linguistic complexity on temporal synchrony perception. A step-wise change in TBW-width, upper asymptote and slope was reported, suggesting that listeners were less sensitive to asynchronies of linguistically more complex stimuli. On an individual level we reported that individuals with narrow TBWs in the most complex linguistic category (i.e., meaningful sentences) were more inclined to have even more narrow TBWs in the other linguistic categories, whereas individuals with wide TBWs for meaningful sentences were more inclined to have similar or even wider TBWs in the lower linguistic categories.

For reversed words we found a similar TBW width as for pseudowords. However, for reversed words, listeners were significantly more variable in their synchrony and asynchrony judgments within a certain SOA, resulting in smaller asymptote amplitudes and shallower slopes. This most likely indicates that listeners were less certain about temporal synchrony decisions for reversed speech signals. Maier and colleagues (2011) similarly reported that the upper asymptote of the sigmoid function for synchronous reversed AV speech signals never reached the expected maximum (i.e., close to 100% synchronous reported). This increased uncertainty could be explained by not being familiar with this type of speech signal. Interestingly, Vatakis and Spence (2006) argued that their similar finding of no differences in TBW width between reversed and forward played speech points to reversed speech resembling responses to another language, and therefore not being an unfamiliar stimulus. Navarra and colleagues (2010) then measured the TBW for sentences (in quiet, in English and in Spanish) for individuals who either were proficient in both languages or who had close to no exposure in the second language. They reported no differences in asymptote amplitude between the two languages. However, they did find that the TBW width was more narrow for the

non-native language, but only for participants who were not proficient in the second language. They argued that prior exposure (i.e., linguistic knowledge of the individual) caused these differences in TBW width between participants.

Translating this finding to our stimuli, where we reported no differences in the width of the TBW between reversed words and pseudowords, but a significant increase in TBW width for words compared to either reversed words or pseudowords, seems to concur with the explanation by Navarra and colleagues (2010). The listeners in our task could not use any prior exposure to higher level linguistic information for either reversed words or pseudowords. Furthermore, as we found no differences in TBW width between pseudowords and reversed words, low-level cues like phoneme-viseme connections – which are absent in reversed words but not in pseudowords – most likely did not significantly influence the TBW width. It is more likely that the lexical-semantic information provided by the knowledge of the word created the shift in TBW.

The difference between reversed words and pseudowords for our study was reported in both the asymptote amplitude and slope. As phoneme-viseme connections are the main difference between these two stimulus types, it could be that familiar phoneme-viseme connections led to increased certainty of temporal synchrony judgments. That this was not reported by Navarra and colleagues (2010) is most likely because their 'unfamiliar' language had 14 phoneme-viseme connections in common with the native language (English - Spanish; Whitley, 2002). They used sentence stimuli in their experiment and therefore it was very likely that sufficient familiar phoneme-viseme connections were provided to make temporal synchrony decisions. Alternatively, Kim and Davis (2003) reported that a high correlation between the acoustic envelope and the mouth movements in combination with prelinguistic coherence (temporal synchrony) was needed to result in AV detection benefits, and that not knowing the language could have a negative outcome on the perceptual processing of these stimuli. So, as our reversed words

created changes in the relationship between the acoustic envelope and the mouth movements, and participants were not familiar with these stimuli, a lack of AV integration, as shown in disrupted TBW shapes, could be hypothesized.

Nonetheless, the idea that stimulus familiarity plays a role in the asymptote amplitude and slope remains plausible (Maier et al., 2011). We did not expect our participants to be familiar with either reversed word or pseudoword stimuli; both reported here with the smallest amplitudes and most shallow slopes. However, as the asymptote amplitude and slope was significantly different for reversed words and pseudowords, but not for pseudowords and anomalous sentences, we interpret that stimulus familiarity might not completely explain this asymptote and slope effect. Therefore, it is plausible that not being provided with familiar phoneme-viseme connections in reversed speech stimuli resulted in additional increased uncertainty in the synchrony decision. This idea is further corroborated by presenting individuals with incongruent phoneme-viseme connections (i.e., the McGurk effect, McGurk & Macdonald, 1976). The phoneme-viseme connection for these incongruent stimuli is also disrupted and lower asymptotes for the TBW of these stimuli were observerd (van Wassenhove et al., 2007).

Interestingly, neither of these measures of interest (TBW width, slope, asymptote amplitude) represented an increase of linguistic complexity in the exact order we expected. As defined in Table 5.2, we argued that linguistic complexity increased from reversed words, to pseudowords, to words, to anomalous sentences, to meaningful sentences as the following linguistic characteristics were added: phoneme-viseme connections, lexical knowledge, morpho-syntactic knowledge and predictable linguistic context. However, we found that the TBW shape did not significantly differ for anomalous sentences and words, and if anything there was a trend in the opposite direction, with the anomalous sentences resulting in narrower, shallower, less expressed (i.e., smaller asymptote amplitude) sigmoid functions. From a linguistic viewpoint the difference between these two stimuli was twofold:

(1) the predictable morphosyntactic structure of an anomalous sentence compared to words and (2) more phoneme-viseme exposure with increased exposure in the sentence structure. From a prelinguistic point of view we also added stimulus duration, allowing to accumulate more temporal information (onset /offset and temporal envelope). Vatakis and Spence (2006) argued that this increased linguistic information would widen the TBW, because it results in increased probabilities that the AV signal comes from the same event. However, Maier and colleagues (2011) argued that increased temporal information should lead to more precise temporal estimates.

Analysis of the number of syllables within each word category (words, pseudowords, reversed words) allowed us to gain more insight in these contradictory results. There was no effect of the number of syllables on the TBW width for words and pseudowords. However, there were for reversed words. One syllable reversed words resulted in significantly wider TBWs than 2- or 3- syllable reversed words. As reversed words contained no linguistic information (Sheffert et al., 2002), the increased amount of temporal information in 2- and 3-syllable reversed words might have resulted in more precise decisions. However, for words and pseudowords that have phoneme-viseme connections to inform us about stimulus synchrony, the added temporal information from 1- to 3- syllables might only have been used sparsely. This is in line with the suggestion of hierarchical integration with descending weights allocated to lexical, segmental, and prosodic cues suggested by Mattys and colleagues (2005). They state that only when interpretive conditions are altered due to a lack of contextual and lexical information, lower-level cues become the driving force behind segmentation.

Thus, considering all linguistic elements we manipulated (i.e., phoneme-viseme connections, lexical information, predictable morpho-syntactic structure, and predictable context), and the temporal differences between the linguistic categories (number of syllables in word type stimuli and words vs. sentences), we conclude (1)

that phoneme-viseme connections improve the certainty of the temporal (a)synchrony decision, (2) that lexical-semantic information as provided by words and meaningful sentences results in a higher tolerance for asynchronous stimuli (i.e., wider TBWs), as they are more likely to represent a single AV event, that (3) the meaning of the stimulus might be more of an influence than solely lexical information (pseudowords and anomalous sentences vs. words and meaningful sentences), (4) that the predictable morpho-syntactic structure and predictable context of sentences had little influence on the TBW width, and that (5) exposure to longer stimuli only led to increased sensitivity of temporal asynchrony (i.e., more narrow TBW) when contextual and lexical information was absent (i.e., 1 vs. 2 or 3 syllables in reversed words).

Stevenson and colleagues (2014) formulated that an effect of late-stage integration mechanisms (like phoneme-viseme connections and other linguistic elements) on the TBW shape point to parallel accumulation of evidence, rather than multi-stage integration (Peelle & Sommers, 2015). Our findings could strengthen this reasoning, however it could also be that because we measure what we manipulate (i.e., temporal (a)synchrony), rather than a feature orthogonal to it, that late-stage integration information (i.e., linguistic information) has impacted our judgment, even within a multistage integration model. Therefore, at least in behavioral studies it might be most informative to discuss the work from an angle of prelinguistic and linguistic influences on AV integration, rather than looking at integration stages.

Interestingly, our findings on a group level did not necessarily explain individual variability. By plotting the correlation between the TBW width of meaningful sentences to the TBW width of all other linguistic categories we reported two interesting findings. First, the TBW width of all linguistic levels was significantly correlated to the TBW width of meaningful sentences. Thus individuals who have wider TBWs in one category also have wider TBWs in the other categories, and vice versa. Although significant, this correlation was only of moderate strength between

the width of the TBW of meaningful sentences and reversed words, which is not surprising as these were the outer ends of our linguistic continuum. Second, for all linguistic levels, the slope was significantly different from 1, suggesting that there was no directly proportional relationship between meaningful sentences and the other linguistic categories. The effect was most pronounced for reversed words, but overall we found that the wider a listener's TBW for meaningful sentences, the less likely that the TBWs of the same listener would be narrower in the other linguistic categories. For reversed words it would even mean that individuals with large TBWs for meaningful sentences were more likely to have significantly wider TBWs for reversed speech. These findings suggest that although there is overall a good correlation between individual's performance on temporal (a)synchrony tasks for different linguistic levels, individual performance will uniquely explain the effect that linguistic complexity has on temporal synchrony perception.

A last interesting note to make is that participants were significantly faster (i.e., shorter response times) for stimuli at SOA = 50ms, across linguistic levels. As visual speech naturally precedes auditory speech (Chandrasekaran et al., 2009), this SOA might represent the most realistic scenario and therefore results in the fastest decision.

In sum, we conclude

(1) that high-level linguistic information changes the TBW width, slope and upper asymptote in a stepwise manner, where higher linguistic complexity results in wider, steeper and more expressed sigmoid functions and items that provide meaning (words and meaningful sentences) have a higher tendency to be integrated than meaningless items (reversed words, pseudowords, anomalous sentences),

(2) that low-level linguistic information (phoneme-viseme connections) influences the certainty level of temporal synchrony perception,

(3) that temporal information improves temporal estimates, leading to more narrow TBWs; however, increased stimulus length only has a limited effect in contrast to the linguistic effect on the TBW shape,

and

(4) that individual performance on temporal (a)synchrony tasks is significantly correlated within different linguistic levels, and individual temporal synchrony perception can explain the variability of the effect caused by differences in linguistic complexity of the stimuli.

5.2.4. Potential limitations

We acknowledge that our task set up deviated from most of the previous work assessing temporal synchrony perception in AV speech, and therefore want to address some potential limitations of data interpretation that come with this.

First, this task was assessed remotely. To our knowledge, there currently has only been one report of temporal binding window measures in an online or remote format (Flannagan, Zumbrunn, et al., 2020). Although they used a non-speech illusion as stimuli, they were successful in replicating their in-lab findings to an online format. Nonetheless, we acknowledge the difficulties that come with temporal synchrony measures in an online format as the synchrony is dependent on aspects like network latency and buffering, and each personal device has different playback device performance etc. We tried to largely overcome this by (1) pre-assembling all video files with the audio of both the signal and the noise, (2) reducing the file size while keeping the quality of the sound and video (using https://handbrake.fr/) to induce a reduction of loading time, (3) and assessing the task remote, but offline. The well fitted sigmoid functions, with a visual-leading TBW width in the order of 271 - 306 ms are in line with earlier findings (Hay-McCutcheon et al., 2009; Maier et al.,

2011), and we therefore conclude that our temporal synchrony measure was assessed successfully remotely.

Second, we wanted to make note that our lab did not record any of the stimuli (nor the masker). All audio and video files were validated stimuli obtained via other research laboratories (Al-Salim et al., 2020; Kocins et al., 2022; Grieco-Calub et al., 2023; Holt et al., 2011; Shatzer et al., 2018; Stelmachowicz et al., 2000; Van Engen et al., 2014; SteVi n.s). Although it might not be common to use so many (i.e., eight) different talkers across different levels of the variable of interest, our results showed no talker-specific effect on temporal synchrony perception, and the potential effects of using multiple talkers (i.e., slower response times - Heald & Nusbaum, 2014; or lower speech recognition scores - Kaiser et al., 2003) are not expected to be different between different linguistic levels, which is our interest rather than absolute outcomes.

Third, we are aware that our remote task set-up provided limited control over (1) participant characteristics, (2) computer set-up, (3) quality of the audio and visual signal, and most importantly (4) the actual intensity of the stimuli. By deliberately taking this into account in task development, we aimed to minimize these influences. For example, (1) we had the participants fill out two screening forms so we could compare consistency in report of participant characteristics, (2) our task only worked on computers via Google Chrome or Firefox web browsers. The task would not export correctly in mobile versions (i.e., phones or tablets), nor would it run without crashing in different browsers (e.g., Safari). We believe that this provided us extra control over the actual use of intended hardware. Yet, (3) we had no control over whether the participant was actually honest about not using headphones. We aimed to limit headphone use because of large differences in headphone quality and the use of potential noise cancellation mechanisms. However, as these effects would be similar during the entire task, it would introduce individual variability on an absolute level, but not in regards to the relationship between the different variables

used in the study. We similarly approached the differences in background noise in participants' home environment or even (4) the intensity setting of the computer. As long as these were comparable across the duration of one task (and for all tasks completed) they had little effect on our measure of interest, since this was a comparison measure rather than an absolute measure (Gijbels & Lee, 2023).

Participants had to set a specific speech fragment to a comfortable loudness level and report the intensity level of their computer. As shown in Table 5.1, the intensity settings varied from 25% to 100% (μ = 65.83 % , Q2 = 74 %, sd = 18.56%). Although this is a wide range, noise tolerance is different for different individuals (Weinstein, 1978), and a TMR of -6 dB might therefore be too intense for some individuals at higher intensity levels. Furthermore, post hoc analysis showed no correlation between the intensity setting of the computer and the TBW width ($R_{Pearson}$ = 0.005, p = 0.93), or the response time ($R_{Pearson}$ = 0.19, p = 0.20).

Finally, there were two limitations to our speech stimuli choices. First, it would have been informative to add word strings with no morpho-syntactic structure (e.g., Ball piano working red submarine) to better interpret the effect of stimulus duration and repeated phoneme-viseme exposure. Furthermore, we are aware that reversed speech is not an ideal speech stimulus as it also reverses some information within the temporal speech envelope. Although further analysis about our stimuli are warranted, reversing the envelope might result in different attack/decay characteristics of the speech envelope (Stecker & Hafter, 2000).

5.3. Variability and reliability of temporal synchrony perception between and within individuals

A significant body of literature delves into the perception of temporal synchrony as an indicator of AV integration. More recently, studies have highlighted a profound connection between the TBW measure and specific neurodevelopmental disorders

(for a review: Wallace et al., 2020). However, the significance of these findings hinges on two critical factors: (1) the test-retest reliability of the TBW measures and (2) whether temporal synchrony perception is characteristic to an individual. Interestingly, to the best of our knowledge, there are very limited reports (Flannagan, Zumbrunn et al., 2023 – although not yet peer-reviewed) of the test-retest reliability of the TBW on a group level, or at an individual level. Hence, this study seeks to verify whether the TBW represents an individual trait characterized by substantial intersubject variability but minimal intrasubject variability across sessions, and whether this varies across different levels of linguistic complexity.

5.3.1. Methods

5.3.1.1. Participants

All participants from task 1 (N= 46) were asked to repeat a shorter version of the task for similar financial compensation. Eighteen participants agreed to complete the task again (Table 5.1). The average age was 28.5 (Q2 = 26.5, sd = 6.02, min = 21, max = 40), 4 participants identified as male, 14 as female.

All participants completed this task within a month of the first task.

5.3.1.2. Temporal synchrony measure

Stimuli and presentation mode were identical to task 1 (Section 5.2), with the following exceptions:

1. Only 3 linguistic levels were selected instead of 5: words, pseudowords and reversed words. These three linguistic levels were deemed to be sufficient to show distinct linguistic differences in TBW width, slope, and asymptote amplitude, and any temporal effects caused by stimulus length would be minimized.

2. Participants did not have to complete the hearing screening or the training again at the beginning of this task.

Due to the reduction in levels of linguistic complexity, the participants only had to complete 480 trials, and therefore 6 blocks (2 blocks for each linguistic level). The task could be completed in about 25 minutes.

All instructions were repeated at the start of the task and the intensity level of the computer audio they provided in Task 1 was emailed back to the participant alongside with the task file.

5.3.1.3. Data analysis

Outliers

We used 3 different steps to define outliers.

(1) Residuals were analyzed based on a generalized linear mixed effects model with binomial distribution (Model 5.6: Percentage perceived synchronous ~ test moment * linguistic level + SOA + (1|participant)). The percentage synchrony perception was calculated for each participant, linguistic level, test moment and SOA, based on their synchronous/asynchronous judgments. This resulted in 48 scores (3x2x8) per participant. Participant was included in the model as a random effect. We used a QQplot to analyze the residuals. Although a QQplot of the data is not valuable for logistic regression, a QQplot to interpret the distribution of the residuals to define outliers is adequate.

(2) Scores (% perceived synchronous) were analyzed per participant (and linguistic level) for test moment 1 and 2 to identify which participant had scores more than 2 standard deviations away from the mean score of the group.

(3) Response times were analyzed per participant to identify which participant had response times more than 2 standard deviations away from the mean response time. This was calculated per test moment and linguistic level.

A participant was identified as an outlier if they failed on all 3 steps.

Group level response time

The response time was analyzed between the 2 test moments to see whether overall response time changed with increased familiarity of the task. We ran a linear mixed effects model (Model 5.7: Response time ~ linguistic level * test moment + (1|participant)), to investigate how linguistic level, test moment, and potentially the interaction between these two predicted response time. We implemented participant as a random effect and used linguistic level 'reversed words' and test moment 1 as references (treatment coding). Reversed words were chosen as we expected the largest effect of familiarity in this category.

Group level test-retest reliability of the TBW

To analyze test-retest reliability on a group level we identified 3 linear mixed effects models. We used TBW width (Model 5.8a), slope (Model 5.8b), and asymptote amplitude (Model 5.8c) as dependent variables whereas the right half of the model (independent variables) stayed the same: ~ linguistic level * test moment + (1|participant). We used test moment 1 and linguistic level 'words' as references in our treatment coding, and participants were included as a random effect factor.

Individual level test-retest reliability of the TBW

To assess individual test-retest reliability we used Pearson correlation coefficients to test the correlation between % perceived asynchrony at test moment 1 and 2. For each linguistic level we compared scores (at each SOA) between test moments per participant. Furthermore we correlated the width of the TBW of test moment 1 and 2,

per individual, split up by linguistic level. To show little intersubject variability both correlation coefficients were expected to be high.

Finally, we interpreted the intersubject variability. Combining information of both intrasubject and intersubject variability allowed us to conclude whether the TBW width is characteristic to the individual. To quantify the intersubject variability we interpreted the interquartile range (Q3 - Q1) for each linguistic level.

In order to measure whether the intersubject variability was significantly larger than the intrasubject variability between the two test moments we extracted the variance components from the Model (5.8a), while including test moment in the random effects to interpret intercepts for each participant, and within each participant for the different test moments. We then used a likelihood ratio test to see whether intrasubject variability (between test moments) was significantly smaller than intersubject variability.

5.3.2. Results

5.3.2.1. Outliers

QQplot

The QQplot showed a good fit for our data. From an analysis on an individual level we found that 2 participants (P2 and P16) acted as outliers.

Z-scores

Percentage perceived synchronous

Based on the z-scores of the percentage perceived synchronous, averaged across SOAs but split up by linguistic level, and test moment, 2 participants (P16 and P13) had at least one value more than 2 standard deviations from the mean score.

Response times

The mean response time per trial (across participants, test moments, and linguistic levels) was 3055 ms (Q2 = 2808 ms, sd = 781.18). Based on the z-scores per linguistic level, and test moment, 4 participants (P16, P2, P4, P10) had at least one value more than 2 standard deviations slower than the mean score. Since we had no instructions on response time, and told participants they could take as many breaks as they would like, even within a block, no one was excluded solely based on response times.

However P16 failed all three outlier criteria, and their data was therefore excluded for all further analyses.

5.3.2.2. Group level response time

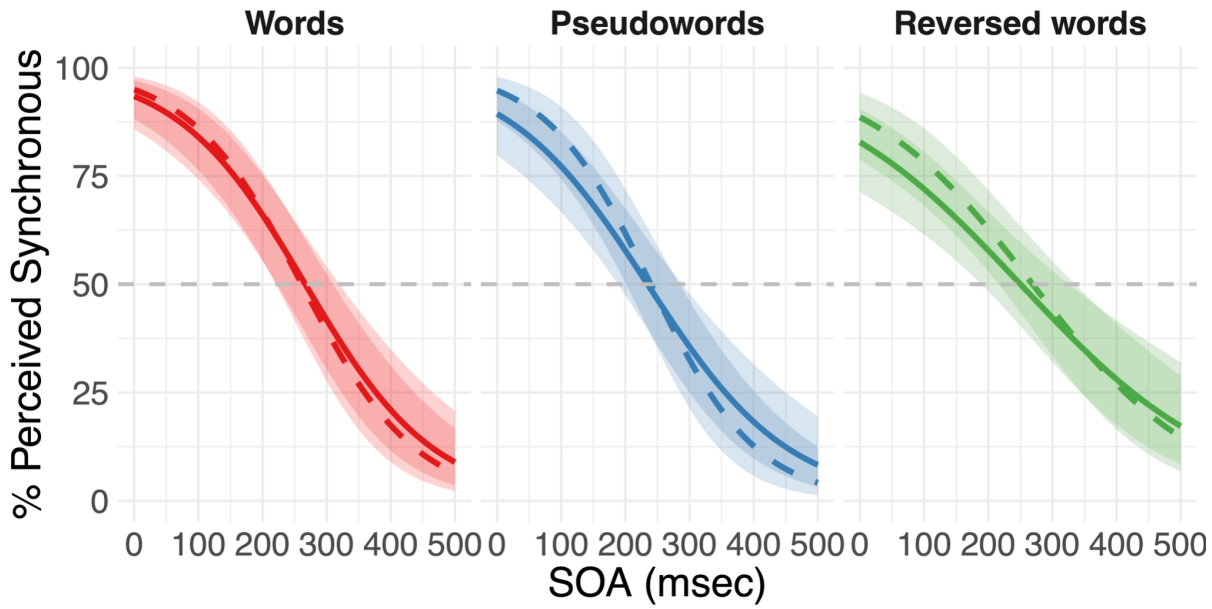
The linear mixed effects model (Model 5.7) showed a main effect of linguistic level and a main effect of test moment on response time. No significant interaction effects were reported. Response times for words and pseudowords were not significantly different; however, responses for both of these linguistic levels were significantly faster (-743 ms, -919 ms) than response times for reversed words (t_{words} = -5.54, $t_{pseudowords}$ = -6.85 and $p < 0.001$). Responses were also significantly faster in test moment 2 (492 ms, t = -3.67, $p < 0.01$). The data also revealed that the response time of individuals mostly conserved across sections, i.e., the slower individuals in test moment 1 were also relatively slower in test moment 2, albeit the absolute response time in test moment 2 was slightly faster due to a potential familiarity effect. This response time pattern was also conserved across linguistic levels.

5.3.2.3. Group level test-retest reliability of the TBW

We used a linear mixed effects model (Model 5.8a) to analyze the influence of test moment and linguistic level (and the interaction of those two) on the width of the TBW. There was no main effect of test moment on the TBW width, suggesting the

TBW width did not significantly differ between test moment 1 and 2. There was a main effect of linguistic level, but only between words and pseudowords ($t = -2.54$, $p = 0.01$), showing that the TBW width of pseudowords was significantly more narrow (32.08 ms). In our larger cohort (Section 5.2) we reported an additional significant difference in TBW width between words and reversed words. We did not find that in this subset of participants. From Fig. 5.3, it could be argued that the TBW width for reversed words of test moment 2 was wider than it was for test moment 1, and this might explain the change from our large cohort in Section 5.2, yet this effect was not significant.

FIGURE 5.3. The Temporal binding window of the 3 different linguistic levels (words, pseudowords, reversed words) for the two different test moments.



Test moment 1 = full line, test moment 2 = dashed line. Percentage perceived synchronous is plotted in function of SOA, ranging from 0 to 500 msec. The sigmoid functions are calculated based on a general linear model with a binomial distribution (logit link function), and the 95% confidence interval is plotted as a shaded band around the sigmoid function. The cross-section of the sigmoid function and the horizontal gray dashed line represent the 50% synchrony perception point. The TBW width is obtained by extrapolating the cross-section of this horizontal line with the sigmoid function to x-axis, resulting in the SOA that corresponds to 50% synchrony perception.

We then applied the same model to the slope of the TBW sigmoid functions by test moment (Model 5.8b). Similar to what we observed in the large cohort (Section 5.2),

there was a significantly more shallow slope for reversed words ($t = 5.46$, $p < 0.001$), and a shallower, but not significantly different, slope for pseudowords compared to words. Furthermore, there was a main effect of test moment, showing the slope was more steep at test moment 2 ($t = -2.40$, $p < 0.05$). There was no significant interaction effect between linguistic level and test moment. Post-hoc analysis showed however that the slope change was more limited for reversed words.

Finally, we analyzed the asymptote amplitude (Model 5.8c). There was a significant difference in asymptote amplitude between reversed words and both word ($t = -6.41$, $p < 0.001$) and pseudoword ($t = -5.37$, $p < 0.001$) type stimuli, but not between words and pseudowords itself. Exactly as reported for the large cohort (Section 5.2), there was a 3% asymptote amplitude difference between words and pseudowords, and the difference between the amplitude of reversed words and words was in the same order as in the previous section (Section 5.2 = 22%, Section 5.3 = 19%). There was no effect of test moment for words, but there was for pseudowords ($t = 2.99$, $p < 0.01$) and reversed words ($t = 2.82$, $p < 0.01$). All 3 linguistic levels showed a more pronounced sigmoid function at test moment 2. There was no significant interaction effect.

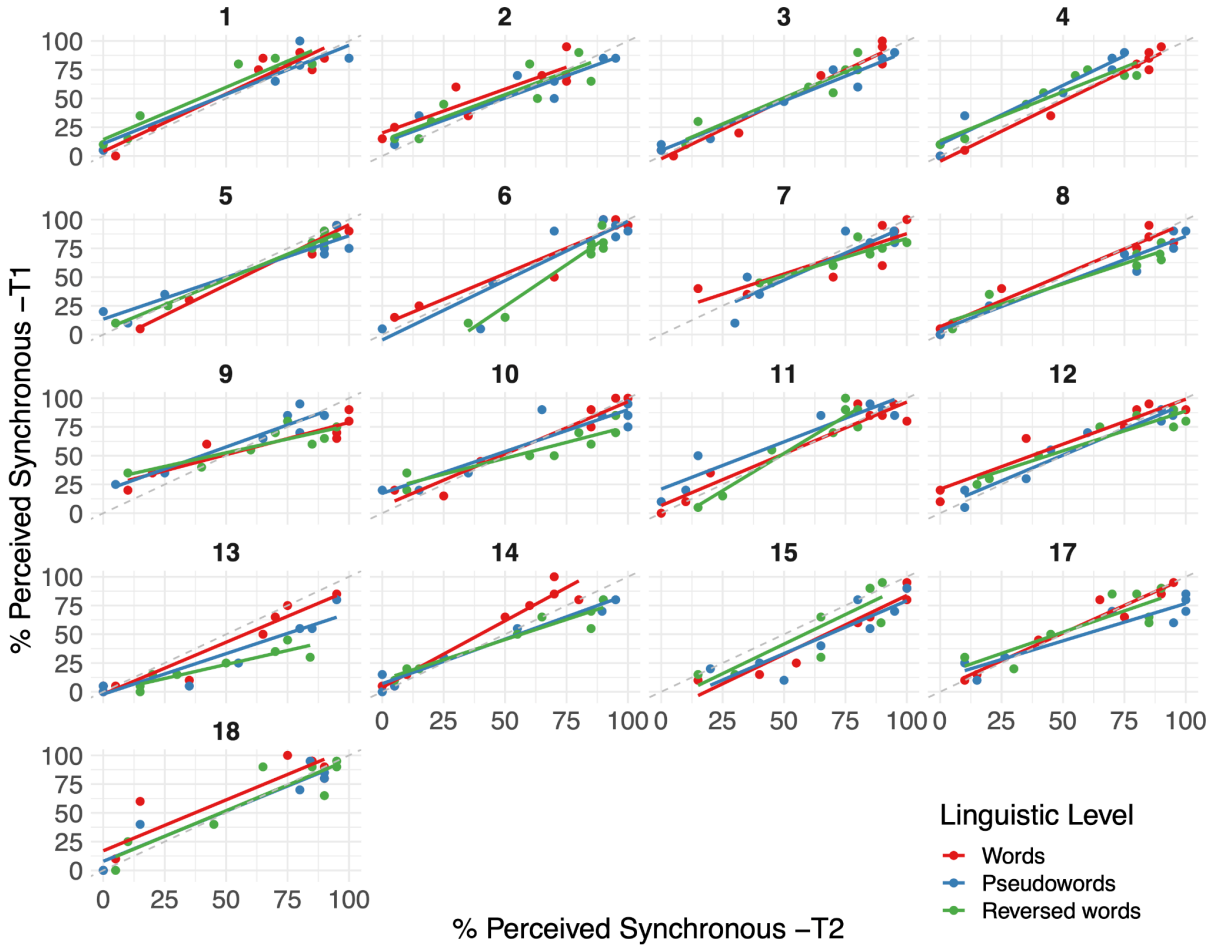
In sum, we found no significant change in TBW width measures between test moment 1 and 2, and the slope and asymptote amplitude improved between test moment 1 and 2.

5.3.2.4. Individual level test-retest reliability of the TBW

We plotted (Fig 5.4) the correlation between test moment 1 and 2 (per linguistic level) for each participant. Each dot represents the percentage synchrony perception for a specific SOA. Given the sigmoid-like nature of the TBW, the responses were rather clustered in the bottom left and top right. Ideally the scores would fall on the diagonal (dashed gray line), indicating that performance on both test moments was identical. Averaged across linguistic levels, we observed a very strong Pearson

correlation coefficient of R = 0.91 (t = 45.06; p < 0.001). Indeed, for most of the participants the correlation plot shows a nice approximation of the diagonal for each linguistic level (R_{words} = 0.93, t = 30.21 ,p < 0.001 ; $R_{pseudowords}$ = 0.92, t_{word} = 27.03 ,p < 0.001; R_{rev_words} = 0.88, t = 21.38 ,p < 0.001).

FIGURE 5.4. Correlational plots of temporal synchrony judgments per individual participant, per linguistic level.

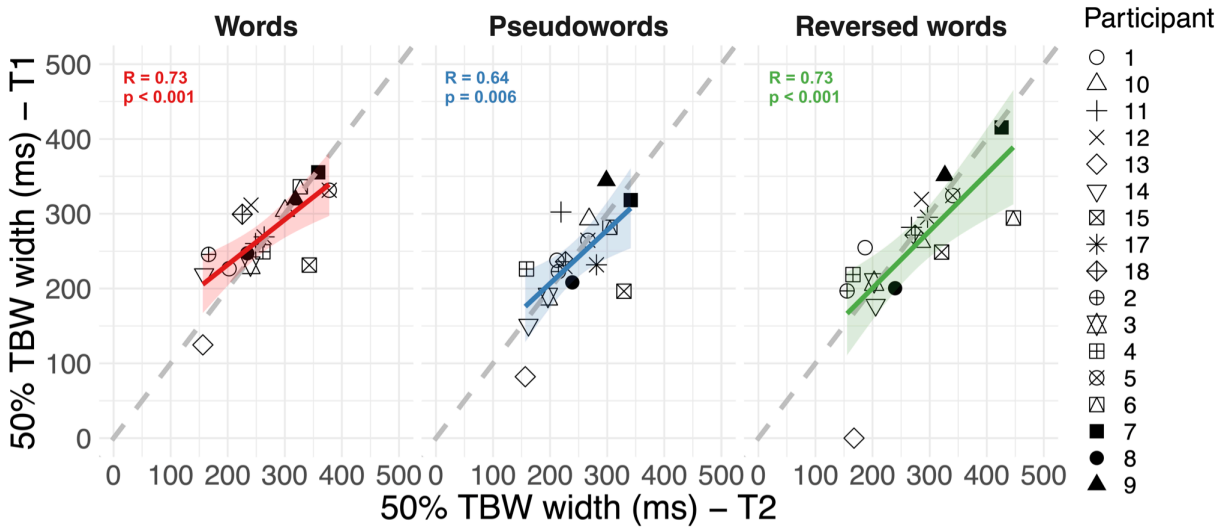


Both axes plot the percentage synchrony perception. The y-axis represents test moment 1 (T1), and the x-axis test moment 2 (T2). The dashed gray diagonal line represents the ideal scenario where performance on both test moments is identical. The linguistic levels are presented by different colors (red: words, blue: pseudowords, green: reversed words). Each individual dot is the percentage simultaneity perception for 1 SOA – as noted earlier, P16 was excluded for analysis.

Our main point of interest was how good the test-retest reliability of the TBW width was. We used Pearson correlation coefficients of the TBW width calculations between the two test moments to estimate this. Across linguistic levels we found a strong

correlation (R = 0.71, t = 6.97 , p < 0.001), which translated to each linguistic level (Fig 5.5: R_{words} = 0.73, t = 4.14 ,p < 0.001; $R_{pseudowords}$ = 0.64, t = 3.25 ,p = 0.006; $R_{rev.words}$ = 0.73, t = 4.17 ,p < 0.001)

FIGURE 5.5. Correlational plots per linguistic level.



Both axes show the TBW width (in ms, based on the 50% synchrony perception point of the sigmoid function). The y-axis represents test moment 1 (T1), and the x-axis test moment 2 (T2). The dashed gray diagonal line represents the ideal scenario where performance on both test moments is identical. The colored line is the regression line per linguistic level, and the shaded area is the 95% confidence interval. Each individual shape is a participant. Pearson correlation coefficients and p-values are plotted in the left top corner.

5.3.2.5. The TBW as an individual trait

A trait is *"a relatively stable, consistent, and enduring internal characteristic that is inferred from a pattern of behaviors in the individual"* (American Psychological Association, n.d.; Merriam-Webster, n.d.). To decide whether the TBW width is an individual trait there thus needs to be little intrasubject variability (within participants) and large intersubject variability (between participants).

We already discussed little intrasubject variability above (5.3.2.4). Additionally, we could observe from Fig 5.5, that individuals that have narrower/wider TBW's in test moment 1, also have narrower/wider TBWs in test moment 2 and that individuals

that have narrower/wider TBWs for one linguistic level, also show narrower/wider TBWs in another linguistic level.

Regarding intersubject variability we look at the range of TBW width across participants. For example for words we find TBWs ranging from 125 ms to 378 ms. Looking at the interquartile range we found a Q3-Q1 difference of 86 ms for words, 73 ms for pseudowords, and 110 ms for reversed words. In the order of average TBW widths between 250 and 300ms, interquartile ranges of 73 - 110 ms should be considered large.

We calculated the variance both within and between participants and found that the variance between participants (S^2 = 3503.18, sd = 59.19) was about five times as large as the variance within participants, between test moments (S^2 = 685.64, sd = 26.18). A likelihood ratio test showed that this difference was significant (χ^2 = 10.44; p = 0.001).

In summary, these results point to large intersubject variability and little intrasubject variability, identifying the TBW width as an individual trait.

5.3.3. Discussion

Currently in the literature, we have limited information about the reliability of the TBW shape within an individual. It was suggested that the PSS – the point where the highest synchrony perception was reached – is variable both between and within individuals (Recio et al., 2019). However, this might have been more of a task effect than a test-retest reliability effect (van Eijck et al., 2008; Vroomen & Stekelenburg, 2011). There are contraindications from an in-lab versus online comparison of the AV flash-beep illusion showing a reliable PSS (R = 0.75; Flanagan, Zumbrunn et al., 2023). This same group further showed a highly reliable (R = 0.93) TBW width within individuals at two test moments. Nonetheless, care in interpretation of the TBW as it relates to illusions is warranted as these might be driven more by the nature of the task rather than integration that would occur for naturalistic speech signals.

Comparing test moment 1 and 2 in our tasks, we reported a main effect of response time. Participants were on average 492 ms faster to respond in test moment 2 for the same stimuli. This increase in response time was not significantly different for any of the linguistic levels. Furthermore, participants who were slower for one linguistic level were also slower on the other levels, and vice versa. These findings were not surprising as participants knew what to expect, and were more familiar with the different types of stimuli in the repeat testing phase.

Regarding the test-retest reliability of the TBW, we reported no main effect of test moment on TBW width. Although there is sufficient evidence that synchrony perception can be trained, narrowing the TBW (McGovern et al., 2022; Powers et al., 2009, Zerr et al., 2019) even by exposure to unisensory stimuli (Stevenson et al., 2013), simple passive repetition of the task was not sufficient to narrow the TBW. Powers and colleagues (2009) indeed showed no difference in visual-leading TBW width between their baseline condition and pretest training on day 1, or between any of the test moments of their passive exposure experiment. If anything, there was a trend for shifting the full TBW (auditory- + visual-leading) to the right, mainly indicating a shift in PSS.

With regards to the slope and amplitude asymptote between test moments, we showed a trend towards a steeper slope and a larger amplitude in test moment 2 for all linguistic levels. Post-hoc analysis showed that the change in asymptote amplitude was not significant for words; however, this might be explained by the already large amplitude for these stimuli in test moment 1.

In sum, we reported replicable TBW width measures between test moment 1 and 2 both on a group and individual level. Nonetheless, the response duration, the slope and asymptote amplitude improved between test moment 1 and 2, but the findings and order of effects were similar to what we reported in our larger cohort (Section 5.2). The larger amplitude of the sigmoid could be explained by increased familiarity

with the rather unfamiliar stimuli: pseudowords and reversed speech, leading to increased certainty of (a)synchrony perception. In the context of causal inference (Körding et al., 2007) and integration of AV stimuli, it could be argued that the decision process can occur more smoothly as higher familiarity with these speech stimuli might lead to increased expectancy that the auditory and visual stimulus come from the same event.

Furthermore, we reported both high intersubject variability (interquartile ranges: 73 - 110 ms), and strong within-subject correlations (R = 0.71) of the TBW width. This combination confirms that the TBW width can be viewed as an individual trait. High intersubject variability of the TBW width is consistent with earlier findings (Conrey and Pisoni, 2006; Stevenson et al., 2012) and strong within-subject correlations have been reported more recently in Flanagan, Zumbrunn and colleagues (2023). We also showed that individuals with wide TBWs in one linguistic category were more likely to have wide TBWs in another linguistic category. Therefore, although there were significant differences in TBW width caused by linguistic complexity of the stimuli, this effect was similar for the participants.

In sum, the current work contributes to the literature

(1) that an AV temporal asynchrony measure can successfully be completed remotely for speech stimuli,
(2) that the linguistic differences reported for the TBW width (Section 5.2) are stable across test moments, however the amplitude of the sigmoid can improve when the task is repeated due to task familiarity,
(3) that ratings of synchrony perception are highly reliable on an individual level (R=0.93),
(4) that the TBW width of two test moments strongly correlates on an individual level (R=0.73),

and

(5) that the TBW width measure is an individual characteristic showing from high intersubject and low intrasubject variability.

5.4. The contribution of linguistic complexity and temporal synchrony to loudness perception of AV speech

The temporal coherence of AV stimuli has garnered considerable attention in the exploration of AV speech perception. Temporal information plays a role in directing the listener where and when to focus (Bernstein et al., 2004), in AV object formation (Bizley et al., 2016), and potentially in loudness enhancement of the auditory signal (Gillmeister & Eimer, 2007). However, loudness perception in the context of AV integration is not well understood. It has been reported that AV signals result in loudness enhancement compared to auditory-only signals (Odgaard et al., 2004). The change in loudness perception reported by Odgaard and colleagues was very robust to task manipulations. This, in combination with the fact that enhancement was reported of a feature that was not experimentally manipulated (i.e., loudness), suggests early-stage sensory integration. Furthermore, Gillmeister and Eimer (2007) suggested that loudness enhancement caused by multisensory stimulus presentation in the auditory and somatosensory domain is dependent on the synchrony of both signals. However, none of these studies used speech stimuli nor is it tested in the presence of competing noise. Therefore, we were interested in how this potential loudness enhancement phenomena translates to synchronous and asynchronous AV speech stimuli, whether this is influenced by linguistic complexity and whether there is any relationship to other measures of AV integration like the TBW width.

5.4.1. Methods

5.4.1.1. Participants

All 46 participants from Task 1 were invited to take part in this remote loudness perception task. Among them, 27 participants consented to participate (Table 5.1).

The average age of the participants was 28.3 years (Q2 = 27, sd = 5.91, min = 21, max = 40), with 4 identifying as male and 23 as female. All participants completed the task within two months of the initial task.

5.4.1.2. Loudness measure

For this loudness assessment, we selected the same three linguistic levels as those in Task 2 (Section 5.3): words, pseudowords, and reversed words. We reduced the SOAs from 8 to 5 levels: 0 ms, 50 ms, 150 ms, 300 ms and 500 ms, but covering the same range as before. The different number of stimuli were also reduced (from 20 to 15), to keep the total number of trials reasonable for one test session. These stimuli were a subselection of the original 20 stimuli, where we had 5 stimuli of each talker, and again 3 talkers total per linguistic level. The number syllables per word category were similarly balanced out across linguistic levels. More about the source of the recordings, the video set up, the stimuli and noise creation, can be found in the general method section (Section 5.1.5).

In order to obtain a wider range of loudness ratings, we introduced sufficient variability in the intensity of the stimuli. Consequently, all stimuli (15 per linguistic level) were presented at each of the 5 SOAs at 3 different TMRs (-3 dB, -6 dB and -9 dB). The intensity of the babble masker was kept constant for all three levels, but the intensity of the stimulus signal was varied. These TMRs, set in increments of 3 dB, were set up so that we have an objective benchmark of subjective loudness rating with objective TMR changes. Additionally, we included 15 trials at TMR = -21 dB and SOA = 0 ms, for each linguistic level. This control condition served as a reference loudness estimate, since it should be close to undetectable for all participants.

Altogether, the task consisted of 720 trials, organized into 9 blocks, each containing 80 trials. These blocks were evenly distributed across the linguistic levels, with 3 blocks per level. The TMRs and SOAs were randomly distributed within these blocks. Additionally, the sequence of these 9 blocks was randomly determined.

Each trial consisted of the presentation of a fixation cross (500 ms), followed by an AV stimulus presentation (4000 ms) and a participant response. The participant response element was different compared to Section 5.2 and 5.3. Under the video presentation the participant was shown the question "*How loud did you perceive the talker of the video?"* The participant was asked to use a slider, positioned beneath the question, to assess the loudness of the talker in the video on a scale ranging from 0 to 100 (with unit increments). As a result we were able to obtain 15 loudness ratings (between 0 and 100) per TMR, SOA, and linguistic level.

Participants were instructed to adjust the intensity of their computers to match the percentage indicated in Task 1. However, determining loudness ratings, particularly in noisy environments, is highly subjective (Berglund & Preis, 1997). To ensure that participants understood the expected range of loudness ratings, we implemented a training phase. During the training, participants were first presented with a sentence at a TMR of -3 dB (SOA = 0ms) and were informed that this represented the maximum loudness level of the task. Accordingly, for such stimuli, we provided feedback that trials like these should be rated in the upper quartile of the slider (75-100%). Subsequently, participants were presented with a stimulus at a TMR of -21 dB (SOA = 0ms), signifying the quietest level of the task. Participants were instructed that stimuli like this should be rated between 0-25%. These examples were followed by 12 practice trials, where participants were randomly presented with 3 sentences per TMR (-3 dB, -6 dB, -9 dB, -21 dB) and were asked to use the slider to rate the loudness of the talker in the video. Following each video, feedback was provided regarding the expected loudness range for their response in quartiles (0-25%, 25-50%, 50-75%, 75-100%).

5.4.1.3. Data analysis

Outliers

To define outliers for this task we calculated z-scores based on the mean and standard deviation of the loudness scores, per linguistic level, SOA and TMR. We then collected a summary of the outlier conditions (z-score > |2|) for each participant. Individuals with more than 5 outlier scores would be considered general outliers and removed from the dataframe.

Response times

Although we did not ask participants to work "as fast as they could" or under any time pressure, we were interested to see whether response times were in any way influenced by the linguistic level, TMR, or SOA of the trials. Therefore we used a linear mixed effects model (Model 5.9: Response time ~ linguistic level * TMR * SOA + (1|participant)), with contrast sum coding used for the TMR and linguistic level, to compare each level to the overall mean performance. Participant was included as a random effect.

Loudness rating by linguistic level

As we had no hypothesis regarding the influence of different levels of linguistic complexity on the loudness ratings, we created a summary table (Table 5.5) for each linguistic level (words, pseudowords, reversed words) and TMR (-21 dB, -9dB, -6dB, -3dB) at a SOA of 0ms. This would give us an idea whether the linguistic category itself (when synchronously presented) would lead to a difference in loudness perception.

We used a linear mixed effects model (Model 5.10: Loudness rating ~linguistic level * TMR +(1|participant)) for SOA = 0ms, to interpret differences in loudness ratings for the three different linguistic levels and 4 TMRs. We used -6dB as a reference TMR,

since this was also the level we used in the other tasks, and words as reference linguistic level. Participant was included as a random effect.

Loudness rating by temporal asynchrony (and linguistic level)

Table 5.6 represents summary statistics (mean, median, standard deviation) of the loudness ratings at 0 and 500 ms for each TMR and linguistic level. This allowed us to directly interpret the loudness rating difference between synchronous and completely asynchronous AV speech stimuli. Henceforth, we will refer to this as the loudness rating drop.

We used a linear mixed effects model (Model 5.11a: Loudness rating ~ (SOA+TMR) * linguistic level + (1|participant)), to interpret the effects of different SOAs on the loudness rating, for each linguistic level (words, pseudowords, reversed words) and each TMR (-21 dB, -9 dB, -6 dB, and -3 dB). We used TMR = -6 dB, SOA= 0 ms, and linguistic level = words as reference variables (treatment coding). Participant was included as a random effect. We initially fitted a three-way interaction: SOA*TMR*linguistic level, but as this three-way interaction unnecessarily further complicates interpretation and no significant differences were found between this model and the model with no interaction effect for SOA*TMR, we excluded this interaction in our final model: Model 5.11a.

We further inspected these changes in loudness rating by SOA on an individual level. We compared (Fig 5.7) the loudness rating for synchronous AV stimuli (SOA = 0ms), to the loudness ratings of all other SOAs, split up by linguistic level. We averaged across TMRs, as there was no interaction effect between SOA and TMR (Model 5.11a). We then calculated the Pearson correlation coefficients and the slope of these correlations, to measure how reliable loudness ratings were between SOAs, within individuals. Furthermore, we calculated the shift in perceived loudness rating from SOA = 0 ms to the other SOAs (50 ms, 150 ms, 300ms, and 500 ms). To do this we measured how far away the linear regression was (at the 50% point) , from the

loudness rating at SOA = 0 ms: Δ 50%. This allowed us to interpret whether a shift, or drop, in loudness rating by SOA was similar for all participants.

Mapping of perceived loudness to perceived TMR

By omitting our reference TMR (-21 dB), we retained 3 TMRs with equal steps (+3dB). To transform our subjective perceptual measure (i.e., loudness rating in %), into an objectively interpretable score (TMR change in dB), we first created individual slopes per participant that represented loudness ratings for synchronous AV stimuli (SOA = 0 ms) by TMR, per linguistic level (Fig 5.8). The stepwise increase in loudness rating for each step of the TMR-scale (i.e., slope), then provided us with a mapping (per individual and as a group) for what percentage of perceived loudness matched with 1dB in TMR change for each individual.

Second, we calculated the change in perceived loudness caused by asynchrony (i.e., by SOA) for each individual. To do so we subtracted the loudness rating of asynchronous AV stimuli (SOA = 50 ms, 150 ms, 300 ms, 500 ms), from the loudness rating for synchronous AV stimuli (SOA = 0 ms). We did this for each linguistic level and each TMR at which we measured loudness ratings. By now dividing the change in perceived loudness for each individual (and SOA) by that individual's slope (see previous paragraph: loudness rating AV 0ms * TMR), we obtained a change in perceived TMR (in dB). This allowed us to map, for each individual and on a group level, what the perceived TMR drop (dB) was for each introduced SOA, split up by (tested) TMR and linguistic level (Fig 5.9).

We reran Model 5.11a, but changed loudness rating into the perceived TMR change (dB), and used TMR (of -3 dB, -6dB, and -9dB) as a continuous variable: Model 5.11b: Perceived TMR change ~ (SOA+TMR) * linguistic level + (1|participant)). Furthermore we now used SOA = 50 ms as a reference (as SOA = 0 ms was 0 for everyone), the word level stayed the reference and the participant was again included as random intercept.

Loudness rating and TBW width

As loudness enhancement experience caused by synchronous stimuli (Odgaard et al., 2004), and the TBW width both are expected to represent early-integration mechanisms of AV integration, it would not be surprising to find a correlation between these two measures of interest on an individual level.

Therefore we correlated (Pearson correlation coefficient) for each linguistic level (i.e., words, pseudowords and reversed words) the TBW width values of Task 1 (Section 5.2) with the perceived TMR drop averaged across TMRs, (as there was no effect of TMR on the perceived TMR drop), for each linguistic level and the absolute loudness rating at SOA = 0ms, and SOA = 500 ms.

5.4.2. Results

5.4.2.1. Outliers

Based on the z-scores of the mean and standard deviation of the loudness scores we identified 9/27 participant who had at least a z-score > |2| in 1 condition. However, for 7/9 participants this was only 1-3 conditions. There were two participants with a noticeably high amount of outliers: 9 and 11 conditions. These two participants were thus subsequently removed for further data-analysis.

5.4.2.2. Response times

The average response time was 4735 ms. No main or interaction effect was reported from the linear mixed effects model (Model 5.9). Visualization of the data suggested a trend for slower response times for larger stimulus asynchronies (i.e., longer SOAs), however this effect was not significant.

5.4.2.3. Loudness rating by linguistic level

Table 5.5 shows loudness ratings by linguistic level (words, pseudowords, reversed words) for each TMR. We selected only synchronous stimulus presentations (SOA = 0 ms) for this analysis to gain a better idea of the intrinsic differences of linguistic complexity on loudness ratings.

We observed a low loudness ratings for TMR -21dB for all linguistic levels (μ = 14.95%, Q2 = 11%). The other 3 TMRs (-3 dB, -6 dB, -9 dB) were more spread out, with mean loudness ratings at $\mu_{TMR=-9dB}$ = 33.77%, $\mu_{TMR=-6dB}$ = 44.59%, and $\mu_{TMR=-3dB}$ = 61.2%. When comparing these 3 TMRs by linguistic level, words and pseudowords similarly increased with increasing TMR (Table 5.5). For reversed words the ratings at TMR = -9 dB and -6 dB stayed closer to the TMR = -21 dB rating, and mean loudness reports for all TMRs were lower. Indeed, the difference in mean loudness ratings between TMR = -9dB and TMR = -3 dB was 28.67% for words, 26.00% for pseudowords, but only 19.74% for reversed words.

A linear mixed effects model (Model 5.10) showed for these synchronous AV stimuli a main effect of TMR for all TMRs compared to TMR = -6dB (t_{-21dB} = -24.14, $p < 0.001$; t_{-9dB} = -9.25, $p < 0.001$; t_{-3dB} = -12.40, $p < 0.001$). Furthermore, across TMRs, synchronous loudness ratings for words were significantly different from both pseudowords (t = -2.70, $p < 0.01$) and reversed words (t = -10.30, $p < 0.001$). Post hoc analysis also showed a main effect between pseudowords and reversed words (t = -7.57, $p < 0.001$). There was a significant interaction effect ($p<0.001$) between TMR and linguistic level, as the loudness ratings for reversed words were significantly lower for all TMRs and more compressed between TMRs compared to words and pseudowords, especially for the highest TMR levels (-6 dB and -3 dB).

TABLE 5.5. Loudness rating for synchronous AV stimuli.

Linguistic Level	TMR (dB)	Peak Loudness rating (%)	Mean	Q2 (Median)	Inter-quartile range	sd
Words	***-21***	8.41	15.20	11	15	13.38
	-9	21.50	33.60	30	30	22.14
	-6	47.00	46.46	46	37	24.52
	-3	74.60	62.27	66	37	25.10
Pseudowords	***-21***	8.61	16.16	11	15	13.29
	-9	18.20	30.15	25	28	20.51
	-6	43.30	42.16	40	37	23.59
	-3	78.80	56.15	57	39	25.61
Reversed words	***-21***	9.20	13.36	11	11	11.92
	-9	12.90	27.05	21	26	20.27
	-6	16.00	34.93	30	36	23.14
	-3	47.70	46.79	46	43	25.97

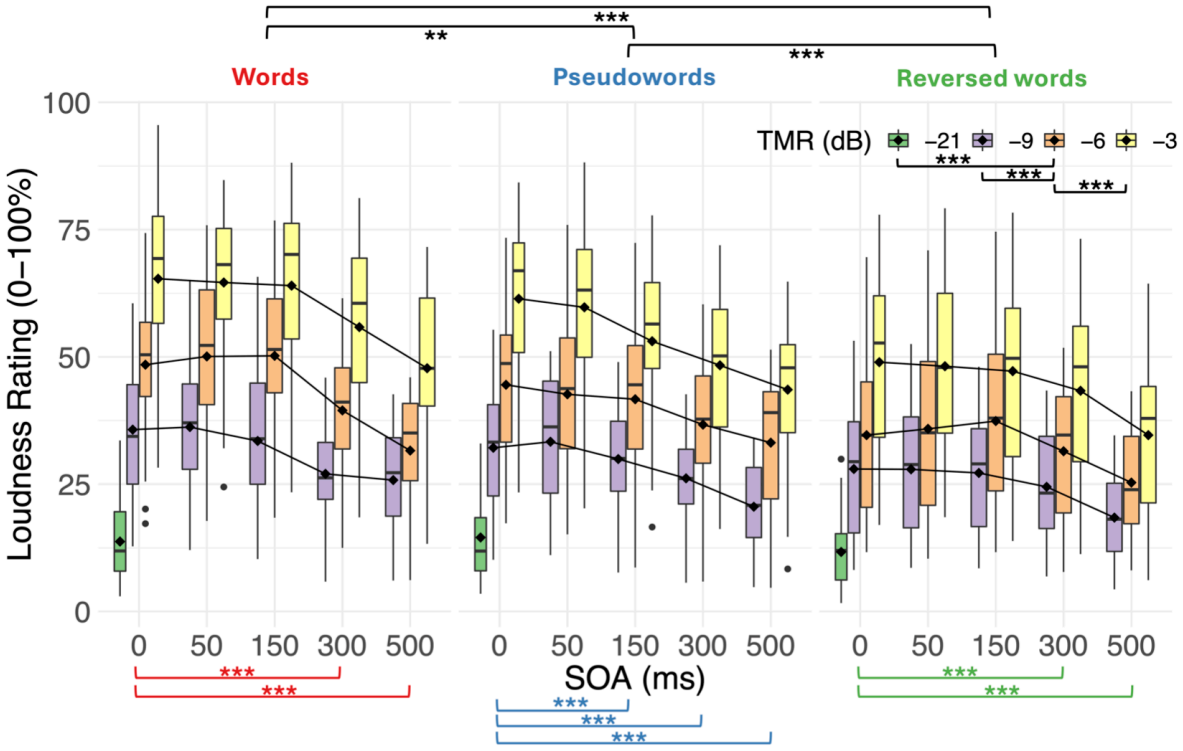
Peak Loudness rating = peak % derived from density function of rating responses. Mean, median, interquartile range and standard deviation of the loudness ratings by linguistic level (words, pseudowords, reversed words) and TMR (-21 dB, -9dB, -6 dB, -3 dB).

5.4.2.4. Loudness rating by temporal asynchrony (and linguistic level)

We investigated the loudness ratings of each participant by SOA, TMR, and linguistic level (Fig 5.6). We observed from Fig 5.6, and confirmed with our linear mixed effects model (Model 5.11a), that there was a main effect of SOA, starting at SOA = 300 ms. Across linguistic levels and TMRs, loudness ratings were not significantly different for synchronous stimuli (SOA = 0 ms), or asynchronous stimuli up to SOA = 150 ms. Beyond this point loudness ratings dropped 9.05% ($t = -9.4$, $p < 0.001$) for SOA = 300ms and 15.00% ($t = -15.55$, $p < 0.001$) for SOA = 500 ms. Indeed, when observing Fig. 5.6 we saw a relatively stable loudness rating up to 150 ms, followed by a significant drop. This is true for all linguistic levels and all TMRs.

Furthermore, there was a main effect of TMR (across SOAs and linguistic levels), with TMR = -21 dB rated -35.97% more quiet than TMR = -6 dB ($t = -23.85$, $p < 0.001$), TMR = -9 dB rated 12.88% more quiet ($t = -17.12$, $p < 0.001$), and TMR = -3d B rated 15.86% louder than TMR = -6 dB ($t = 21.36$, $p < 0.001$). Finally, there was a main effect of linguistic level (across TMR and SOAs), with pseudowords rated 3.52% more quiet than words ($t = -3.10$, $p<0.01$), reversed words 13.56% more quiet than words ($t = -11.89$, $p<0.001$), and 10.04% more quiet than pseudowords ($t = -8.79$, $p < 0.001$).

FIGURE 5.6. Boxplots of loudness ratings of AV speech stimuli for each SOA, plotted by linguistic level and TMR.



SOA = 0 ms, 50 ms, 150 ms, 300 ms ,500 ms. Linguistic level = words, pseudowords, reversed words. TMR = -21 dB, -9dB, -6 dB, -3 dB. The mean is represented as a black square, the median as a black horizontal line. Black dots are significant outliers. Significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

TABLE 5.6. Summary of loudness ratings at SOA = 0 ms and 500 ms per linguistic level and TMR.

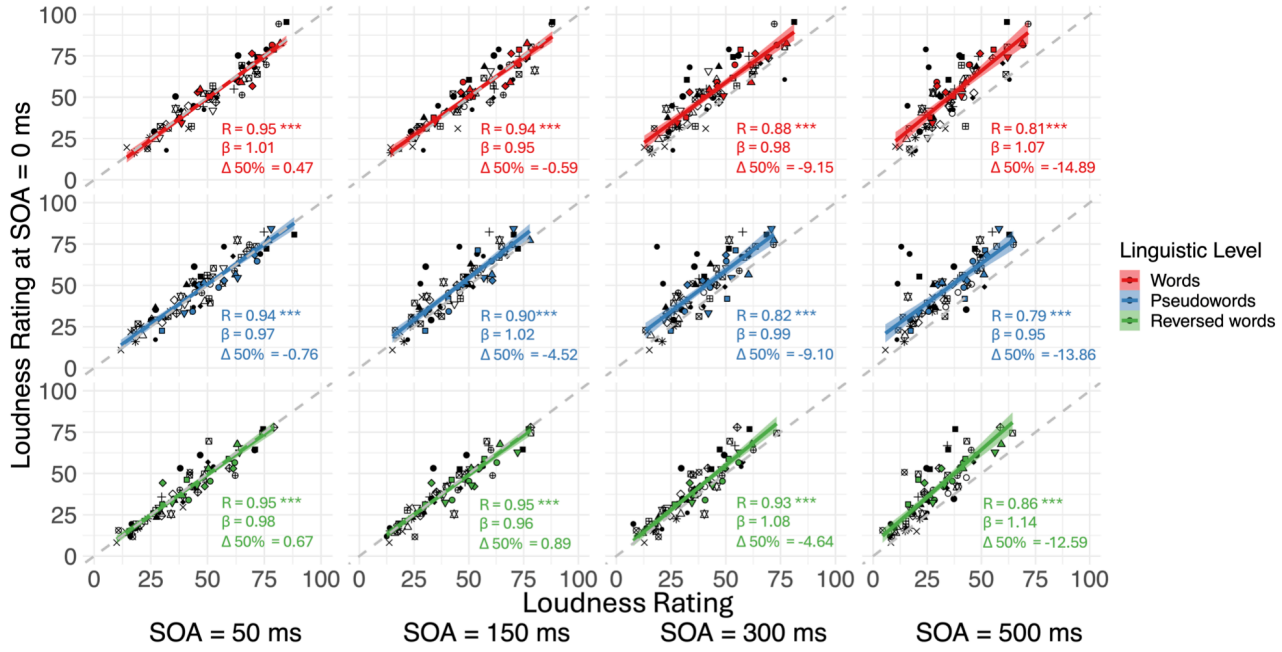
			Loudness rating (%)		
Linguistic Level	**TMR (dB)**	**SOA**	**Mean**	**Median**	**sd**
Words	**-3**	0	67.15	71.00	25.13
		500	48.64	46.00	25.58
	-6	0	48.81	50.00	23.91
		500	34.74	32.00	22.09
	-9	0	37.55	34.00	23.03
		500	25.77	23.00	17.35
Pseudowords	**-3**	0	65.79	69.50	22.16
		500	46.27	42.50	25.49
	-6	0	46.13	44.00	23.99
		500	34.53	32.00	22.28
	-9	0	33.80	32.00	20.67
		500	21.72	17.00	16.70
Reversed words	**-3**	0	51.00	51.00	24.25
		500	36.85	31.00	24.91
	-6	0	36.08	33.00	22.89
		500	27.23	19.00	21.97
	-9	0	30.67	24.00	22.13
		500	19.62	15.00	15.26

Linguistic level: words, pseudowords, reversed words. TMR: -3 dB, -6 dB, -9 dB. Dark gray: mean loudness rating drop comparison between -3 dB and -6dB, light gray mean loudness rating drop comparison between -6 dB and -9 dB.

The interaction effects showed a significant earlier drop for pseudowords compared to the other two linguistic levels, as SOA= 150 ms was already significantly lower ($t = -4.43$, $p < 0.001$) than SOA = 0ms. For reversed words there was a significantly smaller drop for 300 ms ($t = 3.98$, $p < 0.001$) and 500 ms ($t = 2.82$, $p < 0.01$) compared

to SOA = 0 ms, than for words (and pseudowords). When interpreting the interaction of TMR * linguistic level we noted significant interactions between reversed words and both words/pseudowords for all TMRs, as the reversed words were rated significantly more quiet, and the differences between TMRs were more compressed.

FIGURE 5.7 Loudness rating for synchronous AV stimuli by loudness rating for asynchronous AV stimuli, per linguistic level and individual.



Synchronous AV loudness rating (SOA = 0ms) by asynchronous AV loudness rating (SOA = 50 ms, 150 ms, 300 ms, 500 ms), for each linguistic level (red = words, blue = pseudowords, green = reversed words). Every shape represents an individual. As we averaged across TMRs there are 3 data points per individual per plot. The right bottom of each plot shows the Pearson correlations coefficient (Significance $p < 0.001$ (***)), slope (β) and shift in loudness rating caused by stimulus asynchrony (Δ 50%), which represents the horizontal shift from the diagonal (gray dashed line).

A noteworthy observation (Fig 5.6) is that for all linguistic levels we found that the mean loudness rating at 500 ms was very similar to the mean loudness rating at 0 ms of one TMR level down (Table 5.6). This suggests that the loudness rating drop between SOA = 0 ms and 500 ms was in the order of 3dB as this is the difference between TMRs.

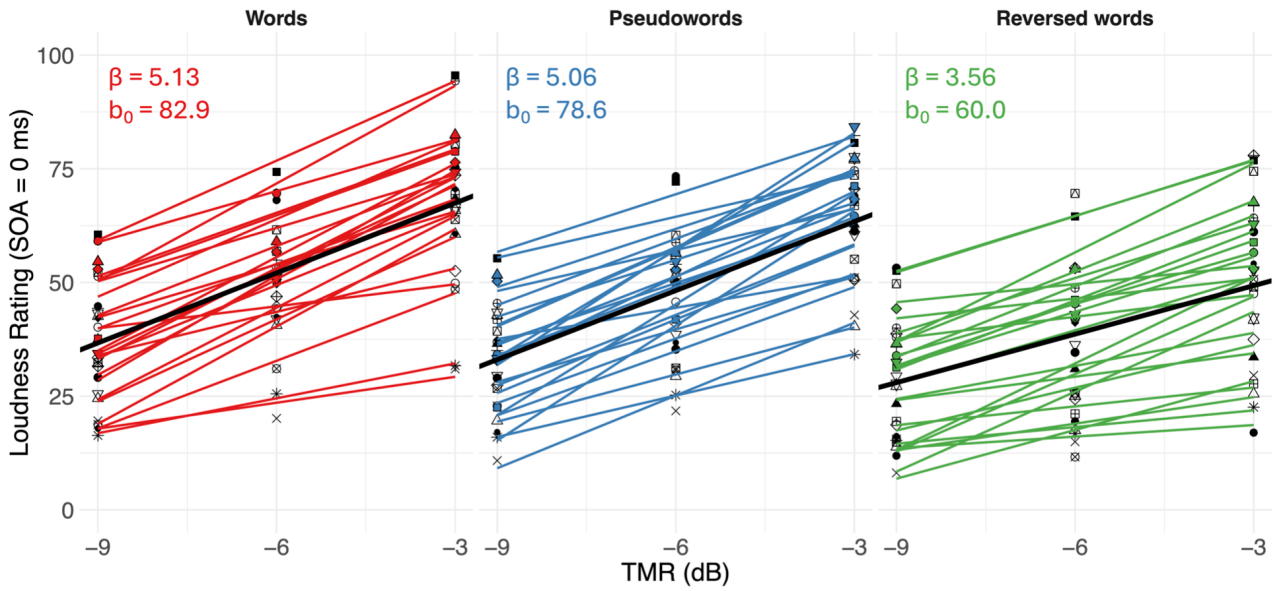
On an individual level, the loudness rating was strongly correlated ($R = 0.79 - 0.95$, $p < 0.001$) between the synchronous AV stimuli (SOA = 0 ms) and different presented asynchronous AV stimuli (SOAs = 50 ms ,150 ms, 300 ms, 500 ms), at all linguistic levels. This means that individuals were highly consistent in their loudness ratings. Furthermore, the slope was close to 1 for all presented plots (Fig 5.7, $\beta = 0.95 - 1.14$), which indicates that individuals with high loudness ratings for synchronous AV stimuli also had high loudness ratings for asynchronous AV stimuli, and vice versa. As the ratings on an individual level were so strongly correlated, it was not surprising that the perceived loudness drop we observed on a group level (Fig 5.6) also showed on an individual level. Fig 5.7 represents an increasing shift of the regression line to the left, parallel to the diagonal (as $\beta = 0.95 - 1.14$), by increasing SOA. For the most asynchronous AV stimuli (SOA = 500 ms), this resulted in a significant loudness rating shift (-12.59% to -14.89%) that was of the same order across linguistic levels. The only difference observed between linguistic levels was SOA = 150 ms for pseudowords. This is similar to what we reported in the linear mixed model for the loudness ratings (Model 5.11a): a significant drop in perceived loudness (-4.52%) for pseudowords at 150 ms, but not for words (-0.59%) or reversed words (0.89%).

5.4.2.5. Mapping of perceived loudness to perceived TMR

First we calculated slopes between loudness ratings for synchronous AV stimuli and TMR, by linguistic level (Fig 5.8). As we only included synchronous AV stimuli, we can make conclusions on a group level similar to Model 5.10. There was a significant increase in loudness rating with increasing TMR (i.e., positive slope) for each linguistic level. Furthermore, the slope was similar for words ($\beta = 5.13$) and pseudowords ($\beta = 5.06$), but shallower for reversed words ($\beta = 3.56$), indicating an increase in TMR resulted in less of a change in loudness rating for reversed words. Furthermore, by comparing the intercept (i.e., projected loudness rating at TMR = 0dB), we observed that words ($b_0 = 82.9\%$) were rated louder than pseudowords ($b_0 =$

78.6%), and both were rated louder than reversed words (b_0 = 60.0%). Furthermore this allowed us to map the perceived loudness rating to a perceived change in TMR (in dB). A slope of β = 5.13 for words can be interpreted as that on average an increase in 1dB of the TMR resulted in a perceived loudness change of 5.18%. Similarly, an increase of 1dB of the TMR for reversed words only resulted in a perceived loudness change of 3.56%.

FIGURE 5.8 Slopes of loudness rating for synchronous AV stimuli by TMR.



Individual slopes are represented as colored lines, whereas the group slope is represented by a black line. The different shapes represent the measured loudness ratings for each subject. Red = Words, blue = Pseudowords, green = Reversed words. The slope (β) and intercept (b_0) are presented in the left top of each plot. y-axis: Perceived loudness rating for synchronous AV stimuli (SOA = 0 ms), x-axis TMR (-3 dB, -6 dB, -9 dB).

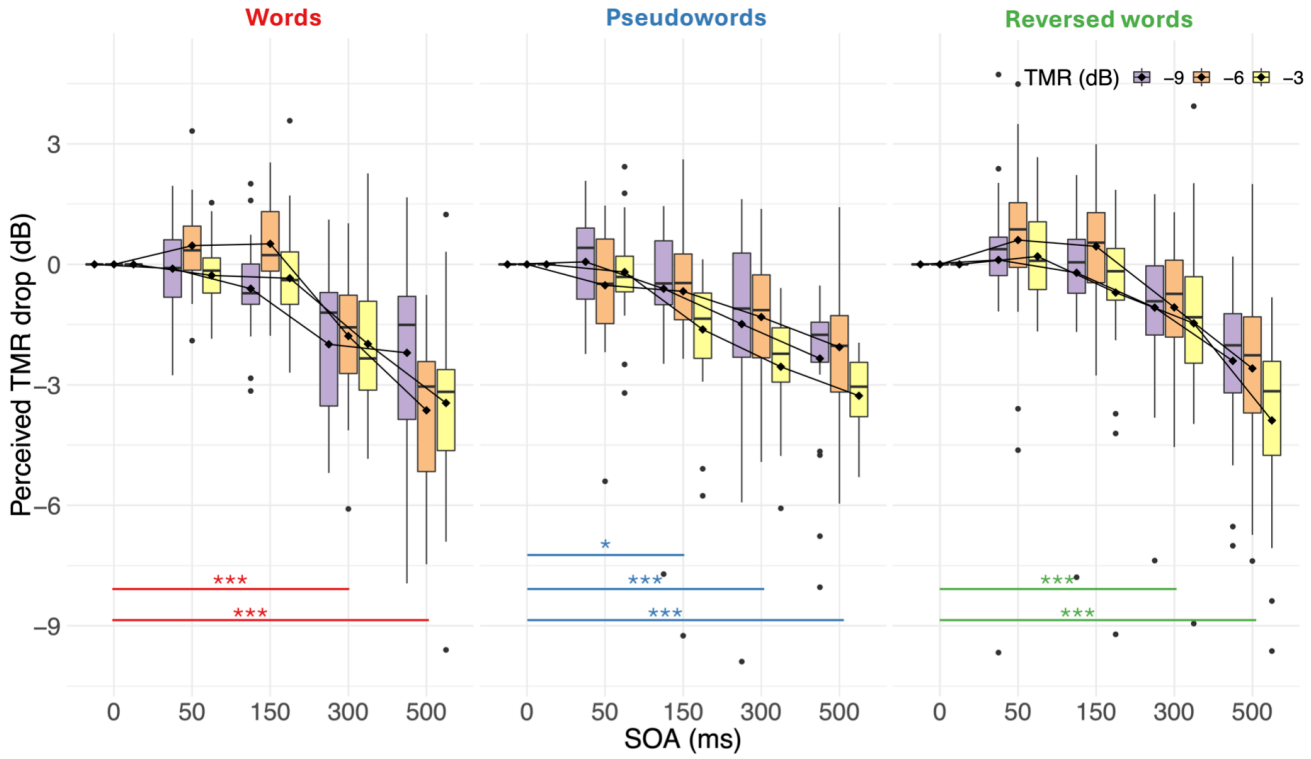
Looking at the individual slopes (Fig 5.8) we observed an increase in loudness rating by TMR for everyone, although the steepness of the slope varies. Individuals also used different 'starting points', or percentages in their ratings, and for each TMR and linguistic level loudness ratings covered about 50% of the total scale (0-100 %) based on differences between individual listeners.

We then transformed the shift in loudness perception between synchronous and completely asynchronous AV stimuli (Δ 50% for SOA = 500 ms in Fig 5.7) to a

change in perceived TMR (dB), by using the slopes calculated for each individual, and on a group level (for synchronous AV stimuli in function of TMR - Fig 5.8). This resulted in a mapping of perceived TMR by change in SOA for each linguistic level and TMR, presented by the stimuli (Fig 5.9). Changing the response variable from loudness rating to perceived TMR change in our model (Model 11b) resulted in a main effect of SOA for 300 ms ($t = -5.82$, $p < 0.001$) and 500 ms ($t = -9.67$, $p < 0.001$) asynchrony, but not for 50 ms and 150 ms. Specifically for words this meant a TMR drop of -1.93 dB at SOA = 300ms and -3.20 dB at SOA = 500 ms. By taking the loudness rating for synchronous AV stimuli (SOA = 0 ms) into account per individual, there was no main effect of linguistic level observed, in contrast to Model 5.11a. Furthermore, there was no main effect of TMR. The only interaction effect observed was also noted in Model 5.11a. The TMR drop for pseudowords at SOA = 150 ms was significantly different from 0 ($t = -2.26$, $p = 0.02$). For pseudowords this then resulted in a drop of -0.75 dB at 150 ms, -1.90 dB at 300 ms and -2.92 dB at 500 ms. For reversed words the TMR mapping resulted in a -1.58 dB change at 300 ms and -3.8 dB at 500 ms.

Thus, by taking into account individual differences for loudness ratings for synchronous AV stimuli, and differences between linguistic levels for synchronous AV stimuli, we reported a main drop of 3dB for all TMRs and all linguistic levels at 500 ms stimulus asynchrony.

FIGURE 5.9 Perceived TMR drop by SOA, per linguistic level and presented TMR.



The perceived TMR drop (dB), based on the slope of the synchronous loudness rating and the TMR is presented in relation to all asynchronous SOAs (50 ms, 150 ms, 300 ms, 500 ms). This is broken up by TMR of the stimuli (-3 dB, -6dB, -9 dB) and linguistic level (words, pseudowords and reversed words). The mean is represented as a black diamond, the median as a black horizontal line. Black dots are outliers. Significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

5.4.2.6. Loudness rating and TBW width

Similar to what we observed for the TBW in Section 5.2 and 5.3, there was a significant difference between words and pseudowords reported in the perceived TMR drop. For the TBW, we observed narrower TBWs for pseudowords than for words. For the perceived TMR drop we observed an earlier drop in perceived loudness for pseudowords than for words.

Surprisingly, by correlating the TBW to the perceived TMR drop on an individual level, we did not find a significant correlation for the TBW width, the slope, or the asymptote amplitude.

The only significant correlation between the loudness ratings and the TBW measures on an individual level was for reversed words. More specifically the amplitude asymptote of the TBW (difference in synchrony perception between 0 and 500 ms) was significantly positively correlated to the loudness rating at 500 ms ($R = 0.44$, $p = 0.02$). This moderate correlation suggests that individuals who rated the completely asynchronous reversed AV words as louder, had larger amplitude asymptotes for the TBW of the reversed AV words.

5.4.3. Discussion

The findings from this task contribute to the relatively unexplored domain of loudness perception in integration of AV speech. Odgaard and colleagues (2004) reported a significant increase in loudness perception when comparing AV to auditory-only stimuli. Gillmeister and Eimer (2007) found that differences in loudness perception for multisensory stimuli (i.e., auditory and somatosensory) were also triggered by temporal asynchrony, with synchronous stimuli being perceived significantly louder. Hence, our curiosity was piqued to explore how this potential loudness enhancement feature translates to synchronous and asynchronous AV speech stimuli, whether this is influenced by linguistic complexity and whether there is any relationship to other measures of AV binding like the TBW width.

To address these questions, we had participants rate the loudness perception of a talker in a video at different TMRs (-21 dB, -9 dB, -6 dB, and -3 dB), different levels of linguistic complexity (words, pseudowords and reversed words) and at different temporal presentations of the audio and video (SOAs ranging from 0 ms to 500 ms, visual-leading).

Initially, we investigated whether there existed an inherent distinction in loudness ratings based on linguistic complexity. To examine this, we exclusively considered the loudness ratings for synchronous AV stimuli presentations. We observed that

across all three linguistic levels, the intended differences were well-represented through the provision of varied TMRs, with the hierarchy -21 dB < -9 dB < -6 dB < -3 dB maintained. Across TMRs, all linguistic levels were significantly different from each other. Specifically, words garnered significantly higher loudness ratings compared to pseudowords, while pseudowords, in turn, were rated louder than reversed words. The relationship between TMR and linguistic level was not different for words and pseudowords, but it was for words/pseudowords and reversed words, as loudness ratings for reversed words were significantly lower for all TMRs and more compressed between TMRs, especially for the highest TMR levels (-6 dB and -3 dB).

These differences in loudness perception based on linguistic complexity, even for synchronous AV stimuli, were rather surprising. Fucci and colleagues (1995) for example showed no difference in loudness ratings between auditory speech stimuli of Hindi and English, for native English speakers. English and Hindi have some phonemes in common, but these are rather limited. Therefore, it could be argued that loudness perception of an unknown language with limited phoneme overlap should fall somewhere between our linguistic categories pseudowords (i.e., identical phoneme-viseme connections but no lexical information) and reversed words (i.e., no common phoneme-viseme connections and no lexical information). Yet, they (Fucci et al., 1995) reported no differences in loudness ratings between the two languages, and we observed differences in both categories compared to words. As they did not measure loudness in background noise, we might be tapping into different mechanisms of loudness judgment. Furthermore the differences reported in reversed speech (i.e., lower and more compressed loudness ratings) could potentially be an unintended result of our stimulus manipulation. Loudness perception is dependent on the speed of the attack and decay characteristics of the speech envelope (Stecker & Hafter, 2000), and although we need to further analyze our sample of reversed words used in this task to quantify this, flipping the attack

and decay characteristics in reversed speech could lead to differences in loudness perception. However, this was not the case for our pseudowords. The only differences between words and pseudowords were linguistic complexity, and potentially familiarity with the stimulus. As we reported in Section 5.3 that the slope and asymptote amplitude envelope of the temporal binding window significantly improved for pseudowords in the second test session, it is likely that this unfamiliarity effect of pseudowords already disappeared in a second test session. This makes linguistic complexity in itself most likely the driver for the reported differences.

Loudness perception is studied well in function of objective intensity, frequency cues, hearing loss and binaural hearing (Meunier et al., 2020), however we know little of experienced loudness perception of speech stimuli, and especially of differences in linguistic influences on these speech stimuli. Future work should therefore focus on perceived differences in loudness based on linguistic characteristics of the speech signal.

We further analyzed loudness perception as a function of temporal asynchrony. We found similarly to the findings of Gillmeister and Eimer (2007), but now in the AV modality (instead of auditory-somatosensory), that temporal asynchrony had a significant effect on loudness ratings. Across linguistic levels and TMRs, we noted that the loudness ratings were not different for synchronous stimuli (SOA = 0ms), or asynchronous stimuli up to SOA = 150 ms; however, beyond 150 ms loudness ratings dropped for all TMRs and all linguistic levels. The effect of linguistic complexity we reported for synchronous stimuli (words > pseudowords > reversed words) was still maintained when averaging across all SOAs, for both synchronous and asynchronous stimuli. However, there was also an interaction between linguistic level and SOAs. There was a significantly earlier drop in loudness rating for pseudowords than for words or reversed words, as it already showed a significant difference at 150 ms. And as we observed that the loudness ratings for reversed

words were more compressed, we also found a significant interaction effect for reversed words and SOA = 300 ms and 500 ms, showing that the drop from synchronous to asynchronous (300 - 500 ms) was more limited in reversed words.

The report that loudness perception is stable up to 150 ms is in the order of magnitude we need our perceptual system to be stable to perceive AV speech, as visual speech inherently precedes auditory speech (Chandrasekaran et al., 2009). Furthermore, although the 50% synchrony perception point of the TBW lies further away (> 150 ms), this is exactly where the TBW functions really start to drop. For example the 75% synchrony point of the TBW in our first task (Section 5.2) is reported at 173 ms for words and 132 ms for pseudowords. For reversed words the 75% point is already located at 56 ms, but this might have more to do with the fact that the upper asymptote for this category already lies at 79%. In line with the TBW width results for words and pseudowords, we also found that the perceived loudness drop for pseudowords occurred earlier than for words. However, further comparisons of the TBW and perceived loudness ratings on an individual level showed no direct relationship on an individual level.

How do we interpret this drop in loudness perception for asynchronous stimuli? We selected TMRs in steps of 3 dB with the intention to map the loudness scale to some objective changes in TMR. As we had no indication of the size, or even the presence of differences in loudness perception for (a)synchronous AV stimuli, the steps of TMR selection were based on the logarithmic nature of the decibel scale. First, steps of 3 dB go together with doubling the signal's power, and second, steps of 3 dB are noticeable for participants (Hassall & Zaveri, 1988). Interestingly, when mapping the perceived loudness rating of synchronous AV stimuli to these changes in TMR we were able to create a mapping of perceived TMRs in dB. When we then used these slopes of each individual to map it to the perceived drop in loudness perception we found on average a perceived TMR drop of +/- 3dB for the most asynchronous stimuli (SOA = 500 ms). This thus suggests that a synchronous presentation of AV

speech signal leads to a 3 dB enhancement in perceived TMR in comparison to completely asynchronous stimuli. Although tested in a very different and unimodal (i.e., auditory-only) context, Egan and colleagues (1961) showed that auditory detection with temporal uncertainty provides thresholds 2-3 dB lower than detection with temporal certainty. So temporal asynchrony and therefore higher temporal uncertainty might indeed provide effects in the order of 3 dB.

What can we derive from this 3 dB drop in perceived TMR through a loudness perception task between synchronous and significantly asynchronous AV speech stimuli? Does temporal synchrony improve perception, or does it improve the ability to filter out the masker stream? Cappelloni and colleagues (2023) studied semantic and temporal influences in AV binding and they compared their results to earlier work from the same lab (Fiscella et al., 2022), who used a slightly different paradigm, with a focus on phoneme-viseme connections and temporal information. Cappelloni and colleagues (2023) concluded that the outcome of both studies differed in increased hit rate (i.e., improved perception) or lowered false alarm rates (i.e., improved ability to filter out the masker), two separate mechanisms associated with changes in sensitivity to AV stimuli, triggered by two different types of incongruencies in their stimuli. Semantic information (i.e., talker identity), as used in Cappelloni et al. (2023), might have been a sufficiently salient cue to maintain selective attention, even without AV binding. As a result they found no benefit based on temporal synchrony, as the talker identity was sufficient to provide information about where and when to listen, resulting in a change in false alarm rate. It might be interesting to do similar stimulus manipulations in both a loudness perception task and a TBW measure, to further disentangle the mechanisms behind our findings.

Looking at it from another angle, this line of reasoning also fits with the *'proposed complementary role of canonical integration operations enabling context-dependent integration'* as suggested by van Atteveldt and colleagues (2014). When presented with visual cues, synchronized firings of groups of neurons in the primary auditory

cortex occur (Schroeder et al., 2008). The neural oscillations in the auditory cortex created by the visual cues shift the phase in a way that when the auditory input arrives the phase will be in a high state of excitability. This process of phase-resetting, in turn, leads to amplification of properties of the auditory signal that results in overall AV enhancement of the incoming signal. Therefore, the loudness effect for synchronous stimuli would be an enhancement effect. However, van Atteveldt and colleagues (2014) also formulate that if there is a second stimulus that differs on a certain dimension (e.g., timing, location, semantics), it will suppress the excitatory response of the first stimulus and therefore influence the normalization signal of the surrounding neurons (i.e., divisive normalization). This would then mean that the loudness of our asynchronous stimuli is actually suppressed. Most likely, the occurrence of phase-resetting and divisive normalization in this process complements each other, and our attenuated loudness experience in the order of 3 dB is a combination of both processes.

Finally, we correlated the size, and the slope of the experienced loudness drop (for the total SOA width), and within the initial frame (between 0 and 150 ms) to the width of the TBW (from Task 1, Section 5.2) on an individual level. We hypothesized that there would be a relationship as they both were expected to represent early-stage integration mechanisms (Odgaard et al., 2004). Furthermore, (1) the findings that the TBW width is an individual trait (Section 5.3.) and (2) the findings that the loudness perception drop occurred earlier for pseudowords as words seemed to corroborate this idea even more, as the TBW was significantly more narrow for pseudowords than words (Section 5.2 and 5.3). Yet, we could not find any significant correlation between any of these measures on an individual level, other than a positive correlation between the amplitude asymptote of the TBW (difference in synchrony perception between 0 and 500 ms) of reversed words with the absolute loudness rating at 500 ms. As we interpreted the asymptote amplitude as a "certainty decision" regarding the (a)synchronicity of the AV speech stimuli, we

could hypothesize that the loudness rating at 500 ms informed us about the certainty of (a)synchrony perception, however it is rather unlikely that this is what an "absolute" loudness rating represents.

One possible explanation for these remarkable similarities between the TBW and the loudness perception on a group level, but not on an individual level, could be that both measures represent different stages of AV integration. Both measures have been reported to represent early-stage integration (TBW: Wallace & Stevenson, 2014, loudness perception: Odgaard et al., 2004). Early-stage integration is argued to be unique from late-stage integration mechanisms in that it is independent of linguistic information (Fiscella et al., 2022), and that it can be measured by an orthogonal feature as AV binding enhances all features of an AV object (Bizley et al., 2006). Our loudness perception measure is especially interesting because loudness perception is an orthogonal feature to the features that bind the AV event in our stimuli. However, as stated by Lee and colleagues (2019), all AV binding (i.e., early-stage integration) is integration, yet not all AV integration (i.e., early- & late-stage integration) is also AV binding. So, although both measures show characteristics of AV binding, late-stage integration mechanisms, like language, might still play a significant role in the tasks assessed. Indeed, for both measures we find significant effects of linguistic complexity, and we also know that the width of the TBW is modulated by both top-down and bottom-up factors (Stevenson et al., 2014; Zhou et al., 2020). Therefore, on a group level these results might nicely match, but on an individual level individual linguistic knowledge, attention, and cognition might influence the results.

Future experiments could help further clarify this. For the TBW it is hard to decide whether temporal synchrony reports are triggered by the task, or if this actually represents AV object formation. However, for loudness perception we could better dissect this problem as loudness being an orthogonal feature, different from the features that are supposed to form the object. So, by for example, not only

manipulating temporal synchrony, but also semantic congruency we could further inspect the influences of early- and late-stage integration.

In sum:

(1) Manipulating temporal synchrony led to a change in loudness perception, resulting in a perceived +/- 3dB TMR drop from synchronous to completely asynchronous (500 ms) visual-leading AV stimuli.
(2) Different levels of linguistic complexity led to significantly different loudness ratings, with the highest linguistic complexity resulting in the loudest perceived signal. However, after taking these differences into account for synchronous AV stimuli, no main effects of linguistic complexity were reported for asynchronous AV stimuli.
(3) On a group level, we noted similar trends between the TBW measure and our loudness perception measure (e.g., narrower TBW window for pseudowords than words and earlier drop in loudness perception).
(4) On an individual level we could not find a correlation between the TBW shape and the loudness drop.
(5) Many questions remain unexplored regarding the loudness effect created by AV speech signals, but the mapping of changes in the subjective perceived loudness measure to a more objective perceived change in TMR (in dB) make this a very promising measure.

5.5. The contribution of linguistic complexity and temporal synchrony to AV speech intelligibility

Speech intelligibility is influenced by the (a)synchrony of the AV signal (e.g., The increased difficulty of understanding speech while watching a poorly synchronized movie or listening in a poorly connected video call.). For a certain range of SOAs (~ 150 - 200 ms; Grant et al., 2004) speech intelligibility stays rather stable; however, for larger SOAs speech intelligibility significantly declines (Grant et al., 2004). Nonetheless, even for large asynchronies, AV speech intelligibility is still enhanced compared to auditory-only performance (Buchwald et al., 2009; Campbell & Dodd, 1980; Grant & Greenberg, 2001). Furthermore, temporal synchrony perception on an individual level might have some power in explaining AV integration (Stevenson et al., 2012). At least for high-level linguistic information like sentences, individuals experiencing greater negative effects of AV asynchrony on AV speech intelligibility demonstrated significantly larger AV benefits for synchronous stimuli (Grant & Seitz, 1998).

How speech intelligibility changes across a range of SOAs (0 - 500 ms), for stimuli of different linguistic complexity, remains unresolved. Both temporal and linguistic characteristics that intrinsically change with increasing linguistic complexity might influence speech intelligibility in different ways. Therefore, we manipulated both linguistic complexity and temporal (a)synchrony in an AV speech intelligibility task. For synchronous AV speech we hypothesized that AV speech intelligibility would improve with increasing linguistic complexity. For asynchronous stimuli we hypothesized that the intelligibility for longer stimuli would be influenced by short-term memory, especially at the wider SOAs. Therefore, intelligibility would drop earlier for longer stimuli.

5.5.1. Methods

5.5.1.1. Participants

All participants in this task, also completed task 1 (Section 5.2) and 3 (Section 5.4). In total, 20 participants consented to the task (Table 5.1), within 5 months of completing the initial task.

5.5.1.2. Speech intelligibility measure

This remote speech intelligibility measure was an open set speech recognition task. Participants listened to either an auditory stimulus or an AV stimulus in a four-talker babble masker. After the video presentation, participants typed into a textbox what they heard as accurately as possible.

We selected three linguistic levels for this task: words, meaningful sentences and anomalous sentences. Although it would have been really interesting to complete this task also for pseudowords and reversed words, to make a better comparison between linguistic levels, and different tasks (Section 5.2 - 5.4), the nature of pseudowords and reversed words does not allow for a straightforward scoring of a 'typed' response.

The AV stimulus presentation was similar to the earlier tasks (Section 5.2 - 5.4). The audio and video of a female talker was presented in a cut-out circle of a black frame. The total video duration was 4000 ms. In the auditory-only presentation, the audio was matched with a 4000 ms black screen with a white star in the middle. We opted for this presentation over a still frame of the talker, as the remote format might accidentally trigger the suspicion that the video incorrectly froze (Gijbels & Lee, 2023), and therefore influence the response of the listener.

All stimuli were presented in four-talker babble fragments (Van Engen et al., 2014), at a TMR of -6 dB similar to Tasks 1,2, and 3. AV stimuli were presented at 5 different SOAs: 0 ms, 50 ms, 150 ms, 300 ms and 500 ms (similar to Section 5.4).

We selected 40 target words per condition. For words this meant that participants had to complete 40 trials per each of the 5 SOAs and 40 trials of auditory-only speech recognition. For meaningful and anomalous sentences, 10 sentences per SOA, and 10 in the auditory-only modality, were presented, as one sentence contained 4 target words. This resulted in a total of 360 trials (240 for words, 60 for meaningful sentences and 60 for anomalous sentences). Stimuli were presented in blocks per linguistic level. The sentence blocks were split up in mini blocks of 30 trials each. The word stimuli were split up in 6 mini blocks of 40 trials each. The different modalities and different SOAs were mixed randomly within blocks. The order of the blocks varied by participant. All stimuli were scored at the word level.

All stimuli were selected from the same cohort of stimuli as used in the previous tasks (Al-Salim et al., 2020; Kocins et al., 2022; Grieco-Calub et al., 2023; Holt et al., 2011; Shatzer et al., 2018; Stelmachowicz et al., 2000; Van Engen et al., 2014). Critically, none of the stimuli selected for this task were used as target or training stimuli in the previous tasks (Section 5.2 -5.4).

A training session was introduced at the beginning of the experiment to make the listener familiar with the task set-up (4 AV stimuli in quiet and 1 auditory-only stimulus) and with the task in noise (5 trials in TMR -6 dB). The stimuli used in the training were not used in the actual experiment.

5.5.1.3. Data analysis

Outliers

No outlier selection was done as the current participants already successfully completed task 1, (2), and 3.

Response times

Response times have very little meaning in an open 'type-what-you-heard' task, and were therefore not analyzed for this section.

Speech intelligibility by linguistic level

To investigate speech intelligibility by linguistic level, we created a subset of the data using only synchronous AV presentations (SOA = 0 ms) and the auditory-only presentations. A linear mixed effects model (Model 12: Speech intelligibility performance ~ linguistic level * modality + (1|participant)) allowed us to interpret the effects of modality and linguistic complexity, and its relationship, on speech intelligibility, for synchronous AV stimuli and auditory-only stimuli. We used AV modality and meaningful sentences as reference levels in our treatment coding. We included participant as a random effect.

As we expected speech intelligibility performance to be significantly more variant in one modality than in a multisensory environment (Altieri & Hudock, 2014), we used a Levene's test to assess the equality of variances between modalities. This test was used since normality needs not to be assumed (Levene, 1960). We similarly interpreted potential differences in variance of speech intelligibility performance by linguistic levels.

Speech intelligibility by temporal asynchrony and linguistic level

Our speech perception tasks obtained information about intelligibility performance in the auditory and AV modality. In the previous model (Model 12), we addressed how performance on both modalities was related to each other, for each linguistic level. When including temporal asynchrony it might be more informative to interpret the AV speech perception benefit (AV - auditory-only). As SOAs have no meaning in a unimodal task (auditory-only), the function of AV performance by SOA stayed the same whether we used AV speech intelligibility scores, or AV speech perception benefit. The only change would be the starting point. Thus, we calculated for each individual participant their AV speech benefit by subtracting auditory performance (for each linguistic level) from AV speech performance (for each specific linguistic level and SOA). This AV benefit score allowed us to better interpret

the effects of the linguistic complexity and temporal asynchrony by taking individual differences in auditory performance into account at SOA = 0 ms.

As we had little indications to what kind of function would fit the data best, and we only measured performance at 5 different SOAs (between 0 and 500 ms), we fitted a Loess function to the data for specific analysis. Loess (i.e., local regression) is a non-parametric strategy for fitting smooth curves to empirical data, using a direct generalization of traditional least-squares methods. Loess is nonparametric in the sense that the fitting technique does not require an a priori specification of the relationship between the dependent and independent variables (Jacoby, 2000). We fitted a Loess function per linguistic level to derive the SOAs at which the AV benefit would be zero.

We had no prediction about the AV benefit by SOA for each linguistic level, we did not assume a linear model fit. Therefore, we performed Wilcoxon rank-sum tests on the AV benefit between the three linguistic levels at each SOA measure. As this test does pairwise comparisons, it accounts for the participant factor, and as it is a non-parametric test, normal distribution is not automatically assumed. Bonferroni correction was used for multiple comparisons.

To compare how well auditory, synchronous AV (SOA = 0 ms), asynchronous AV (SOA = 500 ms) speech intelligibility performance, the difference in performance between synchronous and asynchronous AV stimuli, and AV speech perception benefits correlated between linguistic levels on an individual level we calculated Pearson correlation coefficients.

Speech intelligibility, TBW width, and loudness perception

To compare information of the different tasks (Section 5.2, 5.4, and 5.5) on a group level, while having no hypotheses about their relationship, we decided to explore the data visually by plotting the commonalities between the tasks (Fig 5.13). Therefore, we selected word stimuli for all three tasks, and added both sentence types for the

TBW measure and the speech intelligibility task. Only the TMR of -6 dB was included for these observations. Sigmoid functions were plotted to the TBW shapes (Powers et al., 2009). And Loess functions were plotted for the other two tasks as we had no prior hypotheses about the shape of these functions.

To compare data on an individual level we used Pearson correlation coefficients between the TBW task, the loudness task and the speech intelligibility task for words, meaningful sentences and anomalous sentences. For the intelligibility task we also correlated AV performance at SOA = 0 ms and at SOA = 500 ms, auditory-only performance, the AV speech perception benefit (AV performance (at SOA = 0ms) - auditory-only performance) and the asynchrony difference (AV performance at 0ms - AV performance at 500 ms). For the TBW task, we correlated AV performance at SOA = 0 ms and at SOA = 500 ms, the SOA value for 50% and 75% synchrony perception (TBW width), and the asymptote amplitude (% perceived synchrony at 0ms - % perceived synchrony at 500 ms) . For the loudness task we interpreted the perceived TMR drop.

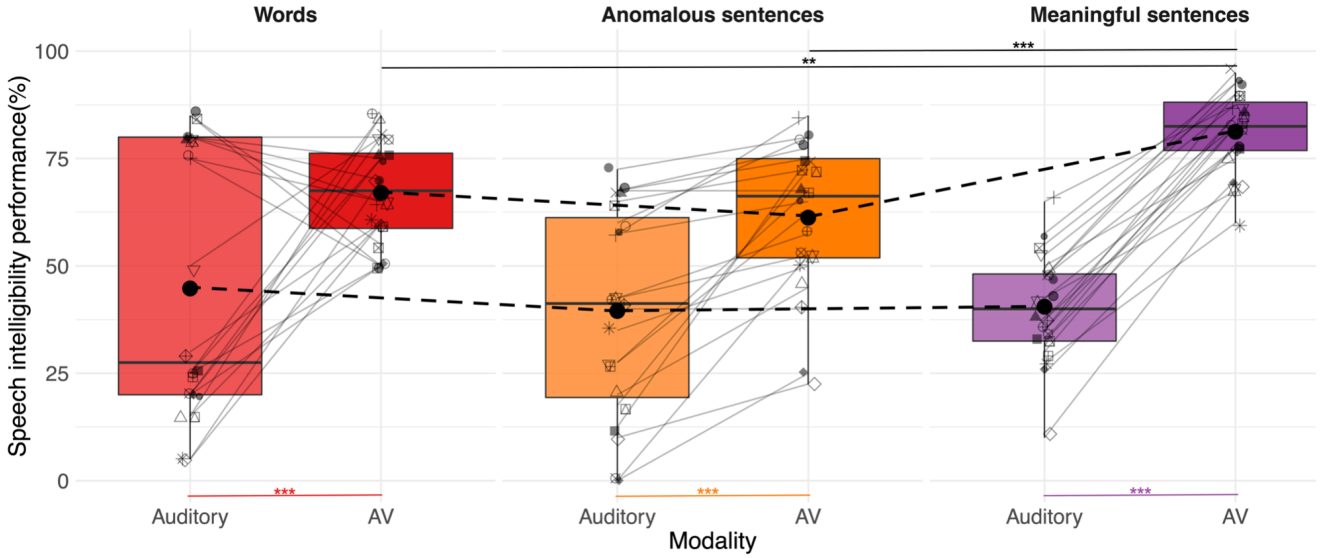
5.5.2. Results

5.5.2.1. Speech intelligibility by linguistic level

Fig 5.10 shows speech intelligibility performance per linguistic level and modality for synchronous AV (SOA = 0 ms) and auditory only stimuli. We observed larger variability in performance in the auditory modality than in the AV modality ($F_{1,118}$ = 11.22, p = 0.001). This was more pronounced for words, than anomalous sentences, than meaningful sentences. However, these variances were not significantly different from each other (p > 0.10). When we looked at the variance of linguistic levels for only the auditory modality, we did find a significant effect ($F_{2,57}$ = 6.38, p = 0.003). The variability of auditory performance for words was so large that it covered the entire AV performance. This indicated that not everyone perceived AV intelligibility benefit (over auditory performance) for words. However, for meaningful sentences

performance was completely distinct between the two modalities showing that all participants benefited from AV speech compared to auditory-only speech on a sentence level.

FIGURE 5.10. Boxplots of speech intelligibility performance by linguistic level and modality for synchronous stimuli.



AV (SOA = 0 ms). The mean is represented as a black circle. The median is a black horizontal line. Individuals are represented by different shapes. Significance: $p < 0.01$ (**), $p < 0.001$ (***)

The linear mixed effects model (Model 12) showed that intelligibility scores were significantly better in the AV modality ($t = -8.44$, $p < 0.001$), across all linguistic levels. It also showed that in the AV modality performance for meaningful sentences was significantly higher than performance for either anomalous sentences ($t = -4.14$, $p < 0.001$) or words ($t = -2.95$, $p = 0.004$). Post-hoc analysis showed no effect between words and anomalous sentences. In the auditory modality, there were no significant differences (based on the mean) between either linguistic categories. Furthermore, there was an interaction effect between linguistic level and modality showing that the improvement in performance by modality was significantly larger for meaningful sentences than for words ($t = 2.71$, $p = 0.008$), or anomalous sentences ($t = 2.78$, $p = 0.007$). Post-hoc analysis showed no significant interaction effect between words and anomalous sentences ($p > 0.10$).

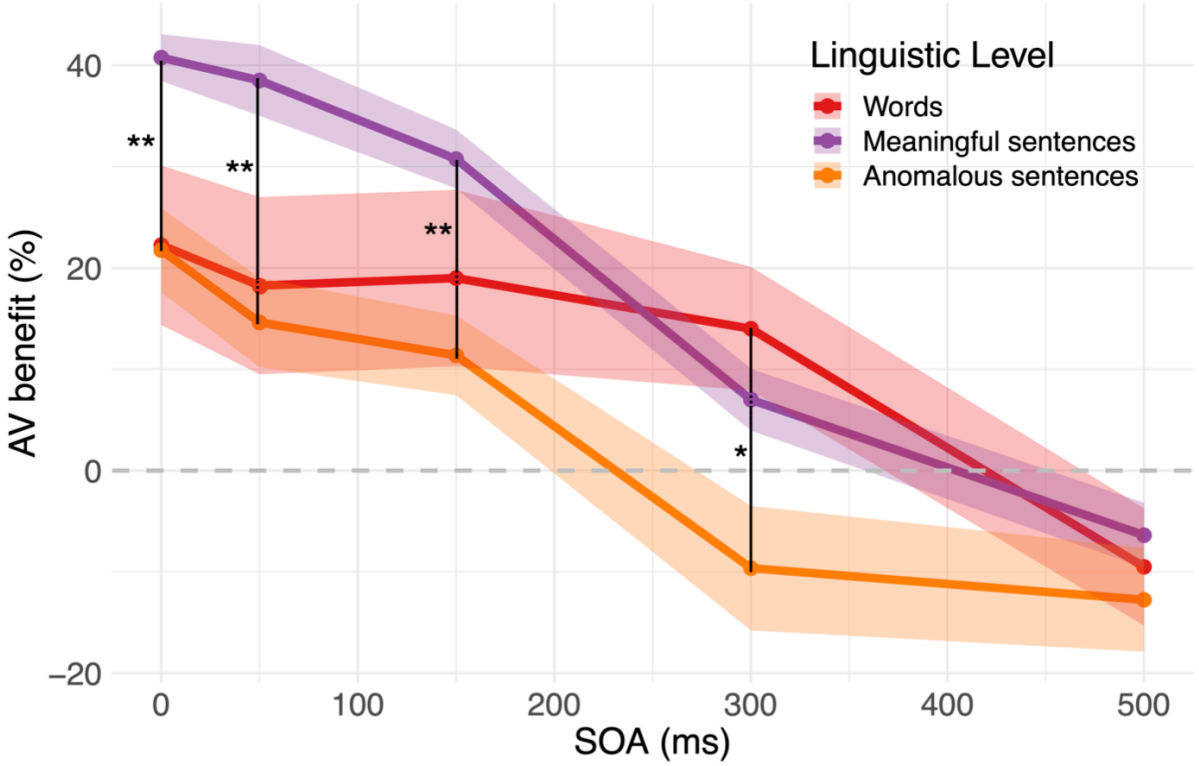
5.5.2.2. Speech intelligibility by temporal asynchrony and linguistic level

Analysis on a group level

We reported in the previous section (Section 5.5.2.1) that there was a significant AV speech perception benefit in all modalities when comparing performance on the speech intelligibility task of auditory stimuli to synchronous AV stimuli. Performance in the auditory modality was not significantly different for any of the linguistic levels and performance in the AV modality only differed with meaningful sentences. Indeed when looking at the AV benefit for synchronous stimuli we found a 22.25% AV speech intelligibility benefit for words, 21.75% for anomalous sentences and 40.75% for meaningful sentences (Fig 5.11). When introducing AV asynchronies we observed a shallow decline until 150 ms for meaningful sentences, followed by a steeper decline until 300 ms and again a more shallow decline between 300 and 500 ms SOA (Fig 5.11). As these reported SOAs were artificially determined by the task, we used a Loess function to calculate whether, and at what SOA the AV speech perception benefit disappeared ((AV - Auditory-only) ≦ 0%). For meaningful sentences the 0%-point was estimated to be reached at an asynchrony of 369 ms. Beyond this point there was a disadvantage of AV stimulus presentations (compared to auditory-only), resulting in a disadvantage of 6.35% at 500 ms. For anomalous sentences we reported a very similar shape, although performance started lower and therefore already crossed the 0% benefit at 236 ms, ending in a 12.75% disadvantage of AV stimulus presentations at 500 ms. We also observed a more noticeable initial drop in performance between 0 and 50 ms SOA (7.13% drop, vs 2.25% for meaningful sentences). Words, however, told a different story. Audiovisual speech perception performance was relatively stable up to 300 ms (8.25% drop), followed by a steep decline, ending in a 9.50% disadvantage of AV stimuli at 500ms. Words crossed the 0% benefit line at 436 ms. Thus, all linguistic levels started with a significant AV speech perception benefit, but for large asynchronies (i.e., 500 ms) all linguistic levels, even the short words, passed the 0% benefit point and showed a

disadvantage of a multisensory stimulus presentation compared to auditory speech information. There was no significant difference between the 0%-point of words and meaningful sentences.

FIGURE 5.11. AV speech perception benefit by SOAs.



AV benefit defined as AV - auditory speech intelligibility performance. SOAs ranging from 0 - 500 ms. 95% confidence interval (shaded area), per linguistic level. The horizontal gray dashed line is the cross-section where AV performance is equal to auditory-only performance. Significance: $p < 0.05$ (*), $p < 0.01$ (**).

To interpret the size of these reported differences between different linguistic levels we used paired samples Wilcoxon tests with Bonferroni correction for the AV benefit (between linguistic levels) for each SOA. We found significant differences between meaningful sentences and anomalous sentences for SOAs 0ms ($p = 0.003$), 50 ms ($p = 0.002$) and 150 ms ($p = 0.002$). There were no differences between words and any of the sentence types for SOAs of 0ms to 150 ms. However, at 300 ms anomalous sentences were significantly different from words ($p = 0.04$), but meaningful sentences were not significantly different from either words or anomalous sentences.

At 500 ms, no significant differences were observed between any of the linguistic levels.

By observing the plot, it is rather surprising that meaningful sentences were not significantly different from words in the 0 - 150 ms range. However, the large variability in the word category, and the strict multiple comparison corrections of our test might explain these more cautiously interpreted results.

Analysis on an individual level

We investigated how well auditory, synchronous AV (SOA = 0 ms), asynchronous AV (SOA = 500 ms) speech intelligibility performance, the difference in performance between synchronous and asynchronous AV stimuli, and AV speech perception benefits correlated on an individual level within and between linguistic levels (Fig 5.12).

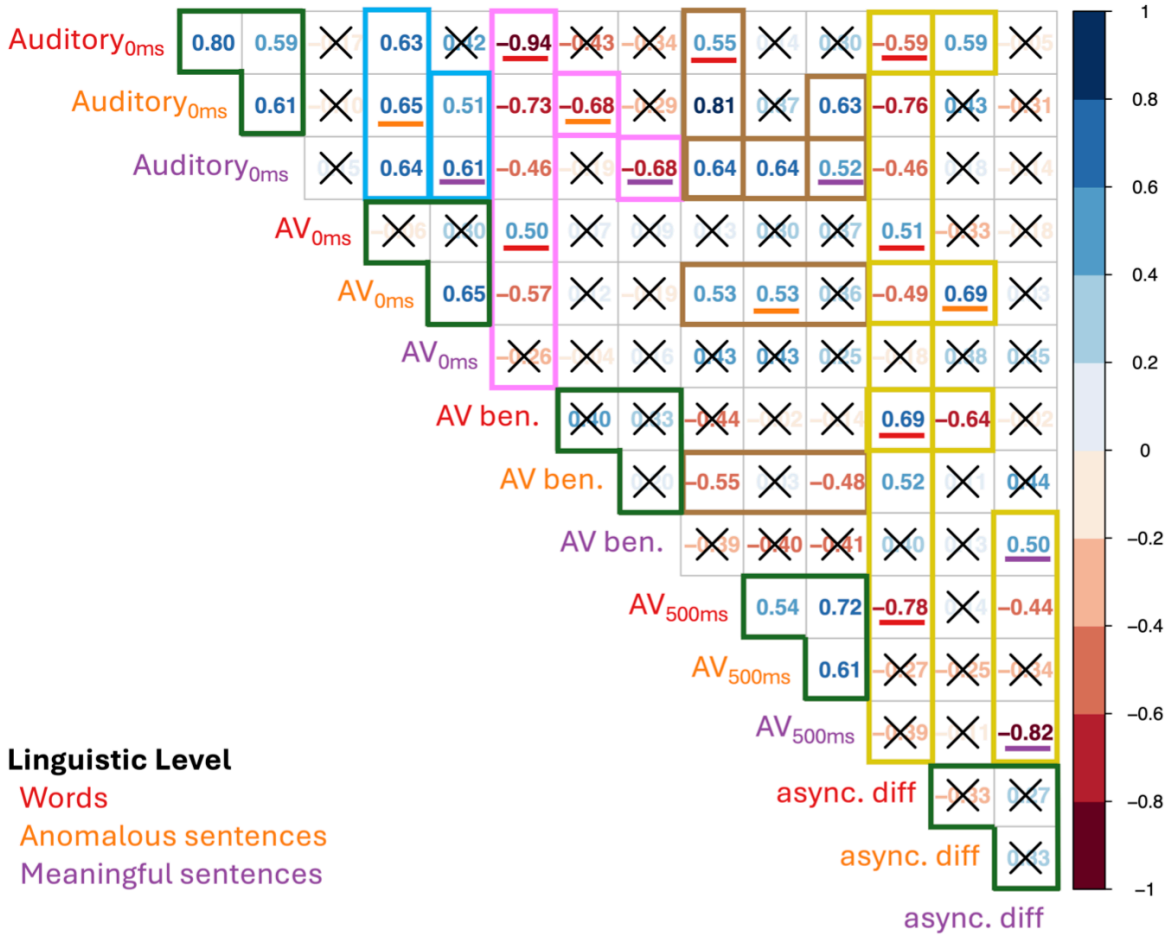
As there are so many comparisons, the plot (Fig 5.12), feels rather chaotic, so we will go through step by step. All values discussed here are significantly correlated, at least at $p < 0.05$.

First, looking at the green boxes, which represent the relationships between the different linguistic levels for each measured category, we found a strong positive correlation between all linguistic levels for the auditory intelligibility measures. Intelligibility for auditorily presented words showed a very high correlation to intelligibility for auditorily presented anomalous sentences ($R = 0.80$). Auditory performance for the two sentence types was highly correlated ($R = 0.61$), and auditory performance between words and meaningful sentences was significant, but moderately correlated ($R = 0.59$). For the synchronous AV measures (AV_{0ms}) a significant, positive correlation ($R = 0.65$) showed, but only between the two sentence types. For AV benefit (AV ben.) – the difference between the synchronous AV and auditory performance – no significant correlation between the different linguistic levels was found. Similarly, no significant correlation between the different linguistic

levels was reported for the drop in intelligibility, caused by AV asynchrony (async diff: AV_{0ms} - AV_{500ms}), however the AV performance for completely asynchronous stimuli was positively correlated between all linguistic levels. Thus, between linguistic levels, both measures that were least likely to be influenced by visual cues (auditory performance and AV performance at SOA = 500 ms), were significantly correlated. However, all measures where AV information was expected to have an influence (AV performance at SOA = 0ms, AV benefit and AV intelligibility difference caused by temporal asynchrony), showed limited correlations between linguistic levels. Thus, changes in intelligibility scores between linguistic levels were most expressed when both auditory and visual cues could be used to the benefit of the listener.

Second, the correlations within the light blue boxes represent the relationship between synchronous AV and auditory measures, within and between linguistic levels. Most notably, synchronous AV performance on a word level was not correlated to the auditory measure, in any of the linguistic categories, not even within the word category. The opposite was reported for anomalous sentences, where synchronous AV intelligibility scores were positively correlated to auditory performance in all linguistic categories. For synchronous AV performance of meaningful sentences we found significant positive correlations with auditory performance on the two sentence types, but not with words. Thus individuals with high performance in the AV modality (for synchronous stimulus presentations) for sentences were more likely to also show high performance in the auditory modality for sentences. That we did not find any correlation for word intelligibility was not surprising as Fig 5.10 showed that AV performance for words completely fell within the range of auditory performance for words. In sum, sentences most likely have characteristics (temporal and/or linguistic) that can be used across modalities, and for both sentence types, resulting in significantly correlated performance.

FIGURE 5.12 Pearson correlation coefficients between the intelligibility measures.



Correlation plot between different aspects of the intelligibility measure for words (red), anomalous sentences (orange) and meaningful sentences (purple) top). Negative correlations are color-coded red, while positive correlations are coded blue. AV ben (pink) = AV speech perception benefit (AV - A) at 0 ms SOA, $Auditory_{0ms}$ = auditory performance for SOA = 0ms, AV_{0ms} (light blue) and AV_{500ms} (brown) = AV performance for SOA = 0 ms and SOA = 500 ms, async. diff (yellow) = the difference in AV intelligibility performance caused by the difference in asynchrony (0 ms - 500 ms). Dark green boxes are correlations between linguistic levels, within a category. Significant values within a linguistic level are underlined in the color of that linguistic level. Only significant values ($p < 0.05$) are presented.

Third, in pink we showed the correlation between AV speech perception benefit and auditory and synchronous AV performance. For both sentence types, we find an equal, strong, negative correlation between the AV benefit and the auditory performance within the same linguistic category (R = -0.68). This suggests that individuals with low auditory performance in that specific sentence category are likely to have high AV speech perception benefits within that same sentence

category. No other correlations were reported to the AV benefit of sentences. In contrast, AV benefit of words was significantly negatively correlated to auditory performance in all linguistic categories. There was a very high correlation within the word category (R = -0.94), hinting that auditory performance might explain most of the variance in AV speech perception benefits for words. Individuals with high AV speech perception benefits for words were more likely to also have higher synchronous AV speech intelligibility scores (R = 0.50), however they were more likely to have lower synchronous AV speech intelligibility scores (R = -0.57) for anomalous sentences. There was also a negative trend between AV benefit for words and synchronous AV performance for meaningful sentences, however this was not significant. Thus, individuals with good auditory performance were more likely to show high AV speech perception benefits. Auditory performance had at each linguistic level a significant role in the size of the AV speech perception benefit, AV speech performance, however, was only informative for AV benefit on a word level. This suggests that certain characteristics of sentences aid to the AV benefit, that cannot be explained by solely perceptual (AV) performance.

As we mentioned a similar trend of correlations between linguistic levels for auditory speech intelligibility and asynchronous AV speech, it was not surprising to find some significantly positive correlations there. Thus, how well an individual understands asynchronous AV speech for words goes hand in hand with auditory performance for all linguistic levels. However the correlation within the same linguistic category is the lowest. We see this for both words and meaningful sentences, and there is not even a significant correlation for anomalous sentences. For anomalous sentences, we found positive correlation to the synchronous AV performance. How to interpret these results is less clear. Most likely listeners try to use asynchronous AV speech to their benefit, but fail and therefore show performance that is even worse than auditory-only performance (Fig 5.11).

Lastly, the correlations within the yellow boxes represent how the drop in intelligibility scores caused by AV asynchrony are correlated to any of the other measures. Again, we find remarkably more significant correlations between the different measures for the asynchrony drop of word stimuli. The more the AV asynchrony influences speech intelligibility for words, the larger the AV benefit for synchronous word stimuli, the higher AV performance, the lower auditory performance and the lower asynchronous AV performance for words. The drop in intelligibility caused by asynchrony thus correlates to all other word measures, as expected by what we previously discussed. This suggests a rather straightforward relationship between synchronous and asynchronous, auditory and AV measures on a word level. However for both sentence types this is not the case. The drop in intelligibility for meaningful sentences is not related to auditory or synchronous AV performance, but mainly to the asynchronous AV performance ($R = -0.82$), and it is not surprising that lower performance for asynchronous sentences correlates to larger drops in intelligibility caused by asynchrony (as we saw in words: $R = -0.78$). However, for anomalous sentences we did not observe this. We did again see that synchronous AV performance resulted in the strongest correlation, now suggesting that higher intelligibility scores for AV anomalous sentences more likely resulted in larger drops caused by temporal asynchrony of the AV stimuli.

Thus, especially within the word category the intelligibility measures were strongly correlated. Individual auditory performance seemed to play a significant role in explaining all other measures within the word category. The size of the AV speech perception benefit on a word level was correlated to both auditory and synchronous AV measures for both words and sentences, however it was not correlated to any of the AV speech perception benefits on a sentence level. The combination of these findings suggest that AV benefits for words can mainly be explained by auditory performance, however on a sentence level other characteristics of the speech signal play a role too. Similarly, asynchronous AV performance correlated well between

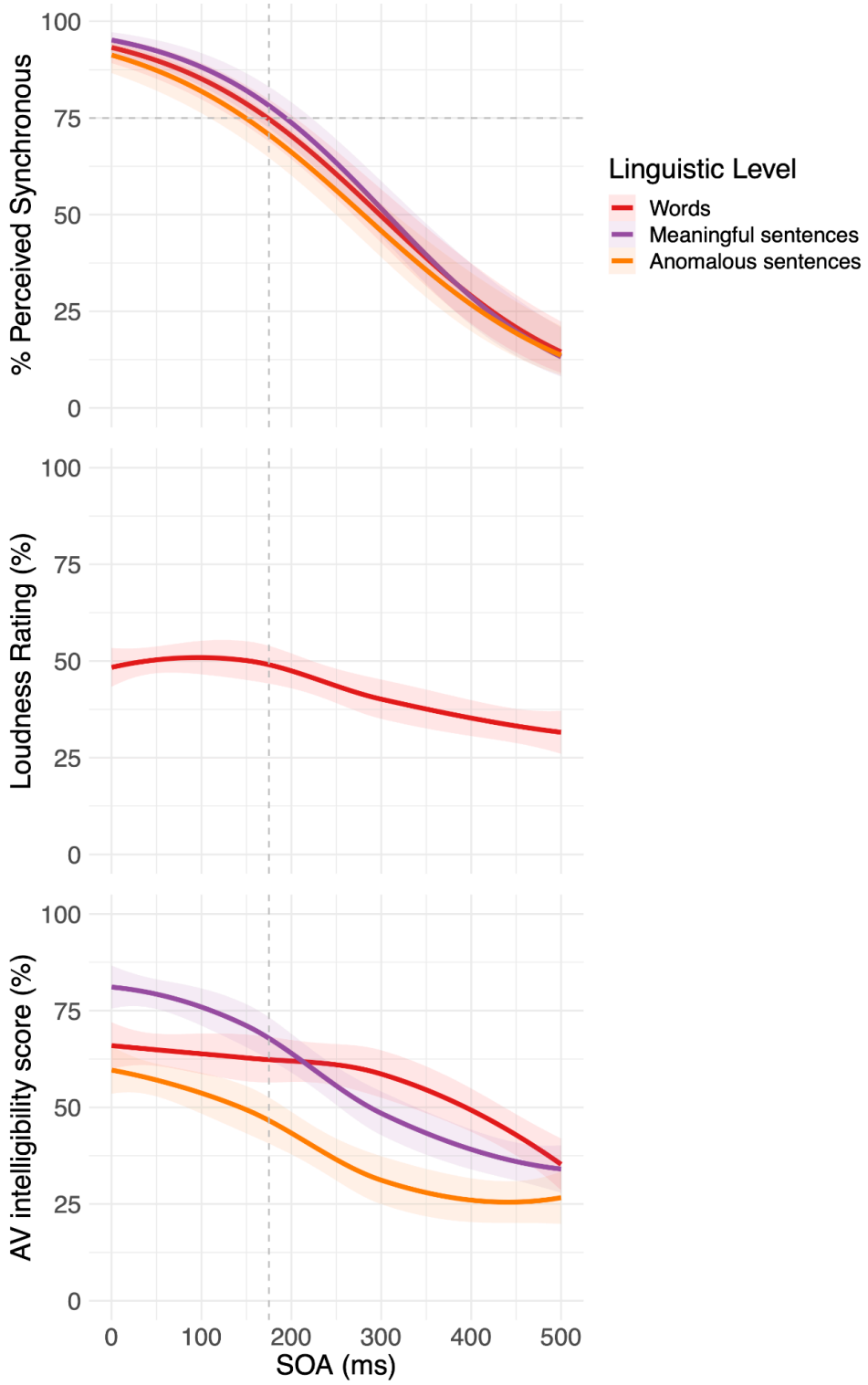
linguistic levels, and was overall well correlated to auditory performance, yet on a sentence level this was less clear.

5.5.2.3. Speech intelligibility, TBW width, and loudness perception

Analysis on a group level

In this dissertation study, we gathered information about AV speech stimuli that were either presented synchronous (SOA = 0ms), or asynchronous, with the visual stimulus preceding, up to 500 ms. The listeners (1) rated these speech stimuli as synchronous or asynchronous (Task 1; Section 5.2), (2) rated how loud they perceived these stimuli Task 3; Section 5.4), and (3) performed a speech intelligibility task on these stimuli (Task 4; Section 5.5). To compare the information from these 3 tasks we selected the linguistic levels that were in common (words, meaningful sentences and anomalous sentences for temporal synchrony perception and intelligibility, and words for loudness rating) at a TMR of -6 dB, as shown in Fig 5.13. By observing the plots, and accumulating information from the different sections, we noted that, on a group level, there were some commonalities between these tasks.

FIGURE 5.13. Performance across the three tasks by SOA.



Top plot: the TBW measure (Section 5.2) for words, meaningful sentences and anomalous sentences, with the % perceived synchrony on the y-axis. Middle plot: loudness measure (Section 5.4) for words at TMR = -6 dB with the loudness rating (%) on the y-axis. Bottom plot: intelligibility task (Section 5.5) for words, meaningful sentences and anomalous sentences, with the AV intelligibility score (%) on the y-axis.

For all tasks, performance was significantly lower at 500 ms asynchrony, than for synchronous stimuli (SOA = 0ms). For both the TBW and the intelligibility task, performance for synchronous stimuli was higher for meaningful sentences than for words, and performance for words was higher than performance for anomalous sentences, although the difference between words and the 2 sentence types was not statistically significant. The influence of linguistic complexity disappeared at 500 ms SOA. For the speech intelligibility task, we also observed sigmoid-like performance for both sentence types, similar to the TBW shapes. However, for words there was a longer constant performance with a later SOA drop.

As the 50% synchronous point as a TBW-width measure is rather random, we could select any other point for comparison (e.g., Stevenson et al., 2012 vs. Grant et al., 2004). We selected the 75% point of perceived synchrony as a comparison point (Fig 5.13, top), as this is where the slope of the TBW really started to steepen. We observed a matching SOA of 173 ms for words, 149 ms for anomalous sentences, and 193 ms for meaningful sentences with this 75% point of perceived synchrony. Interestingly, this is exactly where our loudness perception (Fig 5.13, middle) started to decline, and where the intelligibility for both sentence types notably dropped (Fig 5.13, bottom). For words however, intelligibility performance was still stable at 173 ms SOA.

Analysis on an individual level

To compare performance between tasks on an individual level we used Pearson correlation measures by linguistic level (meaningful sentences, anomalous sentences and words). We already discussed the correlations between the intelligibility measures in Section 5.5.2.2 and will therefore not repeat this.

Within the TBW task the order and direction of effects were fairly consistent between linguistic levels. For example, for all linguistic levels the width of the TBW at the 75% synchrony perception point was significantly positively correlated to both the

upper (R_{words} = 0.71; $R_{mean.sent}$ = 0.69; $R_{anom.sent}$ = 0.59) and lower (R_{words} = 0.57; $R_{mean.sent}$ = 0.57; $R_{anom.sent}$ = 0.50) asymptote. The asymptote amplitude however was not correlated to either the 50% or 75% TBW width, for any linguistic level.

For the loudness measure we interpreted correlations of the perceived TMR drop between linguistic categories (i.e., words, pseudowords and reversed words). At SOA = 300 and 500 ms the TMR drop for the three different word levels was significantly positively correlated (R_{300ms} = 0.58 - 0.78; R_{500ms} = 0.65 - 0.79), but not at 50 ms or 150 ms.

As reported in Section 5.4, we did not find any correlation on an individual level between the perceived TMR drop and any of the TBW measures, potentially indicating they represent different aspects of AV speech integration.

Between the loudness task and the speech intelligibility task we found a significantly positive correlation between the drop in intelligibility caused by temporal asynchrony for words and the perceived TMR drop for words at SOA 150 ms (R = 0.62). There was a similar trend for SOA = 50 ms (R = 0.42), 300 ms (R = 0.45), and 500 ms (R = 0.43), however not significant. This suggests that individuals with larger perceived TMR drops, especially early on (SOA = 150 ms), were more likely to be more influenced by temporal asynchrony in their speech intelligibility. Similarly, and not so surprising, there was a significant negative correlation between the intelligibility for asynchronous words (SOA = 500 ms) and the TMR drop (150 ms: R = -0.49, 300 ms: R = -0.51, 500 ms: R = 0.51), on a word level. This suggests that listeners with small perceived TMR drops were more likely to have better AV speech perception scores for asynchronous AV stimuli. Lastly, but maybe most the most interesting finding was that the TMR drop for words, was significantly negatively correlated to auditory speech intelligibility, not of words, but for anomalous sentences (R_{50ms} = -0.47; R_{150ms} = -0.44; R_{300ms} = -0.53; R_{500ms} = -0.47), suggesting that high auditory performance was correlated to small perceived TMR drops. Going

back to the correlations between the intelligibility measures (Fig 5.12, row 1 and 2) it was not so surprising to report this, as auditory performance for anomalous sentences correlated stronger to asynchronous AV intelligibility for words ($R_{anom.sent}$ = 0.81 vs. R_{words} = 0.55) and the intelligibility drop caused by AV asynchrony ($R_{anom.sent}$ = -0.76 vs. R_{words} = -0.59), than auditory performance for words itself. Thus although it would be really interesting to also obtain perceived TMR scores for both sentence types, we conclude from the current data set that on an individual level the perceived TMR drop caused by asynchrony is related to the influence of asynchrony on word intelligibility.

Between TBW and speech intelligibility tasks, we reported no significant correlations on the word level. For anomalous sentences we found a strong negative correlation between the AV benefit (at 0 ms) and the TBW width (50%: R= -0.67 ; 75%: R= -0.62). This suggested that individuals with narrower TBWs tended to have larger AV benefits for anomalous sentences. This is interesting, as for the intelligibility task we found strong negative correlation between the AV benefit and auditory performance (R = -0.68), but none of the AV measures.

For both sentence types there was a significantly positive correlation between the AV speech intelligibility scores at 0 ms and the asymptote amplitude of the TBW ($R_{anom.sent}$ = 0.53; $R_{mean.sent}$ = 0.54), and a significantly negative correlation with the lower asymptote of the TBW (R = -0.51). This suggests that individuals with more pronounced TBWs (larger asymptote amplitude) were more likely to have better AV speech intelligibility scores for synchronous sentences, which was most likely pulled by lower synchrony ratings at large asynchronies. Comparing this to the speech intelligibility measures we showed that synchronous AV intelligibility for both sentence types was strongly correlated to each other (R = 0.65) and to the auditory performance on both sentence types. Furthermore, for words, intelligibility for synchronous AV speech was not correlated to any of the auditory performances (words or sentences), not to any of the synchronous AV sentences measures, nor to

any of the TBW asymptote measures. Therefore one could argue that the TBW asymptote represents the relationship between auditory and synchronous AV performance. If there is no relationship between synchronous AV and auditory intelligibility, a relationship between AV intelligibility scores and TBW amplitude should also not be expected (i.e., as for words).

Furthermore, for meaningful sentences a significant negative correlation was reported between TBW width at 50% synchrony perception and AV intelligibility for synchronous AV stimuli (R = -0.52) and auditory speech intelligibility (R = -0.56), suggesting that listeners with narrower TBWs (at the 50% synchrony point) had significantly higher intelligibility scores in the both the auditory and the AV modality for meaningful sentences. In contrast to the anomalous sentences, no significant correlation was reported between the TBW width and the AV speech perception benefit.

In sum, on a word level a significant correlation was observed between the perceived TMR drop caused by stimulus asynchrony and speech intelligibility performance influenced by temporal asynchrony: the larger the TMR drop, the larger the intelligibility drop. For the TBW task and the intelligibility task we found no significant correlations on a word level. It is interesting that the most predictive category for our intelligibility measure, and the category that shows correlations to the loudness measure shows no correlations with the TBW. We did report correlations to the asymptote measures of the TBW width and synchronous AV performance for both sentence types. Furthermore a negative correlation between AV benefit and TBW width was reported for anomalous sentences and a negative correlation between auditory and synchronous AV performance, and the TBW width for meaningful sentences. However, the interpretation of this is less clear.

5.5.3. Discussion

To interpret the relationship between linguistic complexity and temporal (a)synchrony of AV speech stimuli we first addressed differences in performance by linguistic category for synchronous stimuli (SOA = 0 ms). Based on findings from Heinrich and Knight (2016) who tested auditory speech intelligibility in words and sentences with low and high context and similar to our work, we would have expected a significant effect of linguistic level with auditory performance being highest for meaningful sentences, followed by words, and lowest performance in the anomalous sentences. This is exactly what we observed for AV speech intelligibility; however, it was not for auditory-only speech intelligibility. A study from Van Engen and colleagues (2014) who tested speech intelligibility performance in both high- and low-context sentences explains the differences between our study and the study from Heinrich and Kneight (2016). Van Engen and colleagues (2014) showed that in the auditory modality predictable semantic context is only beneficial in speech-shaped noise and not in babble. Heinrich and Knight (2016) used speech-modulated noise, where we used four-talker babble. Yet, for AV speech stimuli they reported that semantic context was beneficial in both speech-shaped and babble maskers. Therefore, the combination of added visual information and semantic context could have an increased impact on speech perception (Ma et al., 2009; Tye-Murray et al., 2007b).

The variability in performance in the auditory modality was large, especially for the word stimuli. It was significantly larger than for any of the sentence stimuli. Closer observation of auditory performance in the word category showed that listeners did either not so well (below 27.5% correct, as shown by the median), or really well (above 75 % correct). Whereas for both sentence types performance was more evenly spread around the mean. This was a surprising finding, as we would expect linguistic knowledge of the individual to play a larger role for stimuli with more linguistic information, leading to larger variability for sentences than for words

(Grant et al., 1998). The stimuli of our task might have been easy enough for all native English speakers to complete, limiting expected individual differences in linguistic knowledge. A follow-up study with English-language-learners could help clarify this.

In the AV modality performance was significantly better ($p < 0.001$) for all linguistic levels compared to auditory-only performance, resulting in a significant AV speech perception benefit. For both sentence types almost all individuals showed AV speech perception benefits, however not for the word stimuli. Performance in the AV modality was also less variable than in the auditory modality, and there were no significant differences in variability in AV performance between the three categories. AV speech intelligibility scores were significantly higher for meaningful sentences than either of the other two categories. There was no significant difference between words and anomalous sentences.

Our findings suggest that individuals might have different strategies to complete an auditory speech-in-noise task for words than for sentences. Some individuals could understand the word in the background babble really well, whereas others could not. Furthermore, the individuals who did not have AV speech perception benefits for words, still showed benefits for sentences. Thus, different individuals might have different strategies based on different levels of linguistic information for integration of AV speech signals.

Incorporating temporal asynchrony in the analysis, we found a relatively stable window of AV speech intelligibility scores at all linguistic levels, for SOAs up to 150 ms. We noted a slight, but non significant drop in AV intelligibility scores in this range, that was more pronounced for anomalous sentences, especially in the first 50 ms. After 150 ms, intelligibility scores for both sentence types dropped significantly, and after 300 ms a more shallow decline continued. Grant and colleagues (2004), similarly reported stable AV speech intelligibility scores for visual-leading sentence

stimuli up to 200 ms. These differences in "breaking-point" between their (200 ms) and our study (150 ms), could be explained by the stimulus presentation. Grant and colleagues (2004) manipulated speech intelligibility by using band-pass filtered speech. We, however, presented our stimuli in a babble masker, and Pandey and colleagues (1986) showed that speech intelligibility is affected earlier by temporal asynchrony for stimuli in background noise. Nonetheless, there is consensus that speech intelligibility is stable up to a few 100 ms (Gordon-Salant et al., 2017; Grant and Greenberg, 2001; Grant et al., 2004; Pandey et al., 1986).

Furthermore, we noted that for sentence stimuli the influence of temporal asynchrony on AV speech intelligibility performance presented in a sigmoid-like shape, similar to the TBW, and speech intelligibility started to significantly drop around a similar SOA as where the TBW dropped. For both the intelligibility and the TBW task we also reported that performance was influenced significantly earlier for anomalous sentences. As suggested by Vatakis and Spence (2006), higher linguistic complexity might lead to a larger tolerance for AV asynchronies (i.e., wider TBWs), resulting in a higher tendency to integrate more asynchronous AV stimuli of higher linguistic complexity. This is corroborated by our results, as the AV speech perception benefit disappeared at SOAs of 236 ms for anomalous sentences but only at 369 ms meaningful sentences.

For words, performance was stable up to 300 ms, followed by a significant decline towards 500 ms. This remarkable difference with the sentence type stimuli could be explained by the temporal nature of the stimuli. If the asynchrony is clearly noticeable (SOA = 300 to 500 ms), but still shorter than the length of the total stimulus (+/- 2500 ms for sentences), the visual signal will not provide useable temporal cues similar to synchronous stimuli, but it will also not result in a priming effect, given the asynchronous overlap and the mental capacity that comes with sentence type stimuli. However, for word stimuli significantly preceding visual information might serve as a priming cue for the incoming auditory signal

(Buchwald et al., 2009). Buchwald and colleagues (2009) reported a priming effect for visual word stimuli leading the auditory stream with 500 ms, but it might even be present in larger asynchronies (up to 1600 ms; Campbell & Dodd, 1980). We could adjudicate the persisting AV speech perception benefit from 150 - 300 ms for word stimuli to this priming effect. However, we found a disadvantage of a multisensory stimulus presentation compared to auditory-only performance at 500 ms. In fact, for all stimuli (both words and sentences) a negative AV benefit showed at 500 ms. The difference in our task with the priming studies could again be explained by background noise and the task. When there is no background noise (like in Buchwald et al., 2009 & Campbell & Dodd, 1980), the visual stimulus can serve as a priming cue as long as our mental capacity can keep up with the visual presentation. However, in a noisy environment it would be harder to retain the visual and auditory information. Furthermore, the expectancy of temporal information provided by the AV stimulus in our task – given the large number of synchronously perceived stimuli and the random order of stimulus presentation –, in combination with the babble masker, might actually result in the listener trying to, unsuccessfully, pick one of the auditory streams of the background signal, to match the video and therefore interfere with the actual stimulus presentation when the auditory stream is significantly delayed.

In the comparison of the TBW width and AV speech intelligibility performance on an individual level we reported a significantly negative correlation ($R = -0.52$) for synchronous meaningful AV sentences. This is in line with the findings of Hay-McCutcheon and colleagues (2009; $R = -0.44$) and Conrey and Pisoni (2006; $R = -0.47$), and suggests that listeners who obtained higher AV sentence intelligibility scores tended to be more sensitive to detecting small differences in the relative timing between the auditory and visual signals (Conrey & Pisoni, 2006). Interestingly, we did not find this relationship to be significant for anomalous sentences ($R = -0.37$), nor words ($R = 0.00$). As stimulus duration in itself has a role

(i.e., negative correlation) in sensitivity to temporal synchrony perception (Maier et al., 2011), it is not surprising to find that this correlation is stronger for sentence type stimuli than for words. However, that there were significant differences between meaningful and anomalous sentences, suggests that semantic predictability of the meaningful sentences influenced the relationship between AV speech intelligibility and sensitivity to temporal asynchrony. For anomalous sentences, and only for anomalous sentences, we reported a strong negative correlation (R = -0.67) between the TBW width and the AV speech perception benefit (AV performance (SOA = 0ms) - auditory performance). Stevenson and colleagues (2012) showed that narrower TBWs go hand in hand with increased strength of integration of AV syllables. This could also explain the negative correlation we reported. However we tend to be cautious in the interpretation of the findings from Stevenson and colleagues (2012), as their conclusion is based on the assumption that AV illusions like the McGurk effect (McGurk & Macdonald, 1976) measure AV speech integration, which has been contested more recently (for a review: Alsius et al., 2018).

The more interesting question is, why did we only find a correlation between the TBW width and AV speech perception benefits for anomalous sentences, but not for meaningful sentences and words? For anomalous sentences the AV speech perception benefit (for synchronous stimuli) was negatively correlated to width of the TBW, but none of the AV intelligibility measures (synchronous, asynchronous, or the drop in intelligibility caused by asynchrony) were correlated to the AV speech perception benefit. For meaningful sentences the AV speech perception benefit was not correlated to the TBW width, but individuals whose AV speech intelligibility was more influenced by AV asynchrony did show larger AV speech perception benefits. Furthermore, the TBW width of meaningful sentences was correlated to both auditory and AV performance for synchronous meaningful sentences, yet auditory and AV performance for synchronous meaningful sentences was not correlated to the intelligibility drop caused by AV asynchrony. It becomes even more interesting

when we add words. Intelligibility for words showed no correlation to the TBW measures at all. However the drop in intelligibility caused by asynchronous AV word presentations was correlated to all intelligibility measures (AV benefit, auditory performance, and AV performance for both synchronous and asynchronous words).

How do we explain these differences between linguistic levels and why were only sentence level stimuli correlated to the TBW measures? In a task like this where integration of the AV signal is slightly forced upon the listener, as they were asked to do a speech intelligibility task for AV stimuli, the listener will at least try to use both auditory and AV information to improve speech intelligibility, regardless of whether it is perceived synchronous or not. Listeners would be most successful in doing so for word stimuli as words are less influenced by temporal asynchrony due to the short nature of the stimuli and less cognitive capacity needed to retain the information from both modalities, even when presented asynchronously. On a sentence level, this kind of late-stage integration is more difficult due to the temporal nature of a sentence. For sentence stimuli and more specifically, for anomalous sentences, individuals will rely mainly on more early-stage integration cues as posited by the TBW. Therefore, it is not surprising to find a correlation between AV speech intelligibility benefit measures and the TBW width, for sentences, on an individual level. The linguistic information provided by the sentence might then determine whether this has an effect of both auditory and AV speech intelligibility (i.e., meaningful sentences), or the AV speech perception benefit (i.e., anomalous sentences).

In conclusion:

(1) The access to linguistic information resulted in better speech intelligibility performance for synchronous AV meaningful sentences than for anomalous sentences or words. Auditory speech intelligibility was not significantly influenced by the linguistic complexity of the stimulus.
(2) Anomalous sentences were most affected by temporal asynchrony of the AV speech signal. Linguistic information provided by words and meaningful sentences resulted in stable speech intelligibility performance at longer SOAs.
(3) The specific characteristics of word stimuli (i.e., most likely short duration) allowed to integrate AV stimuli for SOAs significantly beyond the point of synchrony perception.
(4) On a word level, the perceived TMR drop caused by asynchronous presentation of AV speech stimuli correlated to speech intelligibility performance for asynchronous speech stimuli. Future work should explore this relationship on a sentence level.
(5) Temporal synchrony perception was correlated with speech intelligibility, but only on a sentence level.

and

(6) Linguistic characteristics of the speech stimuli explained AV speech perception benefits beyond sensitivity to stimulus (a)synchrony on an individual level.

5.6. General Discussion and Conclusion

The current chapter contributes to the literature in exploring the intricate relationship between prelinguistic and linguistic information of the AV integration process. Additionally, this work is distinctive in that all tasks were assessed successfully remotely, with adult listeners located across the United States. Sourcing stimuli from eight different research groups with results validated already in

previous studies (Al-Salim et al., 2020; Kocins et al., 2022; Lalonde, 2019; Grieco-Calub et al., 2023; Holt et al., 2011; Shatzer et al., 2018; Stelmachowicz et al., 2000; Van Engen et al., 2014) and sentences from the SteVi Speech test corpus (STeVi Speech Test Video Corpus. (n.d.). Sensimetrics' Speech) allowed us to select from a large sets of recordings, and to provide listeners with multiple talkers in each category to lower effects that could be specific to the individual recordings.

To study prelinguistic and linguistic contributions in the AV integration process we assessed a series of four tasks. By using similar task manipulations (prelinguistic and linguistic) and a within-subject design across all tasks, we allowed for a comprehensive interpretation when summarizing findings. Forty-six younger adults (age 21 – 40) completed the first task, and a self-selected subset of this group completed the remaining three tasks. The prelinguistic information was manipulated by varying the temporal synchrony of the AV speech signals. We presented either synchronous AV speech signals, or asynchronous AV speech signals where the visual modality could lead up to 500 ms. The linguistic information was manipulated by using five different categories of linguistic complexity (Table 5.2), each contributing a new step of linguistic information to the signal. At the most basic level we used reversed word stimuli as they contained no linguistic information other than being a speech signal with a temporal match between the audio and video. At the highest level we used sentences with a predictable semantic context. These meaningful sentences provided information about phoneme-viseme connections, lexical, semantical and morpho-syntactic context. In total we defined five distinct stimulus categories: reversed words, pseudowords, words, anomalous sentences, and meaningful sentences.

Temporal synchrony manipulations are a common tool to study AV integration (Zhou et al., 2020 for review), or any type of multisensory integration (e.g., auditory-somatosensory: Navarra et al., 2007; auditory-vestibular: Chang et al., 2012; visual-somatosensory: Fujisaki & Nishida, 2009), and have shown their value in

neurodiverse populations (Wallace & Stevenson, 2014). Audiovisual temporal (a)synchrony tasks use both non-speech and speech stimuli, in artificial (Van den Burg et al., 2011), naturalistic (Vatakis & Spence, 2006), or illusionary contexts (e.g., non-speech: Donohue et al., 2015; speech: Woynaroski et al., 2007). Temporal (a)synchrony is a valuable stimulus manipulation as it has a natural connection to AV speech perception. Face-to-face conversational speech inherently contains visual-leading temporal asynchronies (~150 ms) when the sound reaches the listener's ears (Chandrasekaran et al., 2009) and our perceptual system has this window of tolerance to AV asynchronies (i.e., the Temporal Binding Window; TBW). The width of this window informs us up to what point AV signals are being perceived as AV events and might explain how well each individual listener can integrate AV speech signals (Stevenson et al., 2012; Zhou et al., 2020). Furthermore, AV temporal integration plays a role in language processing and language development. For example, AV temporal processing difficulties have been established in children with developmental language disorders (Kaganovich, 2017). Yet, the current literature of AV temporal processing lacks an analysis of systematic changes that come with different linguistic characteristics in the speech signal (e.g., phoneme-viseme information, lexical, semantical, and morpho-syntactic context).

Therefore, in a first task we examined the role of linguistic information on AV temporal synchrony perception (Section 5.2). We found that high-level (i.e., lexical and semantic) linguistic information changes the TBW shape in a stepwise manner, where higher linguistic complexity resulted in a larger tolerance for asynchronous stimuli and more certainty in the (a)synchrony decision making process. Specifically, items that provide meaning (words and meaningful sentences) have a higher tendency to be integrated at larger stimulus asynchronies than meaningless items (reversed words, pseudowords, anomalous sentences). More basic linguistic information provided by the AV speech signal (phoneme-viseme connections) influenced the certainty level of temporal synchrony perception.

By creating a stepwise change in linguistic complexity, changing temporal and, or linguistic information, we were able to explain some contradictory findings about the influence of linguistic information and temporal synchrony perception from previous work (Maier et al., 2011; Vatakis & Spence, 2006). We concluded that added temporal information improves temporal estimates, leading to more narrow TBWs; however, increased stimulus length only has a limited effect in contrast to the linguistic effect on the TBW shape. These findings corroborate the theory of Mattys and colleagues (2005) that only when interpretive conditions are altered due to a lack of contextual and lexical information, lower-level cues become the driving force behind segmentation.

On an individual level, we found that listeners with more narrow TBWs for the most complex linguistic level (i.e., meaningful sentences), tended to show the just described effect where increased linguistic complexity was related to increased TBW width. However, listeners with wide TBWs for meaningful sentences showed little influence of linguistic complexity on the TBW width (Fig 5.2). Furthermore, our speech intelligibility task (Task 4, Section 5.5) showed that individuals with narrower TBWs for meaningful sentences had better intelligibility scores in both the auditory and AV modality (similar to Conrey and Pisoni, 2006; Hay-McCutcheon and colleagues, 2009). Thus, for stimuli that contain all levels of linguistic information (i.e., meaningful sentences), individuals with narrower TBWs tended to have (1) better AV speech intelligibility scores (for synchronous stimuli), (2) better auditory speech intelligibility scores, and (3) were more likely to be influenced by linguistic information in their temporal synchrony perception. Whether this means that within conversational settings, where individuals possess ample temporal and linguistic cues, listeners who are more sensitive to temporal (a)synchrony perception are more likely to use this linguistic information – as their TBW width widens more with increased linguistic complexity – resulting in better speech intelligibility performance; or that individuals with higher linguistic knowledge benefit more from

linguistic information, resulting in better intelligibility scores and asynchrony detection, remains unresolved. That we do not find any of these correlations in anomalous sentences or words corroborates the role of the use of linguistic cues in this process and thus shows an intrinsic relationship between the use of linguistic information in both speech understanding and sensitivity to temporal asynchronies.

For anomalous sentences – but not for meaningful sentences – we reported a negative correlation between the AV speech perception benefit and the TBW width, but no correlation between the TBW width and the auditory or synchronous AV intelligibility scores itself - in contrast to meaningful sentences. Yet, there was a strong correlation between the two sentence types for both synchronous AV ($R = 0.65$) and auditory ($R = 0.61$) speech intelligibility performance, but not for the AV benefit ($R = 0.20$).

However, for the asynchrony effect of AV stimuli, not for temporal synchrony judgment, but for speech intelligibility, the opposite shows. Larger drops in AV intelligibility scores caused by stimulus asynchrony for meaningful sentences were positively correlated ($R = 0.50$) to AV benefit, but neither to auditory nor synchronous AV speech intelligibility, whereas larger drops for anomalous sentences correlated to performance on synchronous AV speech intelligibility ($R = 0.69$), but not to AV benefit.

Thus, for both sentence types auditory and synchronous AV speech intelligibility is strongly correlated (A * AV: $R_{sent.mean} = 0.61$; $R_{sent.anom} = 0.65$), and for both sentence types AV benefit is negatively correlated to auditory performance (A* AV ben: $R = -0.68$). Between both sentence types auditory and synchronous AV performance are strongly correlated (meaningful * anomalous sentences: $R_A = 0.61$; $R_{AV} = 0.65$), while the AV benefit (AV - A) is not correlated between the two sentence types (meaningful * anomalous sentences: $R_{AV\_ben} = 0.20$). Furthermore, sensitivity to (a)synchrony detection (i.e., TBW width) is negatively correlated to AV benefit for

anomalous sentences (R = -0.67), but not to meaningful sentences (R = 0.02) This could potentially mean that the AV benefit for meaningful sentences has to be explained by an additional factor that is not present for anomalous sentences. This is neither auditory intelligibility performance – as this relationship to AV benefit is similar for both sentence types, nor the sensitivity to temporal synchrony perception – as it does not explain AV benefit for meaningful sentences. Could it be that the linguistic influence of predictable semantic context for meaningful sentences explains these findings? Anomalous sentences do not have this predictable context information and thus the size of the AV benefit might therefore be mainly defined by auditory performance and sensitivity to temporal synchrony perception. For meaningful sentences however, we can use predictable context to create AV speech perception benefits, and this has a larger influence than sensitivity to temporal asynchrony; therefore, the TBW width is not correlated to AV speech perception benefits for meaningful sentences. Furthermore, as performance for asynchronous stimuli (SOA = 500 ms) was strongly correlated between both sentence types, but the drop in performance caused by temporal asynchrony (AV intelligibility score at 0 ms - AV intelligibility score at 500 ms) was not, it could be argued that for asynchronous stimuli the use of predictable semantic context was limited, and thus individuals who relied heavier on linguistic information for their AV benefit in meaningful sentences were more impacted (i.e., larger drop in performance) by temporal asynchrony.

So, what about words? The correlation between auditory performance and AV speech perception benefit was very strong (R = -0.94), and no relationship with temporal synchrony perception (i.e., TBW) was reported. Therefore, in words, where no predictable semantic context, and limited temporal information is available, perceptual performance might be the main driver to determine the size of AV speech perception benefit, independent of the sensitivity to temporal asynchrony. This reliance on mainly perceptual information for speech intelligibility in words seemed

plausible as the size of the AV benefit was indeed negatively correlated to auditory speech intelligibility and positively correlated to synchronous AV speech intelligibility. Furthermore, the drop in speech intelligibility caused by temporal asynchrony for AV words was strongly correlated to both auditory and AV speech intelligibility, and the AV speech perception benefit.

Thus, whereas for words the AV speech perception benefit might mainly be determined by perceptual performance, especially auditory performance, for anomalous sentences (that provide more temporal information, yet have no predictable semantic context), the AV benefit was correlated less to auditory-only performance. But sensitivity to AV synchrony also had a role. For meaningful sentences sensitivity to temporal synchrony did not influence the AV speech perception benefit anymore, and more likely predictable semantic context influenced AV speech perception benefits on top of auditory-only performance.

Finally, recall that the TBW width between words, anomalous sentences, and meaningful sentences was strongly correlated for individuals (Section 5.2 and 5.3), and so were the three linguistic levels for the auditory speech intelligibility performance, but only the two sentence types for AV speech intelligibility performance, and none of the linguistic levels for AV speech perception benefits (Section 5.5). Combining these observations, we hypothesize that linguistic information, specifically predictable morpho-syntactic structure and semantic context, drive the observed differences in AV speech intelligibility and AV speech perception benefits, but does not explain the large between-subject variability that we reported for both TBW width (Section 5.2 and 5.3) and auditory speech intelligibility performance (Section 5.5).

This complex relationship between temporal and linguistic information on AV integration mechanisms trigger the question what other fundamental features of AV perception (Section 5.4) drive our experience. If we already know where and when to

attend, there is not necessarily a detection benefit of temporal synchronously perceived AV stimuli. Is it then only linguistic information that improves speech intelligibility performance in noise, or does perceived temporal synchrony lead to other fundamental effects that increase AV speech intelligibility? Reports of enhanced loudness perception in addition to the detection benefit for AV stimuli (Lovelace et al., 2003; Odgaard et al., 2004), made us wonder how either temporal (a)synchrony or the linguistic nature of the stimuli contribute. We found that manipulating temporal synchrony led to a change in loudness perception. By mapping this perceived loudness change to a perceived change in TMR, we transformed a highly subjective measure into a quantifiable objective scale (in dB). The change of our scale then also resulted in the reduction of the influences of linguistic complexity on our loudness task. As a result we found that synchronous AV stimuli were perceived +/- 3dB louder than completely asynchronous (500 ms) visual-leading AV stimuli, for all linguistic levels and TMRs used. Furthermore, on a word level this perceived TMR drop was significantly correlated to the intelligibility drop for asynchronous stimuli, showing the value to further explore how this loudness effect links to individual AV speech integration.

Chapter 6

Concluding Remarks

6.1. Synthesis of findings

Over the course of seven experiments, our work has explored integration mechanisms of AV speech perception in both development (typical, and neurodiverse) and adults, through the application of remote psychophysical methods. Contextualizing our findings within the broader literature, we have drawn several conclusions that are noteworthy for the field.

The work in this dissertation has shown that AV integration measures for speech signals can effectively be conducted in a remote format, for children – as young as age four – and adults. By collecting data from 261 distinct individuals in our developmental study, we illustrated the efficiency of this approach in evaluating psychophysical tasks within a sizable participant pool, which can otherwise be challenging to assemble and very time-consuming. Moreover, for adults, this remote format facilitated the seamless collection of data for four tasks distributed over several months, exhibiting a low attrition rate as individuals could complete the tasks in the comfort of their home (Reips, 2002). We want to conclude with three key take-aways for remote AV perceptual work. First, having an initial personal contact (via video conferencing tools or personal emails) regarding the task content, but also for consent, or to answer any questions, can improve the participant's motivation and helps screening out participants that do not fit the inclusion criteria. Second, given the limited control of the environment it helps to build in general attention checks. Third, creating relative rather than absolute measures in background noise helps minimizing differences between participants that inherently follow from different computer set-ups, quality of audio and video, environmental noise or other factors.

For all seven tasks we focused on the influences of prelinguistic and linguistic information on the AV integration process. In children we tried to minimize any differences in linguistic knowledge that go hand in hand with development (Chomsky, 1965), through the use of a closed-set word intelligibility task that utilized vocabulary required at the age of 3. Whereas for adults, we specifically manipulated different levels of linguistic complexity to measure their effect.

In sum, we conclude that linguistic information significantly influences AV speech perception. For children we reported an effect of age for both auditory and AV speech perception (Fort et al., 2012; Ross et al., 2011). Yet, in contrast to previous literature (Desjardins et al., 1997; Maidment et al., 2015; Ross et al., 2011; Wightman et al., 2006), we found no age-effect for the AV speech perception benefits. Thus after minimizing linguistic requirements, but retaining an AV speech intelligibility task, we found no differences in AV speech perception benefits between four- and fifteen-year-olds. However, this did not mean that there were no large intersubject differences. Performance on the auditory speech-in-noise intelligibility task was a better predictor than age to explain these individual differences. Specifically, even when children with high auditory scores still had room for improvement, they would not use AV speech cues to their benefit.

Interestingly, children with developmental dyslexia, who often have phonological processing problems – an auditory deficit – might rely more on the visual stimulus in their decision making (Francisco et al., 2017). We indeed found that children with developmental dyslexia only relied on their auditory speech intelligibility performance for the best SNRs to decide whether or not to use visual cues. Nonetheless, this did not influence the size of their AV speech perception benefits, nor their auditory or AV speech intelligibility scores.

For our word intelligibility task with adults, we similarly found that AV speech perception benefits highly correlated with the listener's auditory performance, and

thus the AV benefit might mainly be determined by perceptual performance. For anomalous sentences (that provide more temporal information, and morpho-syntactic context, yet have no predictable semantic context), the AV benefit was correlated less to auditory-only performance, but a relationship showed between AV benefit and sensitivity to AV synchrony. For meaningful sentences, sensitivity to temporal synchrony did not influence the AV speech perception benefit anymore, and more likely predictable semantic context influenced AV speech perception benefits on top of their auditory-only performance.

Thus, for both children and adults auditory speech perception performance has been shown to be a significant predictor of the size of the AV speech perception benefit perceived for AV speech in background noise. Yet, prior individual experiences (e.g., auditory phonological processing deficits) and the linguistic information provided by the signal influence the weight of the auditory modality.

We did find developmental differences in the AV integration process of children. Older children were more influenced by semantically incongruent AV speech stimuli than younger children. We argued that older children might employ prior information more to make decisions about the AV speech stimuli (i.e., to recalibrate) and therefore be more influenced by the incongruency of both streams (Rohlf et al., 2020). Similarly we found that adults would be influenced by the temporally incongruent visual information, up to an extent that their intelligibility would drop below auditory performance for large temporal stimulus incongruencies.

Furthermore, for all tasks assessed in adults (i.e., temporal synchrony perception, loudness rating, speech intelligibility), we reported significant effects of temporal asynchrony for all linguistic levels. At around 150 - 200 ms asynchrony, less than 75% of the stimuli would be perceived as synchronous (for all linguistic types), the loudness rating would start to drop, and the intelligibility on a sentence level would start to decline. These findings suggest a critical influence of temporal asynchrony

on AV integration. However, whether this translates to AV speech perception intelligibility depends on the stimulus information.

One of the most interesting findings this dissertation has produced is the change in loudness perception, and the degree it varies based on prelinguistic (temporal asynchrony) and linguistic information. We found that for all stimuli tested (words, pseudowords, and reversed words) temporal asynchrony resulted in a 3 dB drop in target-masker-ratio through a loudness perception task. Furthermore, the more complex the linguistic stimulus, the higher the loudness rating for synchronous AV stimuli.

There is a lot more to unpack from these findings, yet we believe they are promising as they provide information about AV integration without directly measuring AV perceptual mechanisms (in contrast to a temporal synchrony or speech intelligibility measure). Furthermore, the loudness task is an intuitive task to use for a wide age range, a task that is not directly influenced by the linguistic knowledge of the individual, and a task that shows effects of both prelinguistic and linguistic information.

6.2. Directions for future research

In light of our results and the broader literature, we can make several proposals for future research. The shift we hope to see is to really focus on both prelinguistic and linguistic contributions in AV integration and frame observed differences from this viewpoint. For example, rather than stating that integration of AV speech differs across development or in certain neurodiverse populations, it is more informative to state which and how linguistic contributions explain these differences, what can be explained by prelinguistic differences (i.e., perceptual weighting, decisions based on priors, temporal processing), or what is unique to the combination of both.

6.2.1. Expansion of our developmental work

Our results pointed to individual differences in integration of AV speech signals that were not driven by changes in age, but rather by unimodal auditory speech perception performance of the listener. This finding also shed some light on the conflicting literature regarding AV integration of speech signals in individuals with developmental dyslexia (e.g., Ramirez & Mann, 2005; Rüsseler et al., 2015; van Laarhoven et al., 2018 versus, Baart et al., 2012; Francisco et al., 2017; Megnin-Viggars & Goswami, 2013). For these reasons, we believe this paradigm has some potential to be further utilized.

We would like to expand this task to other developmental disorders, more specifically children with Autism Spectrum Disorders (ASD), and children with Developmental Language Disorders (DLD). Both of these neurodiverse populations have consistently shown lower AV speech intelligibility scores and lower AV speech perception benefits (e.g., ASD: Smith & Bennetto, 2007; Stevenson et al., 2014c; DLD: Kaganovich et al., 2015). As children with ASD rely more on auditory cues, based on studies with illusions (de Gelder et al., 1991; Stevenson et al., 2014), and potentially use different decision criteria when confronted with incongruent speech (Magnotti & Beauchamp, 2015), our task which is exactly measuring these aspects could inform us how children with ASD approach this multisensory processing for non-illusionary speech signals. Children with DLD have struggle with processing of incongruent stimuli (Norrix et al., 2007), and as linguistic processing is the main disruptor for children with DLD (5th ed.; DSM–5; American Psychiatric Association, 2013), it would be informative to study how they perform on a task that is still relevant for AV speech, yet does not require complex linguistic processing skills. Our remote AV speech perception task further provides the benefit that there is limited to no contact with a new person or a new environment, which might benefit children with ASD. Furthermore it does not require any speech production to obtain AV speech intelligibility scores, which can benefit both children with ASD and DLD.

Thus, expanding this task to other developmental disorders could potentially create a comprehensive story about how children across development weigh the perceptual streams, and use their priors (as suggested by the Causal Inference Model ; Körding et al., 2007).

6.2.2. Loudness perception as a measure of AV integration

Loudness perception in the context of integration of AV speech is barely touched upon (Odgaard et al., 2003; 2004). Yet, our work (Section 5.4 - 5.5) has shown that measuring loudness perception of AV speech in noise can provide information about AV integration without directly measuring the features of interest (in contrast to a temporal synchrony or speech intelligibility measure).

We believe that this measure provides information about linguistic processing of AV speech for synchronous AV stimuli, and about prelinguistic processing of AV speech when manipulating important prelinguistic characteristics like temporal synchrony perception. A measure of loudness perception stands orthogonal to either linguistic or prelinguistic manipulations and therefore makes it a useful tool to untangle AV integration of speech signals (Bizley et al., 2016; Maddox et al., 2015).

First, rating loudness by moving a slider is a concept that does not require highly developed cognitive or computer skills, nor does it actively require to implement linguistic knowledge to make a decision. Therefore this task would be appropriate across a large range of the lifespan (age 3 - 4 and up), without having to make additional task manipulations based on age groups, as there is no bottom or ceiling performance on this task. Therefore, this measure could uniquely contribute to the literature of integration of AV speech across the lifespan.

Second, this task is ideal to assess in neurodiverse populations for similar reasons. Again, it would work both for children and adults. But more interestingly, the ease of the task and the indirect measure of linguistic and prelinguistic contributions in

AV speech perception can help us understand neurodevelopmental disorders without relying on baseline unimodal processing, nor on the language skills of the individual.

Third, by assessing this task to non-native English speakers (i.e., ideally different levels of English language learners), we could learn more about how prior experience with language influences critical features, like a 3 dB loudness enhancement caused by synchronous AV speech, in integration of AV speech.

Although this loudness measure on its own might already provide sufficient information to quantify individual differences in AV integration, we do believe there is benefit from adding both temporal synchrony measures and intelligibility measures to the test battery to gain a more comprehensive understanding.

Finally, there are some task manipulations that could be made to further enhance our understanding of the mechanisms of AV integration:

(1) By adding word strings, anomalous sentences, meaningful sentences, and maybe even short paragraphs to our step-wise linguistic continuum for this task we could better understand the influence of stimulus length, separately from the linguistic changes.
(2) By adding more salient stimulus incongruencies to our linguistic and temporal manipulations (e.g., talker identity mismatch or phoneme-viseme incongruencies), we could identify whether this loudness enhancement is specific to AV object formation. This is especially interesting because the change in loudness perception cannot be "forced" by our tasks requirements, like in temporal synchrony perception (Vatakis & Spence, 2007), or speech intelligibility measures.
(3) By manipulating both the language of the stimulus and the language of the masker, we could identify how prelinguistic and linguistic manipulations lead to enhancement or reduction of the loudness perception, to an increase in

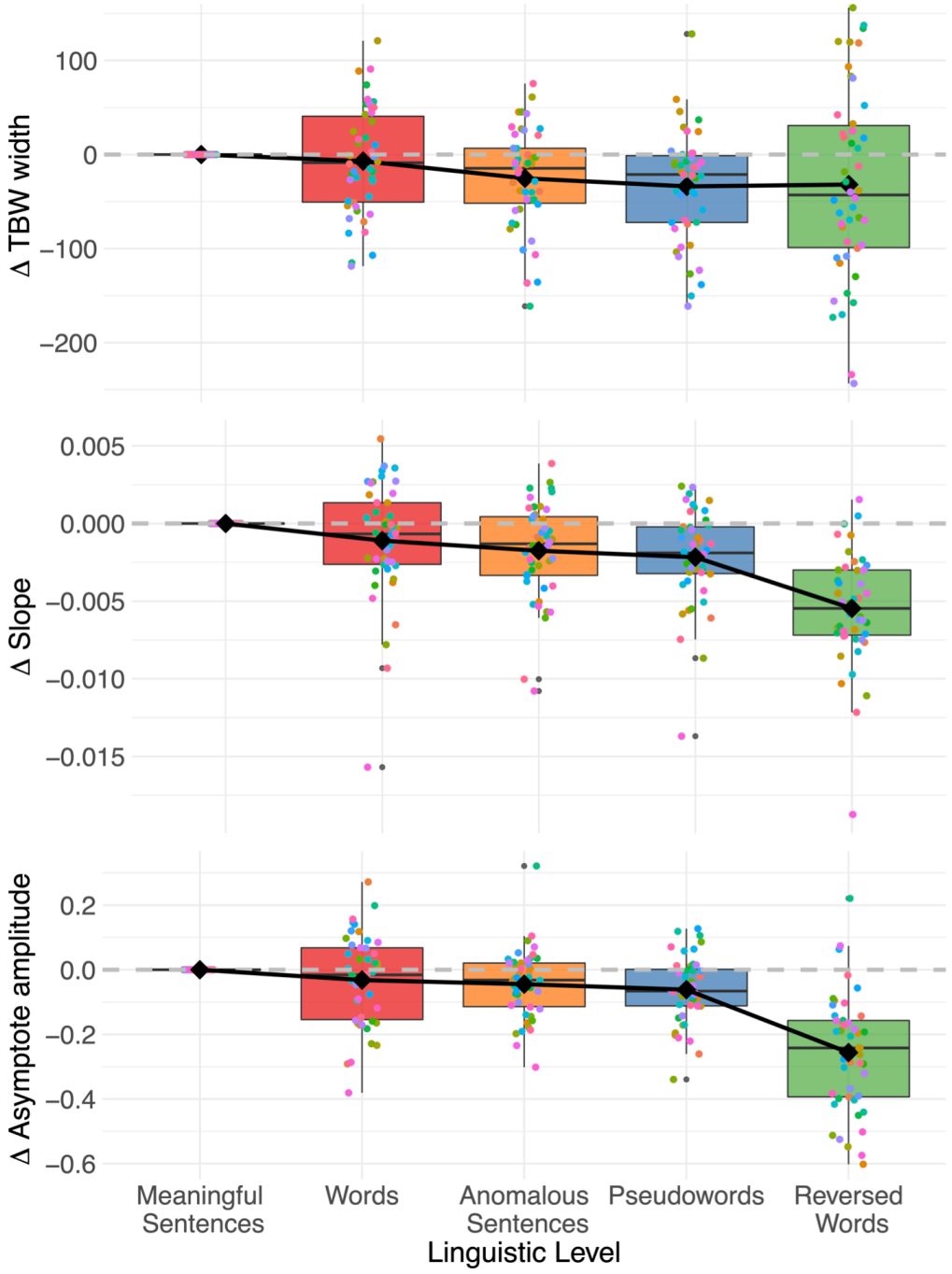
hit-rate or a reduction in false alarm rates (Cappelloni et al., 2023), or to changes in phase-resetting or divisive normalization (van Atteveldt et al., 2014).

"The important thing is not to stop questioning. Curiosity has its own reason for existing."

– Albert Einstein –

Supplemental Figure

FIGURE A. Shift in performance for TBW width, slope and asymptote amplitude, in contrast to meaningful sentences.



Boxplots representing change in TBW width (in ms), slope, or asymptote amplitude compared to meaningful sentences for all other linguistic levels (words, anomalous sentences, pseudowords and reversed words). Mean = black diamond, Mean = horizontal line. The gray dashed line represents the score for meaningful sentences. The colored dots represent individual participants.

Bibliography

Al-Salim, S., Moeller, M. P., & McGregor, K. K. (2020). Performance of Children with Hearing Loss on an Audiovisual Version of a Nonword Repetition Task. *Language, Speech & Hearing Services in Schools*, *51*(1), 42–54. https://doi.org/10.1044/2019_LSHSS-OCHL-19-0016

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology : CB*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. G. R. (1999). The intermodal representation of speech in newborns. *Developmental Science*, *2*(1), 42–46. https://doi.org/10.1111/1467-7687.00052

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *PLoS Biology*, *17*(4), e3000210–e3000210. https://doi.org/10.1371/journal.pbio.3000210

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual Integration of Speech Falters under High Attention Demands. *Current Biology*, *15*(9), 839–843. https://doi.org/10.1016/j.cub.2005.03.046

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited. *Multisensory Research*, *31*(1–2), 111–144. https://doi.org/10.1163/22134808-00002565

Alsius, A., Wayne, R. V., Paré, M., & Munhall, K. G. (2016). High visual resolution matters in audiovisual speech perception, but only for some. *Attention, Perception & Psychophysics*, *78*(5), 1472–1487. https://doi.org/10.3758/s13414-016-1109-4

Altieri, N. (2010). *Toward a unified theory of audiovisual integration in speech perception* (Vol. 71, Issue 9). ProQuest Dissertations Publishing.

Altieri, N., & Hudock, D. (2014). Assessing variability in audiovisual speech integration skills using capacity and accuracy measures. *International Journal of Audiology*, *53*(10), 710–718. https://doi.org/10.3109/14992027.2014.909053

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology*, *2*, 238–238. https://doi.org/10.3389/fpsyg.2011.00238

Andrea Hillock-Dunn, & Mark T. Wallace. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence: Audiovisual simultaneity. *Developmental Science*, *15*, 688–696. https://doi.org/10.1111/j.1467-7687.2012.01171.x

Assmann, P. F. (1996). Tracking and glimpsing speech in noise: Role of fundamental frequency. *The Journal of the Acoustical Society of America*, *100*(4_Supplement), 2680–2680. https://doi.org/10.1121/1.416961

Atilgan, H., & Bizley, J. K. (2021). Training enhances the ability of listeners to exploit visual information for auditory scene analysis. *Cognition*, *208*, 104529–104529. https://doi.org/10.1016/j.cognition.2020.104529

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. *Neuron (Cambridge, Mass.)*, *97*(3), 640-655.e4. https://doi.org/10.1016/j.neuron.2017.12.034

Auer, E. T. (2002). The Influence of the Lexicon on Speech Read Word Recognition: Contrasting Segmental and Lexical Distinctiveness. *Psychonomic Bulletin & Review*, *9*(2), 341–347. https://doi.org/10.3758/BF03196291

Baart, M., de Boer-Schellekens, L., & Vroomen, J. (2012). Lipread-induced phonetic recalibration in dyslexia. *Acta Psychologica*, *140*(1), 91–95. https://doi.org/10.1016/j.actpsy.2012.03.003

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, *130*(1), 31–43. https://doi.org/10.1016/j.cognition.2013.09.006

Bahrick, L. E. (1988). Intermodal learning in infancy: learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development*, *59*(1), 197–209. https://doi.org/10.2307/1130402

Bahrick, L. E. (2010). Intermodal Perception and Selective Attention to Intersensory Redundancy: Implications for Typical Social Development and Autism. In *The Wiley-Blackwell Handbook of Infant Development* (pp. 120–166). Wiley-Blackwell. https://doi.org/10.1002/9781444327564.ch4

Bahrick, L. E., & Lickliter, R. (2000). Intersensory Redundancy Guides Attentional Selectivity and Perceptual Learning in Infancy. *Developmental Psychology*, *36*(2), 190–201. https://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., & Lickliter, R. (2001). Audio-Visual Events.

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior*, *30*, 153–187.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. *Multisensory development*, 183-206

Ballingham, T., & Cienkowski, K. M. (2004). Visual enhancement in consonant identification by younger and older adults. *Journal of the Academy of rehabilitative Audiology*, 37, 11-21.

Bargones, J. Y., & Werner, L. A. (1994). Adults Listen Selectively; Infants Do Not. *Psychological Science*, *5*(3), 170–174. https://doi.org/10.1111/j.1467-9280.1994.tb00655.x

Barutchu, A., Danaher, J., Crewther, S. G., Innes-Brown, H., Shivdasani, M. N., & Paolini, A. G. (2010). Audiovisual integration in noise by children and adults. *Journal of Experimental Child Psychology*, *105*(1), 38–50. https://doi.org/10.1016/j.jecp.2009.08.005

Başkent, D., & Bazo, D. (2011). Audiovisual Asynchrony Detection and Speech Intelligibility in Noise With Moderate to Severe Sensorineural Hearing Impairment. *Ear and Hearing*, *32*(5), 582–592. https://doi.org/10.1097/AUD.0b013e31820fca23

Bast, J., & Reitsma, P. (1998). Analyzing the Development of Individual Differences in Terms of Matthew Effects in Reading: Results From a Dutch Longitudinal Study. *Developmental Psychology*, *34*(6), 1373–1399. https://doi.org/10.1037/0012-1649.34.6.1373

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. *The handbook of child language*, *30*, 96-151.

Baum, S. H., & Stevenson, R. A. (2017). Shifts in Audiovisual Processing in Healthy Aging. *Current Behavioral Neuroscience Reports*, *4*(3), 198–208. https://doi.org/10.1007/s40473-017-0124-7

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, *95*, 40–67. https://doi.org/10.1016/j.specom.2017.07.001

Benítez-Barrera, C. R., Grantham, D. W., & Hornsby, B. W. Y. (2020). The Challenge of Listening at Home: Speech and Noise Levels in Homes of Young Children With Hearing Loss. *Ear and Hearing*, *41*(6), 1575–1585. https://doi.org/10.1097/AUD.0000000000000896

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B, Methodological*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berger, C. C., & Ehrsson, H. H. (2013). Mental Imagery Changes Multisensory Perception. *Current Biology*, *23*(14), 1367–1372. https://doi.org/10.1016/j.cub.2013.06.012

Berglund, B., & Preis, A. (1997). Is perceived annoyance more subject-dependent than perceived loudness?. *Acta Acustica united with Acustica*, *83*(2), 313-319

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual Speech Binding: Convergence or Association? In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 203–223). MIT Press.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech Perception without Hearing. *Perception & Psychophysics*, *62*(2), 233–252. https://doi.org/10.3758/BF03205546

Bertelson, P., & De Gelder, B. (2004). The psychology of multimodal perception. *Crossmodal space and crossmodal attention*, 141-177.

Biel, A. L., & Friedrich, E. V. C. (2018). Why You Should Report Bayes Factors in Your Transcranial Brain Stimulation Studies. *Frontiers in Psychology*, *9*, 1125–1125. https://doi.org/10.3389/fpsyg.2018.01125

Binnie, C. A., Jackson, P. L., & Montgomery, A. A. (1976). Visual Intelligibility of Consonants: A Lipreading Screening Test with Implications for Aural Rehabilitation. *The Journal of Speech and Hearing Disorders*, *41*(4), 530–539. https://doi.org/10.1044/jshd.4104.530

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and Visual Contributions to the Perception of Consonants. *Journal of Speech and Hearing Research*, *17*(4), 619–630. https://doi.org/10.1044/jshr.1704.619

Birch, H. G., & Belmont, L. (1964). Auditory-Visual Integration In Normal and Retarded Readers. *American Journal of Orthopsychiatry*, *34*(5), 852–861. https://doi.org/10.1111/j.1939-0025.1964.tb02240.x

Bishop, C. W., & Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PloS One*, *6*(8), e24016–e24016.https://doi.org/10.1371/journal.pone.0024016

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews. Neuroscience*, *14*(10), 693–707. https://doi.org/10.1038/nrn3565

Bizley, J. K., & King, A. J. (2012). What can multisensory processing tell us about the functional organization of auditory cortex?. *The neural bases of multisensory processes*

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends in Neurosciences (Regular Ed.)*, *39*(2), 74–85. https://doi.org/10.1016/j.tins.2015.12.007

Bjorklund, D. (2005). *Children's thinking: Cognitive development and individual differences* (4th ed.). Australia ; Belmont, CA: Thomson/Wadsworth.

Blau, V., Reithler, J., van Atteveldt, N., Seitz, J., Gerretsen, P., Goebel, R., & Blomert, L. (2010). Deviant processing of letters and speech sounds as proximate cause of reading failure: a functional magnetic resonance imaging study of dyslexic children. *Brain (London, England : 1878)*, *133*(3), 868–879. https://doi.org/10.1093/brain/awp308

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [Computer program]. Version 6.1.41, retrieved 25 March 2021 from http://www.praat.org/

Boets, B., Vandermosten, M., Poelmans, H., Luts, H., Wouters, J., & Ghesquière, P. (2011). Preschool impairments in auditory processing and speech perception uniquely predict future reading problems. *Research in Developmental Disabilities*, *32*(2), 560–570. https://doi.org/10.1016/j.ridd.2010.12.020

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of Spoken Words by Native and Non-Native Listeners: Talker-, Listener-, and Item-Related Factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. https://doi.org/10.1121/1.427952

Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking Clearly for Children With Learning Disabilities: Sentence Perception in Noise. *Journal of Speech, Language, and Hearing Research*, *46*(1), 80–97. https://doi.org/10.1044/1092-4388(2003/007)

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (1st ed.). The MIT Press. https://doi.org/10.7551/mitpress/1486.001.0001

Brooks, C. J., Chan, Y. M., Anderson, A. J., & McKendrick, A. M. (2018). Audiovisual Temporal Perception in Aging: The Role of Multisensory Integration and Age-Related Sensory Loss. *Frontiers in Human Neuroscience*, *12*, 192–192. https://doi.org/10.3389/fnhum.2018.00192

Buchwald, A. B., Winters, S. J., & Pisoni, D. B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*(4), 580–610. https://doi.org/10.1080/01690960802536357

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. https://doi.org/10.1002/dev.20032

Buss, E., Hall, J. W., & Grose, J. H. (2011). Development of auditory coding as reflected in psychophysical performance. In *Human auditory development* (Springer Handbook of Auditory Research, pp. 107–136). New York, NY: Springer New York.

Buss, E., Leibold, L. J., & Hall, J. W. (2016). Effect of response context and masker type on word recognition in school-age children and adults. *The Journal of the Acoustical Society of America*, *140*(2), 968–977. https://doi.org/10.1121/1.4960587

Cabeza, R., Anderson, N. D., Locantore, J. K., & McIntosh, A. R. (2002). Aging Gracefully: Compensatory Brain Activity in High-Performing Older Adults. *NeuroImage (Orlando, Fla.)*, *17*(3), 1394–1402. https://doi.org/10.1006/nimg.2002.1280

Calvert, G. A. (2001). Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex (New York, N.Y. 1991)*, *11*(12), 1110–1123. https://doi.org/10.1093/cercor/11.12.1110

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campbell, R., Iversen, S. D., & David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, *10*(12), 2619–2623. https://doi.org/10.1097/00001756-199908200-00033

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Steven C. R. Williams, McGuire, P. K., Peter W. R. Woodruff, Iversen, S. D., & David, A. S. (1997). Activation of Auditory Cortex During Silent Lipreading. *Science (American Association for the Advancement of Science)*, *276*(5312), 593–596. https://doi.org/10.1126/science.276.5312.593

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

Campbell, C. S., & Massaro, D. W. (1997). Perception of Visible Speech: Influence of Spatial Quantization. *Perception (London)*, *26*(5), 627–644. https://doi.org/10.1068/p260627

Campbell, R. (1998). Speechreading: Advances in Understanding its Cortical Bases and Implications for Deafness and Speech Rehabilitation. *Scandinavian Audiology*, *27*(4), 80–86. https://doi.org/10.1080/010503998420694

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, *32*(1), 85–99. https://doi.org/10.1080/00335558008248235

Cappelloni, M. S., Mateo, V. S., & Maddox, R. K. (2023). Performance in an Audiovisual Selective Attention Task Using Speech-Like Stimuli Depends on the Talker Identities, But Not Temporal Coherence. *Trends in Hearing*, *27*,. https://doi.org/10.1177/23312165231207235

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal Inference in the Multisensory Brain. *Neuron (Cambridge, Mass.)*, *102*(5), 1076-1087.e8. https://doi.org/10.1016/j.neuron.2019.03.043

Cappelloni, M. S., Shivkumar, S., Haefner, R. M., & Maddox, R. K. (2019). Task-uninformative visual stimuli improve auditory spatial discrimination in humans but not the ideal observer. *PloS One*, *14*(9), e0215417–e0215417. https://doi.org/10.1371/journal.pone.0215417

Carroll, J. M., & Snowling, M. J. (2004). Language and phonological skills in children at high risk of reading difficulties. *Journal of Child Psychology and Psychiatry*, *45*(3), 631–640. https://doi.org/10.1111/j.1469-7610.2004.00252.x

Castellanos, I., Vaillant-Molina, M., Lickliter, R., & Bahrick, L. E. (2006). Intersensory redundancy educates infants' attention to amodal information in unimodal stimulation. *Poster presented at the International Society for Developmental Psychobiology, Atlanta, GA.*

Cecere, R., Gross, J., Willis, A., & Thut, G. (2017). Being First Matters: Topographical Representational Similarity Analysis of ERP Signals Reveals Separate Networks for Audiovisual Temporal Binding Depending on the Leading Sense. *The Journal of Neuroscience*, *37*(21), 5274–5287. https://doi.org/10.1523/JNEUROSCI.2926-16.2017

Chan, Y. M., Pianta, M. J., & McKendrick, A. M. (2014). Older age results in difficulties separating auditory and visual signals in time. *Journal of Vision (Charlottesville, Va.)*, *14*(11), 13–13. https://doi.org/10.1167/14.11.13

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech: e1000436. *PLoS Computational Biology*, *5*(7). https://doi.org/10.1371/journal.pcbi.1000436

Chang, N.-Y. N., Uchanski, R. M., & Hullar, T. E. (2012). Temporal integration of auditory and vestibular stimuli. *The Laryngoscope*, *122*(6), 1379–1384. https://doi.org/10.1002/lary.23329

Chen, A., Wijnen, F., Koster, C., & Schnack, H. (2017). Individualized Early Prediction of Familial Risk of Dyslexia: A Study of Infant Vocabulary Development. *Frontiers in Psychology*, *8*, 156–156. https://doi.org/10.3389/fpsyg.2017.00156

Chomsky, N. (1965). Persistent Topics in Linguistic Theory. *Diogenes (English Ed.)*, *13*(51), 13–20. https://doi.org/10.1177/039219216501305102

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-Visual Speech Perception and Aging. *Ear and Hearing*, *23*(5), 439–449. https://doi.org/10.1097/00003446-200210000-00006

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of Open-Set and Closed-Set Task Demands on Spoken Word Recognition. *Journal of the American Academy of Audiology*, *17*(5), 331–349. https://doi.org/10.3766/jaaa.17.5.4

Conrey, B., & Pisoni, D. B. (2006). Auditory-Visual Speech Perception and Synchrony Detection for Speech and Nonspeech Signals. *The Journal of the Acoustical Society of America*, *119*(6), 4065–4073. https://doi.org/10.1121/1.2195091

Conway, C. M., & Christiansen, M. H. (2009). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology*, *21*(4), 561–580. https://doi.org/10.1080/09541440802097951

Coplan, J., & Gleason, J. R. (1988). Unclear speech: recognition and significance of unintelligible speech in preschool children. *Pediatrics (Evanston)*, *82*(3), 447–452.

Corbin, N. E., Bonino, A. Y., Buss, E., & Leibold, L. J. (2016). Development of Open-Set Word Recognition in Children: Speech-Shaped Noise and Two-Talker Speech Maskers. *Ear and Hearing*, *37*(1), 55–63. https://doi.org/10.1097/AUD.0000000000000201

Cox, C. M. M., Keren-Portnoy, T., Roepstorff, A., & Fusaroli, R. (2022). A Bayesian meta-analysis of infants' ability to perceive audio–visual congruence for speech. *Infancy*, *27*(1), 67–96. https://doi.org/10.1111/infa.12436

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *Quarterly Journal of Experimental Psychology (2006)*, *63*(2), 260–274. https://doi.org/10.1080/17470210902888809

Cunningham, J., Nicol, T., Zecker, S. G., Bradlow, A., & Kraus, N. (2001). Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement. *Clinical Neurophysiology*, *112*(5), 758–767. https://doi.org/10.1016/S1388-2457(01)00465-5

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Ooijen, B. van. (2000). Constraints of Vowels and Consonants on Lexical Selection: Cross-Linguistic Comparisons. *Memory & Cognition*, *28*(5), 746–755. https://doi.org/10.3758/BF03198409

Davies, R., Kidd, E., & Lander, K. (2009). Investigating the psycholinguistic correlates of speechreading in preschool age children. *International Journal of Language & Communication Disorders*, *44*(2), 164–174. https://doi.org/10.1080/13682820801997189

de Boer-Schellekens, L., Eussen, M., & Vroomen, J. (2013). Diminished sensitivity of audiovisual temporal order in autism spectrum disorder. *Frontiers in Integrative Neuroscience*, *7*, 8–8. https://doi.org/10.3389/fnint.2013.00008

de Gelder, B., & Vroomen, J. (1998). Impaired Speech Perception in Poor Readers: Evidence from Hearing and Speech Reading. *Brain and Language*, *64*(3), 269–281. https://doi.org/10.1006/brln.1998.1973

de Gelder, B., Vroomen, J., & Bertelson, P. (1998). Upright but not inverted faces modify the perception of emotion in the voice. *Current Psychology of Cognition*, *17*, 1021-1032.

de Gelder, B. de, Vroomen, J., & van der Heide, L. (1991). Face recognition and lip-reading in autism. *European Journal of Cognitive Psychology*, *3*(1), 69–86. https://doi.org/10.1080/09541449108406220

DeLoss, D. J., & Andersen, G. J. (2015). Aging, Spatial Disparity, and the Sound-Induced Flash Illusion. *PloS One*, *10*(11), e0143773–e0143773. https://doi.org/10.1371/journal.pone.0143773

Déry, C., Campbell, N. K. J., Lifshitz, M., & Raz, A. (2014). Suggestion overrides automatic audiovisual integration. *Consciousness and Cognition*, *24*, 33–37. https://doi.org/10.1016/j.concog.2013.12.010

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Desjardins, R. N., Rogers, J., & Werker, J. F. (1997). An Exploration of Why Preschoolers Perform Differently Than Do Adults in Audiovisual Speech Perception Tasks. *Journal of Experimental Child Psychology*, *66*(1), 85–110. https://doi.org/10.1006/jecp.1997.2379

Devergie, A., Grimault, N., Gaudrain, E., Healy, E. W., & Berthommier, F. (2011). The effect of lip-reading on primary stream segregation. *The Journal of the Acoustical Society of America*, *130*(1), 283–291. https://doi.org/10.1121/1.3592223

Dey, A., & Sommers, M. S. (2015). Age-Related Differences in Inhibitory Control Predict Audiovisual Speech Perception. *Psychology and Aging*, *30*(3), 634–646. https://doi.org/10.1037/pag0000033

Dias, J. W., McClaskey, C. M., & Harris, K. C. (2021). Audiovisual speech is more than the sum of its parts: Auditory-visual superadditivity compensates for age-related declines in audible and lipread speech intelligibility. *Psychology and Aging*, *36*(4), 520–530. https://doi.org/10.1037/pag0000613

Ding, N., & Simon, J. Z. (2011). Cortical neural coding of speech in simple and complex auditory scenes. *The Journal of the Acoustical Society of America*, *129*(4_Supplement), 2383–2383. https://doi.org/10.1121/1.3587731

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Dixon, N. F., & Spitz, L. (1980). The Detection of Auditory Visual Desynchrony. *Perception (London)*, *9*(6), 719–721. https://doi.org/10.1068/p090719

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Dodd, B. J. (1972). Effects of social and vocal stimulation on infant babbling. *Developmental Psychology*, *7*(1), 80–83. https://doi.org/10.1037/h0032694

Dodd, B., & Burnham, D. (1988). Processing Speechread Information. *Volta Review*, *90*(5), 45-.

Dodd, B., Mcintosh, B., Erdener, D., & Burnham, D. (2008). Perception of the auditory-visual illusion in speech perception by children with phonological disorders. *Clinical Linguistics & Phonetics*, *22*(1), 69–82. https://doi.org/10.1080/02699200701660100

Donohue, S. E., Green, J. J., & Woldorff, M. G. (2015). The effects of attention on the temporal integration of multisensory stimuli. *Frontiers in Integrative Neuroscience*, *9*, 32–32. https://doi.org/10.3389/fnint.2015.00032

Dubas, C., Porter, H., McCreery, R. W., Buss, E., & Leibold, L. J. (2023). Speech-in-speech recognition in preschoolers. *International Journal of Audiology*, *62*(3), 261–268. https://doi.org/10.1080/14992027.2022.2035833

Duffy, M. E. (2002). Methodological Issues In Web-based Research. *Journal of Nursing Scholarship*, *34*(1), 83–88. https://doi.org/10.1111/j.1547-5069.2002.00083.x

Dye, M. W. G., & Bavelier, D. (2010). Differential development of visual attention skills in school-age children: Perceptual Learning. Part II. *Vision Research (Oxford)*, *50*(4), 452–459.

Eden, G. F., Stein, J. F., Wood, M. H., & Wood, F. B. (1995). Verbal and visual problems in reading disability. *Journal of Learning Disabilities*, *28*(5), 272-. https://doi.org/10.1177/002221949502800503

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. (1961). Operating Characteristics, Signal Detectability, and the Method of Free Response. *The Journal of the Acoustical Society of America*, *33*(8), 993–1007. https://doi.org/10.1121/1.1908935

Elliott, L. L. (1979). Performance of Children Aged 9 to 17 Years on a Test of Speech Intelligibility in Noise Using Sentence Material with Controlled Word Predictability. *The Journal of the Acoustical Society of America*, *66*(3), 651–653. https://doi.org/10.1121/1.383691

Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing statistical learning across perceptual modalities in infancy: An investigation of underlying learning mechanism(s). *Developmental Science*, *22*(6), e12847-n/a. https://doi.org/10.1111/desc.12847

Englund, N., & Behne, D. M. (2020). Perception of audiovisual infant directed speech. *Scandinavian Journal of Psychology*, *61*(2), 218–226. https://doi.org/10.1111/sjop.12599

Eramudugolla, R., Henderson, R., & Mattingley, J. B. (2011). Effects of audio–visual integration on the detection of masked speech and non-speech sounds. *Brain and Cognition*, *75*(1), 60–66. https://doi.org/10.1016/j.bandc.2010.09.005

Erber, N. P. (1969). Interaction of Audition and Vision in the Recognition of Oral Speech Stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. https://doi.org/10.1044/jshr.1202.423

Erber, N. P. (1975). Auditory-Visual Perception of Speech. *The Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Erber, N. P. (1979). Auditory-Visual Perception of Speech with Reduced Optical Clarity. *Journal of Speech and Hearing Research*, *22*(2), 212–223. https://doi.org/10.1044/jshr.2202.212

Erber, N. P. (1982). "Glenondale auditory screening procedure," in Auditory Training (Alexander Graham Bell Association, Washington, DC), pp.47–71.

Erdener, D. (2016). Basic to applied research: the benefits of audio-visual speech perception research in teaching foreign languages. *Language Learning Journal*, *44*(1), 124–132. https://doi.org/10.1080/09571736.2012.724080

Erdener, D., & Burnham, D. (2018). Auditory–visual speech perception in three- and four-year-olds and its relationship to perceptual attunement and receptive vocabulary. *Journal of Child Language*, *45*(2), 273–289. https://doi.org/10.1017/S0305000917000174

Erickson, L. C., & Newman, R. S. (2017). Influences of Background Noise on Infants and Children. *Current Directions in Psychological Science : A Journal of the American Psychological Society*, *26*(5), 451–457. https://doi.org/10.1177/0963721417709087

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162-. https://doi.org/10.1016/j.tics.2004.02.002

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental Brain Research*, *208*(3), 447–457. https://doi.org/10.1007/s00221-010-2495-9

Facoetti, A., Lorusso, M. L., Paganoni, P., Cattaneo, C., Galli, R., Umiltà, C., & Mascetti, G. G. (2003). Auditory and visual automatic attention deficits in developmental dyslexia. *Brain Research. Cognitive Brain Research*, *16*(2), 185–191. https://doi.org/10.1016/S0926-6410(02)00270-7

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews. Neuroscience*, *9*(4), 292–303. https://doi.org/10.1038/nrn2258

Fantz, R. L. (1963). Pattern Vision in Newborn Infants. *Science (American Association for the Advancement of Science)*, *140*(3564), 296–297. https://doi.org/10.1126/science.140.3564.296

Fernald, A. (1992). Human maternal vocalizations to infants as biologically relevant signals - an evolutionary perspective. In *The adapted mind - evolutionary psychology and the generation of culture* (pp. 391–428).

Fiscella, S., Cappelloni, M. S., & Maddox, R. K. (2022). Independent mechanisms of temporal and linguistic cue correspondence benefiting audiovisual speech processing. *Attention, Perception & Psychophysics*, *84*(6), 2016–2026. https://doi.org/10.3758/s13414-022-02440-3

Fiser, J., & Aslin, R. N. (2002). Statistical Learning of New Visual Feature Combinations by Infants. *Proceedings of the National Academy of Sciences - PNAS*, *99*(24), 15822–15826. https://doi.org/10.1073/pnas.232472899

Fisher, C. G. (1968). Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. https://doi.org/10.1044/jshr.1104.796

Flanagan, L., Zumbrunn, N. M., Hirst, R., & McGovern, D. (2023). Assessing the reliability of an online measure of the temporal binding window of audiovisual integration.

Fleming, J. T., Maddox, R. K., & Shinn-Cunningham, B. G. (2021). Spatial alignment between faces and voices improves selective attention to audio-visual speech. *The Journal of the Acoustical Society of America*, *150*(4), 3085–3100. https://doi.org/10.1121/10.0006415

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2010). The word superiority effect in audiovisual speech perception. *Speech Communication*, *52*(6), 525–532. https://doi.org/10.1016/j.specom.2010.02.005

Fort, M., Spinelli, E., Savariaux, C., & Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *International Journal of Behavioral Development*, *36*(6), 457–467. https://doi.org/10.1177/0165025412447752

Fox, E. (1994). Grapheme-phoneme correspondence in dyslexic and matched control readers. *The British Journal of Psychology*, *85*(1), 41–53. https://doi.org/10.1111/j.2044-8295.1994.tb02507.x

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H.-P., Russo, N. N., Blanco, D., Saint-Amour, D., & Ross, L. A. (2015). Severe multisensory speech integration deficits in high-functioning school-aged children with Autism Spectrum Disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex (New York, N.Y. 1991)*, *25*(2), 298–312. https://doi.org/10.1093/cercor/bht213

Francisco, A. A., Groen, M. A., Jesse, A., & McQueen, J. M. (2017). Beyond the usual cognitive suspects: The importance of speechreading and audiovisual temporal sensitivity in reading ability. *Learning and Individual Differences*, *54*, 60–72. https://doi.org/10.1016/j.lindif.2017.01.003

Freeman, E. D., & Ipser, A. (2016). Individual differences in multisensory integration and timing. *Electronic Imaging*, *28*(16), 1–4. https://doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-097

Freiherr, J., Lundström, J. N., Habel, U., & Reetz, K. (2013). Multisensory integration mechanisms during aging. *Frontiers in Human Neuroscience*, *7*, 863–863. https://doi.org/10.3389/fnhum.2013.00863

Frissen, I., Vroomen, J., de Gelder, B., & Bertelson, P. (2005). The aftereffects of ventriloquism: Generalization across sound-frequencies. *Acta Psychologica*, *118*(1–2), 93–100. https://doi.org/10.1016/j.actpsy.2004.10.004

Froyen, D., Van Atteveldt, N., Bonte, M., & Blomert, L. (2008). Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds. *Neuroscience Letters*, *430*(1), 23–28. https://doi.org/10.1016/j.neulet.2007.10.014

Frtusova, J. B., & Phillips, N. A. (2016). The Auditory-Visual Speech Benefit on Working Memory in Older Adults with Hearing Impairment. *Frontiers in Psychology*, *7*, 490–490. https://doi.org/10.3389/fpsyg.2016.00490

Frtusova, J. B., Winneke, A. H., & Phillips, N. A. (2013). ERP Evidence That Auditory-Visual Speech Facilitates Working Memory in Younger and Older Adults. *Psychology and Aging*, *28*(2), 481–494. https://doi.org/10.1037/a0031243

Fucci, D., Bettagere, R., Gonzales, M. D., Reynolds, M. E., & Petrosino, L. (1995). Language Familiarity in Magnitude-Estimation Scaling of Loudness by Young Adults. *Perceptual and Motor Skills*, *80*(2), 419–423. https://doi.org/10.2466/pms.1995.80.2.419

Fujisaki, W., & Nishida, S. (2009). Audio—tactile superiority over visuo—tactile and audio—visual combinations in the temporal resolution of synchrony perception: Crossmodal Processing. *Experimental Brain Research*, *198*(2–3), 245–259.

Fuster-Duran, A. (1996). Perception of conflicting audio-visual speech: An examination across Spanish and German. *Speechreading by Humans and Machines: Models, Systems, and Applications*, 135-143.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology*, *22*(7), 615–621. https://doi.org/10.1016/j.cub.2012.02.015

Ganesh, A. C., Berthommier, F., & Schwartz, J. (2018). Audiovisual Binding for Speech Perception in Noise and in Aging. *Language Learning*, *68*(S1), 193–220. https://doi.org/10.1111/lang.12271

Gelbar, N. W., Bray, M., Kehle, T. J., Madaus, J. W., & Makel, C. (2018). Exploring the Nature of Compensation Strategies in Individuals With Dyslexia. *Canadian Journal of School Psychology*, *33*(2), 110–124. https://doi.org/10.1177/0829573516677187

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.

Gibson, F., & Twycross, A. (2008). Editorial: Getting it right for children and young people's health care services. *Journal of Clinical Nursing*, *17*(23), 3081–3082. https://doi.org/10.1111/j.1365-2702.2008.02644.x

Gibson, J. J. (James J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.

Gijbels L., Yeatman J.D., Lalonde K., Doering P., Lee, A.KC (in revision; 2024) Audiovisual speech perception benefits are stable from preschool through adolescence. *Multisensory Research*

Gijbels, L., & Lee, A. K. (2023). How moderation affects remote psychophysical tasks with children. *JASA Express Letters*, *3*(1), 014401–014401.https://doi.org/10.1121/10.0016832

Gijbels, L., Cai, R., Donnelly, P. M., & Kuhl, P. K. (2021). Designing Virtual, Moderated Studies of Early Childhood Development. *Frontiers in Psychology*, *12*, 740290–740290. https://doi.org/10.3389/fpsyg.2021.740290

Gijbels, L., Lee, A. K. C., & Yeatman, J. D. (2024). Children with developmental dyslexia have equivalent audiovisual speech perception performance but their perceptual weights differ. *Developmental Science*, *27*(1), e13431-n/a. https://doi.org/10.1111/desc.13431

Gijbels, L., Yeatman, J. D., Lalonde, K., & Lee, A. K. C. (2021). Audiovisual Speech Processing in Relationship to Phonological and Vocabulary Skills in First Graders. *Journal of Speech, Language, and Hearing Research*, *64*(12), 5022–5040. https://doi.org/10.1044/2021_JSLHR-21-00196

Gillmeister, H., & Eimer, M. (2007). Tactile enhancement of auditory detection and perceived loudness. *Brain Research*, *1160*, 58–68. https://doi.org/10.1016/j.brainres.2007.03.041

Giordano, B. L., Ince, R. A. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *eLife*, *6*. https://doi.org/10.7554/eLife.24763

Goldinger, S. D. (2007, August). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th international congress of phonetic sciences* (pp. 49-54).

Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism (Cambridge, England)*, *12*(3), 385–392. https://doi.org/10.1017/S1366728909990150

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Willison, H. M., & Freund, M. S. (2017). Recognition of asynchronous auditory-visual speech by younger and older listeners: A preliminary study. *The Journal of the Acoustical Society of America*, *142*(1), 151–159. https://doi.org/10.1121/1.4992026

Gordon, M. S., & Allen, S. (2009). Audiovisual Speech in Older and Younger Adults: Integrating a Distorted Visual Signal With Speech in Noise. *Experimental Aging Research*, *35*(2), 202–219. https://doi.org/10.1080/03610730902720398

Gori, M., Campus, C., & Cappagli, G. (2021). Late development of audio-visual integration in the vertical plane. *Current Research in Behavioral Sciences*, *2*, 100043-. https://doi.org/10.1016/j.crbeha.2021.100043

Grant, K. W., & Bernstein, J. G. W. (2019). Toward a model of auditory-visual speech intelligibility. In *Multisensory processes* (Springer Handbook of Auditory Research, pp. 33–57). Cham: Springer International Publishing.

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In *AVSP 2001-International Conference on Auditory-Visual Speech Processing*.

Grant, K. W., & Seitz, P. F. (1998). Measures of Auditory-Visual Integration in Nonsense Syllables and Sentences. *The Journal of the Acoustical Society of America*, *104*(4), 2438–2450. https://doi.org/10.1121/1.423751

Grant, K. W., & Seitz, P.-F. (2000). The Use of Visible Speech Cues for Improving Auditory Detection of Spoken Sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.1288668

Grant, K. W., & Walden, B. E. (1996). Evaluating the Articulation Index for Auditory-Visual Consonant Recognition. *The Journal of the Acoustical Society of America*, *100*(4), 2415–2424. https://doi.org/10.1121/1.417950

Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America*, *121*(2), 1164–1176. https://doi.org/10.1121/1.2405859

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-Visual Speech Recognition by Hearing-Impaired Subjects: Consonant Recognition, Sentence Recognition, and Auditory-Visual Integration. *The Journal of the Acoustical Society of America*, *103*(5), 2677–2690. https://doi.org/10.1121/1.422788

Grant, K. W., Wassenhove, V. van, & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Communication*, *44*(1–4), 43–53. https://doi.org/10.1016/j.specom.2004.06.004

Grassegger, H. (1995). McGurk effect in German and Hungarian listeners. In *proceedings of the international congress of phonetic sciences, Stockholm* (Vol. 4, No. 210, p. 13)

Greenberg, S. (1999). Speaking in shorthand : A syllable-centric perspective for understanding pronunciation variation: Special issue on modeling pronunciation variation foe automatic speech recognition. *Speech Communication*, *29*(2–4), 159–176.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews. Neuroscience*, *5*(11), 887–892. https://doi.org/10.1038/nrn1538

Groen, M. A., & Jesse, A. (2013). Audiovisual speech perception in children and adolescents with developmental dyslexia: No deficit with McGurk stimuli.

Guellaï, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' Sensitivity to the Visual Aspects of Infant-Directed Speech: Evidence From Point-Line Displays of Talking Faces. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(9), 1275–1281. https://doi.org/10.1037/xhp0000208

Guerreiro, M. J. S., Murphy, D. R., & Van Gerven, P. W. M. (2013). Making sense of age-related distractibility: The critical role of sensory modality. *Acta Psychologica*, *142*(2), 184–194. https://doi.org/10.1016/j.actpsy.2012.11.007

Grieco-Calub, T. M., Gordon, K. R., Lalonde, K., Cortez, D. M., Lowry, S. L., & Dwyer, G. A. (2023). Effects of face masks on novel word learning in preschool children. *The Journal of the Acoustical Society of America*, *153*(3_supplement), A329–A329. https://doi.org/10.1121/10.0019030

Hairston, W. D., Wallace, M. T., Vaughan, J. W., Stein, B. E., Norris, J. L., & Schirillo, J. A. (2003). Visual Localization Ability Influences Cross-Modal Bias. *Journal of Cognitive Neuroscience*, *15*(1), 20–29. https://doi.org/10.1162/089892903321107792

Hämäläinen, J. A., Salminen, H. K., & Leppänen, P. H. T. (2013). Basic Auditory Processing Deficits in Dyslexia: Systematic Review of the Behavioral and Event-Related Potential/ Field Evidence. *Journal of Learning Disabilities*, *46*(5), 413–427. https://doi.org/10.1177/0022219411436213

Hasher, L., Zacks, R. T., & Rahhal, T. A. (1999). Timing, Instructions, and Inhibitory Control: Some Missing Factors in the Age and Memory Debate. *Gerontology (Basel)*, *45*(6), 355–357. https://doi.org/10.1159/000022121

Hassall, J. R. Zaveri, 1988. *Acoustic Noise Measurements. Brüel & Kjaer documentation*

Hay-McCutcheon, M. J., Pisoni, D. B., & Hunt, K. K. (2009). Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis. *International Journal of Audiology*, *48*(6), 321–333. https://doi.org/10.1080/14992020802644871

Hayes, E. A., Tiippana, K., Nicol, T. G., Sams, M., & Kraus, N. (2003). Integration of heard and seen speech: a factor in learning disabilities in children. *Neuroscience Letters*, *351*(1), 46–50. https://doi.org/10.1016/S0304-3940(03)00971-6

Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio-visual speech perception. *Frontiers in Psychology*, *5*, 698–698. https://doi.org/10.3389/fpsyg.2014.00698

Heikkilä, J., Lonka, E., Ahola, S., Meronen, A., & Tiippana, K. (2017). Lipreading Ability and Its Cognitive Correlates in Typically Developing Children and Children with Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *60*(3), 485–493. https://doi.org/10.1044/2016_JSLHR-S-15-0071

Heinrich, A., & Knight, S. (2016). The contribution of auditory and cognitive factors to intelligibility of words and sentences in noise. In *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (pp. 37-45). Springer International Publishing

Helfer, K. S. (1998). Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology*, *9*(3), 234–242.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556-. https://doi.org/10.31234/osf.io/fqr49

Hewson, C. M., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 186–191. https://doi.org/10.3758/BF03204763

Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: Age-related changes in multisensory temporal processing. *Neuropsychologia*, *49*(3), 461–467. https://doi.org/10.1016/j.neuropsychologia.2010.11.041

Hirst, R. J., McGovern, D. P., Setti, A., Shams, L., & Newell, F. N. (2020). What you see is what you hear: Twenty years of research using the Sound-Induced Flash Illusion. *Neuroscience and Biobehavioral Reviews*, *118*, 759–774. https://doi.org/10.1016/j.neubiorev.2020.09.006

Hockley, N. S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *The Journal of the Acoustical Society of America*, *96*(5_Supplement), 3309–3309. https://doi.org/10.1121/1.410782

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' Use of Synchronized Visual Information to Separate Streams of Speech. *Child Development*, *76*(3), 598–613. https://doi.org/10.1111/j.1467-8624.2005.00866.x

Holt, R. F., Beer, J., Kronenberger, W. G., & Pisoni, D. B. (2013). Developmental effects of family environment on outcomes in pediatric cochlear implant recipients. *Otology & Neurotology*, *34*(3), 388–395. https://doi.org/10.1097/MAO.0b013e318277a0af

Holt, R. F., Kirk, K. I., & Hay-McCutcheon, M. (2011). Assessing Multimodal Spoken Word-in-Sentence Recognition in Children with Normal Hearing and Children with Cochlear Implants. *Journal of Speech, Language, and Hearing Research*, *54*(2), 632–657. https://doi.org/10.1044/1092-4388(2010/09-0148)

Holt, R., Bruggeman, L., & Demuth, K. (2020). Visual speech cues speed processing and reduce effort for children listening in quiet and noise. *Applied Psycholinguistics*, *41*(4), 933–961. https://doi.org/10.1017/S0142716420000302

Huang-Pollock, C. L., Carr, T. H., & Nigg, J. T. (2002). Development of Selective Attention: Perceptual Load Influences Early Versus Late Attentional Selection in Children and Adults. *Developmental Psychology*, *38*(3), 363–375. https://doi.org/10.1037/0012-1649.38.3.363

Humes, L. E., & Christopherson, L. (1991). Speech Identification Difficulties of Hearing-Impaired Elderly Persons: The Contributions of Auditory Processing Deficits. *Journal of Speech and Hearing Research*, *34*(3), 686–693. https://doi.org/10.1044/jshr.3403.686

Huyse, A., Leybaert, J., & Berthommier, F. (2014). Effects of aging on audio-visual speech integration. *The Journal of the Acoustical Society of America*, *136*(4), 1918–1931. https://doi.org/10.1121/1.4894685

Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. (2010). Visual influences on speech perception in children with autism. *Autism : The International Journal of Research and Practice*, *14*(4), 305–320. https://doi.org/10.1177/1362361309353615

Jacoby, W. G. (2000). Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, *19*(4), 577–613. https://doi.org/10.1016/S0261-3794(99)00028-1

Jaime M., Bahrick L.E., & Lickliter R.. (2010). The Critical Role of Temporal Synchrony in the Salience of Intersensory Redundancy During Prenatal Development: CRITICAL ROLE OF TEMPORAL SYNCHRONY. *Infancy*, *15*, 61–82. https://doi.org/10.1111/j.1532-7078.2009.00008.x

Jerger, J., Speaks, C., & Trammell, J. L. (1968). A new approach to speech audiometry. *The Journal of Speech and Hearing Disorders*, *33*(4), 318–328. https://doi.org/10.1044/jshd.3304.318

Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., & Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: A new multimodal picture–word task. *Journal of Experimental Child Psychology*, *102*(1), 40–59. https://doi.org/10.1016/j.jecp.2008.08.002

Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. (2014). Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology*, *126*, 295–312. https://doi.org/10.1016/j.jecp.2014.05.003

Jones, J. A., & Munhall, K. G. (1997). Effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*(4), 13-19

Jones, S. A., & Noppeney, U. (2021). Ageing and multisensory integration: A review of the evidence, and a computational perspective. *Cortex*, *138*, 1–23. https://doi.org/10.1016/j.cortex.2021.02.001

Jones, S. A., Beierholm, U., Meijer, D., & Noppeney, U. (2019). Older adults sacrifice response speed to preserve multisensory integration performance. *Neurobiology of Aging*, *84*, 148–157. https://doi.org/10.1016/j.neurobiolaging.2019.08.017

Kaganovich, N., Schumaker, J., Macias, D., & Gustafson, D. (2015). Processing of audiovisually congruent and incongruent speech in school-age children with a history of specific language impairment: a behavioral and event-related potentials study. *Developmental Science*, *18*(5), 751–770. https://doi.org/10.1111/desc.12263

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and Lexical Effects on Audiovisual Word Recognition by Adults With Cochlear Implants. *Journal of Speech, Language, and Hearing Research*, *46*(2), 390–404. https://doi.org/10.1044/1092-4388(2003/032)

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic Structure and Meaning Organize Neural Oscillations into a Content-Specific Hierarchy. *The Journal of Neuroscience*, *40*(49), 9467–9475. https://doi.org/10.1523/JNEUROSCI.0302-20.2020

Kawase, T., Yahata, I., Kanno, A., Sakamoto, S., Takanashi, Y., Takata, S., Nakasato, N., Kawashima, R., & Katori, Y. (2016). Impact of Audio-Visual Asynchrony on Lip-Reading Effects -Neuromagnetic and Psychophysical Study. *PloS One*, *11*(12), e0168740–e0168740. https://doi.org/10.1371/journal.pone.0168740

Keetels, M., & Vroomen, J. (2007). No effect of auditory-visual spatial disparity on temporal recalibration. *Experimental Brain Research*, *182*(4), 559–565. https://doi.org/10.1007/s00221-007-1012-2

Kim, J., & Davis, C. (2003). Hearing Foreign Voices: Does Knowing What is Said Affect Visual-Masked-Speech Detection? *Perception (London)*, *32*(1), 111–120. https://doi.org/10.1068/p3466

King, C., Warrier, C. M., Hayes, E., & Kraus, N. (2002). Deficits in auditory brainstem pathway encoding of speech sounds in children with learning problems. *Neuroscience Letters*, *319*(2), 111–115. https://doi.org/10.1016/S0304-3940(01)02556-3

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical Effects on Spoken Word Recognition by Pediatric Cochlear Implant Users. *Ear and Hearing*, *16*(5), 470–481. https://doi.org/10.1097/00003446-199510000-00004

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Kishon-Rabin, L., & Henkin, Y. (2000). Age-Related Changes in the Visual Perception of Phonologically Significant Contrasts. *British Journal of Audiology*, *34*(6), 363–374. https://doi.org/10.3109/03005364000000152

Knecht, H. A., Nelson, P. B., Whitelaw, G. M., & Feth, L. L. (2002). Background Noise Levels and Reverberation Times in Unoccupied Classrooms: Predictions and Measurements. *American Journal of Audiology*, *11*(2), 65–71. https://doi.org/10.1044/1059-0889(2002/009)

Knowland, V. C. P., Evans, S., Snell, C., & Rosen, S. (2016). Visual Speech Perception in Children with Language Learning Impairments. *Journal of Speech, Language, and Hearing Research*, *59*(1), 1–14. https://doi.org/10.1044/2015_JSLHR-S-14-0269

Kocins, K. E., Van Engen, K. J., Brown, V. A., McClannahan, K., & Peelle, J. (2023). Impact of clear face masks on audiovisual speech intelligibility and subjective listening effort with normal hearing young adults. *The Journal of the Acoustical Society of America*, *153*(3_supplement), A168–A168. https://doi.org/10.1121/10.0018541

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PloS One*, *2*(9), e943–e943. https://doi.org/10.1371/journal.pone.0000943

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *The American Psychologist*, *59*(2), 105–117. https://doi.org/10.1037/0003-066X.59.2.105

Król, M. E. (2018). Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study. *PloS One*, *13*(3), e0194491-. https://doi.org/10.1371/journal.pone.0194491

Kroos, C. (2007). Auditory-visual speech analysis: In search of a theory. In *Proceedings of the 16th International Congress of Phonetics Sciences, Saarbrcken, Germany* (pp. 279-284).

Kubicek, C., Gervain, J., Hillairet de Boisferon, A., Pascalis, O., Lœvenbruck, H., & Schwarzer, G. (2014). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behavior & Development*, *37*(4), 644–651. https://doi.org/10.1016/j.infbeh.2014.08.010

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*(6), 812–822. https://doi.org/10.1016/0959-4388(94)90128-7

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., & Meltzoff, A. N. (1982). The Bimodal Perception of Speech in Infancy. *Science (American Association for the Advancement of Science)*, *218*(4577), 1138–1141. https://doi.org/10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The Intermodal Representation of Speech in Infants. *Infant Behavior & Development*, *7*(3), 361–381. https://doi.org/10.1016/S0163-6383(84)80050-8

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological Evidence of Illusory Audiovisual Speech Percept in Human Infants. *Proceedings of the National Academy of Sciences - PNAS*, *105*(32), 11442–11445. https://doi.org/10.1073/pnas.0804275105

Kushnerenko, E., Tomalski, P., Ballieux, H., Potton, A., Birtles, D., Frostick, C., & Moore, D. G. (2013). Brain responses and looking behavior during audiovisual speech integration in infants predict auditory speech comprehension in the second year of life. *Frontiers in Psychology*, *4*, 432–432. https://doi.org/10.3389/fpsyg.2013.00432

Kyle, F. E., Campbell, R., Mohammed, T., Coleman, M., & MacSweeney, M. (2013). Speechreading Development in Deaf and Hearing Children: Introducing the Test of Child Speechreading. *Journal of Speech, Language, and Hearing Research*, *56*(2), 416–426. https://doi.org/10.1044/1092-4388(2012/12-0039)

Lachs, L., & Pisoni, D. B. (2004). Cross-Modal Source Information and Spoken Word Recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(2), 378–396. https://doi.org/10.1037/0096-1523.30.2.378

Lalonde, K. (2019, January). Effects of Natural Variability in Cross-Modal Temporal Correlations on Audiovisual Speech Recognition Benefit. In *INTERSPEECH* (pp. 2260-2264).

Lalonde, K., & Holt, R. F. (2015). Preschoolers Benefit from Visually Salient Speech Cues. *Journal of Speech, Language, and Hearing Research*, *58*(1), 135–150. https://doi.org/10.1044/2014_JSLHR-H-13-0343

Lalonde, K., & Holt, R. F. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *The Journal of the Acoustical Society of America*, *139*(4), 1713–1723. https://doi.org/10.1121/1.4945590

Lalonde, K., & McCreery, R. W. (2020). Audiovisual Enhancement of Speech Perception in Noise by School-Age Children Who Are Hard of Hearing. *Ear and Hearing*, *41*(4), 705–719. https://doi.org/10.1097/AUD.0000000000000830

Lalonde, K., & Werner, L. A. (2019). Infants and Adults Use Visual Cues to Improve Detection and Discrimination of Speech in Noise. *Journal of Speech, Language, and Hearing Research*, *62*(10), 3860–3875. https://doi.org/10.1044/2019_JSLHR-H-19-0106

Lalonde, K., & Werner, L. A. (2021). Development of the Mechanisms Underlying Audiovisual Speech Perception Benefit. *Brain Sciences*, *11*(1), 49-. https://doi.org/10.3390/brainsci11010049

Lambert, V., & Glacken, M. (2011). Engaging with children in research: Theoretical and practical implications of negotiating informed consent/assent. *Nursing Ethics*, *18*(6), 781–801. https://doi.org/10.1177/0969733011401122

Lee, A. K., & Wallace, M. T. (2019). Visual influence on auditory perception. *Multisensory Processes: The Auditory Perspective*, 1-8.

Lee, A. K., Maddox, R. K., & Bizley, J. K. (2019). An object-based interpretation of audiovisual processing. Multisensory processes: the auditory perspective, 59-83.

Lee, A. K., Maddox, R., & Bizley, J. (2023). Towards audiovisual scene analysis and object formation. *The Journal of the Acoustical Society of America*, *154*(4_supplement), A265–A265. https://doi.org/10.1121/10.0023478

Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in Psychology*, *5*, 868–868. https://doi.org/10.3389/fpsyg.2014.00868

Legault, I., Gagné, J.-P., Rhoualem, W., & Anderson-Gosselin, P. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. *International Journal of Audiology*, *49*(12), 904–911. https://doi.org/10.3109/14992027.2010.509112

Leibold, L. J., & Buss, E. (2019). Masked Speech Recognition in School-Age Children. *Frontiers in Psychology*, *10*, 1981–1981. https://doi.org/10.3389/fpsyg.2019.01981

Levene, H. (1960). Robust tests for equality of variances. *Contributions to probability and statistics*, 278-292.

Lewis, D., Schmid, K., O'Leary, S., Spalding, J., Heinrichs-Graham, E., & High, R. (2016). Effects of Noise on Speech Recognition and Listening Effort in Children with Normal Hearing and Children with Mild Bilateral or Unilateral Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *59*(5), 1218–1232. https://doi.org/10.1044/2016_JSLHR-H-15-0207

Lewkowicz, D. J. (1996). Perception of Auditory-Visual Temporal Synchrony in Human Infants. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(5), 1094–1106. https://doi.org/10.1037/0096-1523.22.5.1094

Lewkowicz, D. J. (2000). The Development of Intersensory Temporal Perception: An Epigenetic Systems/Limitations View. *Psychological Bulletin*, *126*(2), 281–308. https://doi.org/10.1037/0033-2909.126.2.281

Lewkowicz, D. J. (2010). Infant Perception of Audio-Visual Speech Synchrony. *Developmental Psychology*, *46*(1), 66–77. https://doi.org/10.1037/a0015579

Lewkowicz, D. J., & Flom, R. (2014). The Audiovisual Temporal Binding Window Narrows in Early Childhood. *Child Development*, *85*(2), 685–694. https://doi.org/10.1111/cdev.12142

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences - PNAS*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Lewkowicz, D. J., & Kraebel, K. S. (2004). The Value of Multisensory Redundancy in the Development of Intersensory Perception.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development*, *37*(2), 90–94. https://doi.org/10.1177/0165025412467582

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162. https://doi.org/10.1016/j.jecp.2014.10.006

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Brain Research. Cognitive Brain Research*, *17*(2), 447–453. https://doi.org/10.1016/S0926-6410(03)00160-5

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11-14 years). *British Journal of Educational Psychology*, *70*(4), 505–518. https://doi.org/10.1348/000709900158272

Ma, K. S. T., & Schnupp, J. W. H. (2023). The unity hypothesis revisited: can the male/female incongruent McGurk effect be disrupted by familiarization and priming? *Frontiers in Psychology*, *14*, 1106562–1106562. https://doi.org/10.3389/fpsyg.2023.1106562

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS One*, *4*(3), e4638–e4638. https://doi.org/10.1371/journal.pone.0004638

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage (Orlando, Fla.)*, *21*(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

MacEachern, M. R. (2000). On the visual distinctiveness of words in the English lexicon. *Journal of Phonetics*, *28*(3), 367–376. https://doi.org/10.1006/jpho.2000.0119

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131–141. https://doi.org/10.3109/03005368709077786

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295. https://doi.org/10.1017/S0305000900006449

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *eLife*, *4*. https://doi.org/10.7554/eLife.04995

Magnotti, J. F., & Beauchamp, M. S. (2015). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, *22*(3), 701–709. https://doi.org/10.3758/s13423-014-0722-2

Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Computational Biology*, *13*(2), e1005229–e1005229. https://doi.org/10.1371/journal.pcbi.1005229

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*, *4*, 798–798. https://doi.org/10.3389/fpsyg.2013.00798

Maidment, D. W., Kang, H. J., Stewart, H. J., & Amitay, S. (2015). Audiovisual Integration in Children Listening to Spectrally Degraded Speech. *Journal of Speech, Language, and Hearing Research*, *58*(1), 61–68. https://doi.org/10.1044/2014_JSLHR-S-14-0044

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual Asynchrony Detection in Human Speech. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(1), 245–256. https://doi.org/10.1037/a0019952

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing Spoken Words: The Importance of Word Onsets. *Journal of Experimental Psychology. Human Perception and Performance*, *15*(3), 576–585. https://doi.org/10.1037/0096-1523.15.3.576

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of Asynchronous and Conflicting Visual and Auditory Speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777–1786. https://doi.org/10.1121/1.417342

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology*, *41*(1), 93–113. https://doi.org/10.1016/0022-0965(86)90053-6

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review: Speech Recognition in Adverse Condiditions. *Language and Cognitive Processes*, *27*(7–8), 953–978.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology. General*, *134*(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

McCreery, R. W., & Stelmachowicz, P. G. (2011). Audibility-based predictions of speech recognition for children and adults with normal hearing. *The Journal of the Acoustical Society of America*, *130*(6), 4070–4081. https://doi.org/10.1121/1.3658476

McCreery, R. W., Miller, M. K., Buss, E., & Leibold, L. J. (2020). Cognitive and Linguistic Contributions to Masked Speech Recognition in Children. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3525–3538. https://doi.org/10.1044/2020_JSLHR-20-00030

McCreery, R. W., Spratford, M., Kirby, B., & Brennan, M. (2017). Individual differences in language and working memory affect children's speech recognition in noise. *International journal of audiology*, *56*(5), 306-315.

McCreery, R., Ito, R., Spratford, M., Lewis, D., Hoover, B., & Stelmachowicz, P. G. (2010). Performance-Intensity Functions for Normal-Hearing Adults and Children Using Computer-Aided Speech Perception Assessment. *Ear and Hearing*, *31*(1), 95–101. https://doi.org/10.1097/AUD.0b013e3181bc7702

McFadden, D. (1998). Sex differences in the auditory system: Gonadal hormones and sex differences in behavior. *Developmental Neuropsychology*, *14*(2–3), 261–298.

McGovern, D. P., Burns, S., Hirst, R. J., & Newell, F. N. (2022). Perceptual training narrows the temporal binding window of audiovisual integration in both younger and older adults. *Neuropsychologia*, *173*, 108309–108309. https://doi.org/10.1016/j.neuropsychologia.2022.108309

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature (London)*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech Categorization Develops Slowly Through Adolescence. *Developmental Psychology*, *54*(8), 1472–1491. https://doi.org/10.1037/dev0000542

Mealings, K. T., Demuth, K., Buchholz, J. M., & Dillon, H. (2015). The effect of different open plan and enclosed classroom acoustic conditions on speech perception in Kindergarten children. *The Journal of the Acoustical Society of America*, *138*(4), 2458–2469. https://doi.org/10.1121/1.4931903

Megnin-Viggars, O., & Goswami, U. (2013). Audiovisual perception of noise vocoded speech in dyslexic and non-dyslexic adults: The role of low-frequency visual modulations. *Brain and Language*, *124*(2), 165–173. https://doi.org/10.1016/j.bandl.2012.12.002

Megnin, O., Flitton, A., R.G. Jones, C., de Haan, M., Baldeweg, T., & Charman, T. (2012). Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing. *Autism Research*, *5*(1), 39–48. https://doi.org/10.1002/aur.231

Meijer, D., Veselič, S., Calafiore, C., & Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex*, *119*, 74–88. https://doi.org/10.1016/j.cortex.2019.03.026

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, *365*(2), 350–354. https://doi.org/10.1016/0006-8993(86)91648-3

Meredith, M., Nemitz, J., & Stein, B. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience*, *7*(10), 3215–3229. https://doi.org/10.1523/jneurosci.07-10-03215.1987

Meronen, A., Tiippana, K., Westerholm, J., & Ahonen, T. (2013). Audiovisual Speech Perception in Children with Developmental Language Disorder in Degraded Listening Conditions. *Journal of Speech, Language, and Hearing Research*, *56*(1), 211–221. https://doi.org/10.1044/1092-4388(2012/11-0270)

Meunier, S., Van Eeckhoutte, M., & Moore, B. C. J. (2021). Editorial: Loudness: From Neuroscience to Perception. *Frontiers in Psychology*, *12*, 785093–785093. https://doi.org/10.3389/fpsyg.2021.785093

Michel, A. (2020). Cognition and Perception: Is There Really a Distinction?. *APS Observer*, *33*

Michon, M., Boncompte, G., & López, V. (2020). Electrophysiological Dynamics of Visual Speech Processing and the Role of Orofacial Effectors for Cross-Modal Predictions. *Frontiers in Human Neuroscience*, *14*, 538619–538619. https://doi.org/10.3389/fnhum.2020.538619

Middelweerd, M. J., & Plomp, R. (1987). The Effect of Speechreading on the Speech-Reception Threshold of Sentences in Noise. *The Journal of the Acoustical Society of America*, *82*(6), 2145–2147. https://doi.org/10.1121/1.395659

Mihalik, A., & Noppeney, U. (2020). Causal Inference in Audiovisual Perception. *The Journal of Neuroscience*, *40*(34), 6600–6612. https://doi.org/10.1523/JNEUROSCI.0051-20.2020

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of experimental psychology*, *41*(5), 329

Mitchel, A. D., & Weiss, D. J. (2011). Learning across Senses: Cross-Modal Effects in Multisensory Statistical Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(5), 1081–1091. https://doi.org/10.1037/a0023700

Mittag, M., Thesleff, P., Laasonen, M., & Kujala, T. (2013). The neurophysiological basis of the integration of written and heard syllables in dyslexic adults. *Clinical Neurophysiology*, *124*(2), 315–326. https://doi.org/10.1016/j.clinph.2012.08.003

Mohammed, T., Campbell, R., Macsweeney, M., Barry, F., & Coleman, M. (2006). Speechreading and its association with reading among deaf, hearing and dyslexic individuals. *Clinical Linguistics & Phonetics*, *20*(7–8), 621–630. https://doi.org/10.1080/02699200500266745

Monson, B. B., Rock, J., Schulz, A., Hoffman, E., & Buss, E. (2019). Ecological cocktail party listening reveals the utility of extended high-frequency hearing. *Hearing Research*, *381*, 107773–107773. https://doi.org/10.1016/j.heares.2019.107773

Moore, B. C. (2014). *Auditory processing of temporal fine structure: Effects of age and hearing loss*. World Scientific.

Moore, J. K. (2002). Maturation of Human Auditory Cortex: Implications for Speech Perception. *Annals of Otology, Rhinology & Laryngology*, *111*(5_suppl), 7–10. https://doi.org/10.1177/00034894021110S502

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory Integration in Macaque Visual Cortex Depends on Cue Reliability. *Neuron (Cambridge, Mass.)*, *59*(4), 662–673. https://doi.org/10.1016/j.neuron.2008.06.024

Morrell, C. H., Gordon-Salant, S., Pearson, J. D., Brant, L. J., & Fozard, J. L. (1996). Age- and gender-specific reference ranges for hearing level and longitudinal changes in hearing level. *The Journal of the Acoustical Society of America*, *100*(4 Pt 1), 1949–1967. https://doi.org/10.1121/1.417906

Möttönen, R., Schürmann, M., & Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neuroscience Letters*, *363*(2), 112–115. https://doi.org/10.1016/j.neulet.2004.03.076

Möttönen, R., Tiippana, K., Sams, M., & Puharinen, H. (2011). Sound Location Can Influence Audiovisual Speech Perception When Spatial Attention Is Manipulated. *Seeing and Perceiving*, *24*(1), 67–90. https://doi.org/10.1163/187847511X557308

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal Constraints on the McGurk Effect. *Perception & Psychophysics*, *58*(3), 351–362. https://doi.org/10.3758/BF03206811

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, *15*(2), 133-137.

Musacchia, G., Sams, M., Nicol, T., & Kraus, N. (2006). Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, *168*(1–2), 1–10. https://doi.org/10.1007/s00221-005-0071-5

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America*, *132*(2), 1061–1077. https://doi.org/10.1121/1.4728187

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2015). Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America*, *137*(1), 362–377. https://doi.org/10.1121/1.4904536

Nath, A. R., Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *The Journal of Neuroscience*, *31*(39), 13963–13971. https://doi.org/10.1523/JNEUROSCI.2605-11.2011

Nava, E., & Pavani, F. (2013). Changes in Sensory Dominance During Childhood: Converging Evidence From the Colavita Effect and the Sound-Induced Flash Illusion. *Child Development*, *84*(2), 604–616. https://doi.org/10.1111/j.1467-8624.2012.01856.x

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, *11*(1), 4–11. https://doi.org/10.1016/j.inffus.2009.04.001

Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Adaptation to audiotactile asynchrony. *Neuroscience Letters*, *413*(1), 72–76. https://doi.org/10.1016/j.neulet.2006.11.027

Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science*, *9*(5), 454–464. https://doi.org/10.1111/j.1467-7687.2006.00512.x

Nelson, P. B., & Soli, S. (2000). Acoustical Barriers to Learning: Children at Risk in Every Classroom. *Language, Speech & Hearing Services in Schools*, *31*(4), 356–361. https://doi.org/10.1044/0161-1461.3104.356

Neuman, A. C., Wroblewski, M., Hajicek, J., & Rubinstein, A. (2010). Combined Effects of Noise and Reverberation on Speech Recognition Performance of Normal-Hearing Children and Adults. *Ear and Hearing*, *31*(3), 336–344. https://doi.org/10.1097/AUD.0b013e3181d3d514

Newmark, M., Merlob, P., Bresloff, I., Olsha, M., & Attias, J. (1997). Click Evoked Otoacoustic Emissions: Inter-aural and Gender Differences in Newborns. *Journal of Basic and Clinical Physiology and Pharmacology*, *8*(3), 133–140. https://doi.org/10.1515/JBCPP.1997.8.3.133

Nittrouer, S., & Boothroyd, A. (1990). Context Effects in Phoneme and Word Recognition by Young Children and Older Adults. *The Journal of the Acoustical Society of America*, *87*(6), 2705–2715. https://doi.org/10.1121/1.399061

Norrix, L. W., Plante, E., Vance, R., & Boliek, C. A. (2007). Auditory-Visual Integration for Speech by Children With and Without Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1639–1651. https://doi.org/10.1044/1092-4388(2007/111)

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving Developmental Research Online: Comparing In-Lab and Web-Based Studies of Model-Based Reinforcement Learning. *Collabra. Psychology*, *6*(1). https://doi.org/10.1525/collabra.17213

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in Visual, Auditory, and Audiovisual Perception of Space: e1004649. *PLoS Computational Biology*, *11*(12). https://doi.org/10.1371/journal.pcbi.1004649

Odegaard, B., Wozny, D. R., & Shams, L. (2017). A simple and efficient method to enhance audiovisual binding tendencies. *PeerJ (San Francisco, CA)*, *5*, e3143–e3143. https://doi.org/10.7717/peerj.3143

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2003). Cross-modal enhancement of perceived brightness: Sensory interaction versus response bias. *Perception & Psychophysics*, *65*(1), 123–132. https://doi.org/10.3758/BF03194789

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective & Behavioral Neuroscience (Print)*, *4*(2), 127–132. https://doi.org/10.3758/CABN.4.2.127

Okada, K., Venezia, J. H., Matchin, W., Saberi, K., & Hickok, G. (2013). An fMRI Study of Audiovisual Speech Perception Reveals Multisensory Interactions in Auditory Cortex. *PloS One*, *8*(6), e68959–e68959. https://doi.org/10.1371/journal.pone.0068959

Owens, E., & Blazek, B. (1985). Visemes Observed by Hearing-Impaired and Normal-Hearing Adult Viewers. *Journal of Speech and Hearing Research*, *28*(3), 381–393. https://doi.org/10.1044/jshr.2803.381

Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition*, *125*(3), 353–364. https://doi.org/10.1016/j.cognition.2012.08.003

Pandey, P. C., Kunov, H., & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research*, *26*(1), 27-.

Pasupathy, A. (2015). The neural basis of image segmentation in the primate brain. *Neuroscience*, *296*, 101–109. https://doi.org/10.1016/j.neuroscience.2014.09.051

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior & Development*, *22*(2), 237–247. https://doi.org/10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2002). Infants' Ability to Match Dynamic Phonetic and Gender Information in the Face and Voice. *Journal of Experimental Child Psychology*, *81*(1), 93–115. https://doi.org/10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191–196. https://doi.org/10.1111/1467-7687.00271

Peelle, J. E., & Davis, M. H. (2012). Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology*, *3*, 320–320. https://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*(Jul), 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pekkola, J., Laasonen, M., Ojanen, V., Autti, T., Jääskeläinen, I. P., Kujala, T., & Sams, M. (2006). Perception of matching and conflicting audiovisual speech in dyslexic and fluent readers: An fMRI study at 3 T. *NeuroImage (Orlando, Fla.)*, *29*(3), 797–807. https://doi.org/10.1016/j.neuroimage.2005.09.069

Pelham, W. E., & Ross, A. O. (1977). Selective Attention in Children with Reading Problems: A Developmental Study of Incidental Learning. *Journal of Abnormal Child Psychology*, *5*(1), 1-.

Peng, Z. E., Waz, S., Buss, E., Shen, Y., Richards, V., Bharadwaj, H., Stecker, G. C., Beim, J. A., Bosen, A. K., Braza, M. D., Diedesch, A. C., Dorey, C. M., Dykstra, A. R., Gallun, F. J., Goldsworthy, R. L., Gray, L., Hoover, E. C., Ihlefeld, A., Koelewijn, T., … Venezia, J. H. (2022). FORUM: Remote testing for psychological and physiological acoustics. *The Journal of the Acoustical Society of America*, *151*(5), 3116–3128. https://doi.org/10.1121/10.0010422

Petrini, K., Jones, P. R., Smith, L., & Nardini, M. (2015). Hearing Where the Eyes See: Children Use an Irrelevant Visual Cue When Localizing Sounds. *Child Development*, *86*(5), 1449–1457. https://doi.org/10.1111/cdev.12397

Pichora-Fuller, M.K. (2006). Perceptual Effort and Apparent Cognitive Decline: Implications for Audiologic Rehabilitation. *Seminars in Hearing*, *27*(4), 284–293. https://doi.org/10.1055/s-2006-954855

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, *6*(4), 203–205. https://doi.org/10.3758/BF03207017

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences - PNAS*, *106*(26), 10598–10602. https://doi.org/10.1073/pnas.0904134106

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual Training Narrows the Temporal Window of Multisensory Binding. *The Journal of Neuroscience*, *29*(39), 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

Proverbio, A. M., Massetti, G., Rizzi, E., & Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific Reports*, *6*(1), 30423–30423. https://doi.org/10.1038/srep30423

Ramirez, J., & Mann, V. (2005). Using auditory-visual speech to probe the basis of noise-impaired consonant-vowel perception in dyslexia and auditory neuropathy. *The Journal of the Acoustical Society of America*, *118*(2), 1122–1133. https://doi.org/10.1121/1.1940509

Ramus, F., & Szenkovits, G. (2008). What phonological deficit? *Quarterly Journal of Experimental Psychology (2006)*, *61*(1), 129–141. https://doi.org/10.1080/17470210701508822

Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, *17*(12), 1784–1792. https://doi.org/10.1038/nn.3865

Recio, R. S., Cravo, A. M., de Camargo, R. Y., & van Wassenhove, V. (2019). Dissociating the sequential dependency of subjective temporal order from subjective simultaneity. *PloS One*, *14*(10), e0223184–e0223184. https://doi.org/10.1371/journal.pone.0223184

Reips, U. D. (2001). The Web Experimental Psychology Lab: five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers*, *33*(2), 201–211. https://doi.org/10.3758/BF03195366

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology*, *49*(4), 243–256. https://doi.org/10.1027//1618-3169.49.4.243

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Lawrence Erlbaum Associates, Inc.

Remez, R. E., & Rubin, P. E. (1982). Perception of voice pitch in sinusoidal imitations of speech. *The Journal of the Acoustical Society of America*, *71*(S1), S96–S96. https://doi.org/10.1121/1.2019657

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing Developmental Science via Unmoderated Remote Research with Children. *Journal of Cognition and Development*, *21*(4), 477–493. https://doi.org/10.1080/15248372.2020.1797751

Rohe, T., & Noppeney, U. (2015a). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision (Charlottesville, Va.)*, *15*(5), 22–22. https://doi.org/10.1167/15.5.22

Rohe, T., & Noppeney, U. (2015b). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception: e1002073. *PLoS Biology*, *13*(2). https://doi.org/10.1371/journal.pbio.1002073

Rohe, T., & Noppeney, U. (2016). Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. *Current Biology*, *26*(4), 509–514. https://doi.org/10.1016/j.cub.2015.12.056

Rohlf, S., Li, L., Bruns, P., & Röder, B. (2020). Multisensory integration develops prior to crossmodal recalibration. *Current Biology*, *30*(9), 1726-1732

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78

Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In *The Handbook of Speech Perception* (pp. 51–78). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757024.ch3

Rosenblum, L. D. (2008). Speech Perception as a Multimodal Phenomenon. *Current Directions in Psychological Science : A Journal of the American Psychological Society*, *17*(6), 405–409. https://doi.org/10.1111/j.1467-8721.2008.00615.x

Rosenblum, L. D., & Saldaña, H. M. (1996). An Audiovisual Test of Kinematic Primitives for Visual Speech Perception. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(2), 318–331. https://doi.org/10.1037/0096-1523.22.2.318

Ross L.A., Molholm S, Blanco D, Gomez-Ramirez M., Saint-Amour D, & FoxeJ.J. (2011). The development of multisensory speech perception continues into the late childhood years: Development of audiovisual speech perception. *The European Journal of Neuroscience*, *33*, 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Ross, L. A., Del Bene, V. A., Molholm, S., Frey, H.-P., & Foxe, J. J. (2015). Sex differences in multisensory speech processing in both typically developing children and those on the autism spectrum. *Frontiers in Neuroscience*, *9*, 185–185. https://doi.org/10.3389/fnins.2015.00185

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex (New York, N.Y. 1991)*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Rüssler, J., Gerth, I., Heldmann, M., & Münte, T. F. (2015). Audiovisual perception of natural speech is impaired in adult dyslexics: An ERP study. *Neuroscience*, *287*(257), 55–65. https://doi.org/10.1016/j.neuroscience.2014.12.023

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science (American Association for the Advancement of Science)*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Salinas-Marchant, C., & MacLeod, A. A. N. (2022). Audiovisual speech perception in children: a scoping review. *Speech, Language and Hearing*, *25*(4), 433–449. https://doi.org/10.1080/2050571X.2021.1923302

Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*(1–2), 75–87. https://doi.org/10.1016/S0167-6393(98)00051-X

Sarantopoulos, G., Lykoudis, S., & Kassomenos, P. (2014). Noise levels in primary schools of medium sized city in Greece. *The Science of the Total Environment*, *482–483*, 493–500. https://doi.org/10.1016/j.scitotenv.2013.09.010

Scarborough, H. S. (1998). Predicting the Future Achievement of Second Graders with Reading Disabilities: Contributions of Phonemic Awareness, Verbal Memory, Rapid Naming, and IQ. *Annals of Dyslexia*, *48*(1), 115–136. https://doi.org/10.1007/s11881-998-0006-5

Schepers, I. M., Yoshor, D., & Beauchamp, M. S. (2015). Electrocorticography Reveals Enhanced Visual Cortex Responses to Visual Speech. *Cerebral Cortex (New York, N.Y. 1991)*, *25*(11), 4103–4110. https://doi.org/10.1093/cercor/bhu127

Schrank, F. A., Wendling, B. J., Flanagan, D. P., & McDonough, E. M. (2018). The Woodcock–Johnson IV Tests of Early Cognitive and Academic Development. *Contemporary intellectual assessment: Theories, tests, and issues*, 283-301.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, *12*(3), 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Schwartz, J.-L. (2010). A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America*, *127*(3), 1584–1594. https://doi.org/10.1121/1.3293001

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind (Cambridge, Mass.)*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Seewald, R. C., Ross, M., Giolas, T. G., & Yonovitz, A. (1985). Primary Modality for Speech Perception in Children with Normal and Impaired Hearing. *Journal of Speech and Hearing Research*, *28*(1), 36–46. https://doi.org/10.1044/jshr.2801.36

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, *21*(4), 427–444. https://doi.org/10.1016/S0095-4470(19)30229-3

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology*, *5*, 323–323. https://doi.org/10.3389/fpsyg.2014.00323

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia*, *45*(3), 561–571. https://doi.org/10.1016/j.neuropsychologia.2006.01.013

Setti, A., Burke, K. E., Kenny, R., & Newell, F. N. (2013). Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. *Frontiers in Psychology*, *4*, 575–575. https://doi.org/10.3389/fpsyg.2013.00575

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences (Regular Ed.)*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, *16*(17), 1923–1927. https://doi.org/10.1097/01.wnr.0000187634.68504.bb

Shapiro, N. T., Hippe, D. S., & Ramírez, N. F. (2021). How Chatty Are Daddies? An Exploratory Study of Infants' Language Environments. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3242–3252. https://doi.org/10.1044/2021_JSLHR-20-00727

Shatzer, H., Shen, S., Kerlin, J. R., Pitt, M. A., & Shahin, A. J. (2018). Neurophysiology underlying influence of stimulus reliability on audiovisual integration. *The European Journal of Neuroscience*, *48*(8), 2836–2848. https://doi.org/10.1111/ejn.13843

Shaw, K. E., & Bortfeld, H. (2015). Sources of Confusion in Infant Audiovisual Speech Perception Research. *Frontiers in Psychology*, *6*, 1844–1844. https://doi.org/10.3389/fpsyg.2015.01844

Shaywitz, S. E. (1998). Dyslexia. *The New England Journal of Medicine*, *338*(5), 307–312. https://doi.org/10.1056/NEJM199801293380507

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to Recognize Talkers From Natural, Sinewave, and Reversed Speech Samples. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(6), 1447–1469. https://doi.org/10.1037/0096-1523.28.6.1447

Sheskin, M., and Keil, F. (2018, December 30). A video chat platform for developmental research. *TheChildLab.com*.https://doi.org/10.31234/osf.io/rn7w5

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Shore, D. I., Spence, C., & Klein, R. M. (2001). Visual Prior Entry. *Psychological Science*, *12*(3), 205–212. https://doi.org/10.1111/1467-9280.00337

Sills, S. J., and Song, C. (2002). Innovations in survey research. Social Science Computer Review, 20(1), 22-30.

Sloutsky, V. M., & Napolitano, A. C. (2003). Is a Picture Worth a Thousand Words? Preference for Auditory Modality in Young Children. *Child Development*, *74*(3), 822–833. https://doi.org/10.1111/1467-8624.00570

Smayda, K. E., Van Engen, K. J., Maddox, W. T., & Chandrasekaran, B. (2016). Audio-Visual and Meaningful Semantic Context Enhancements in Older and Younger Adults. *PloS One*, *11*(3), e0152773–e0152773. https://doi.org/10.1371/journal.pone.0152773

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, *48*(8), 813–821. https://doi.org/10.1111/j.1469-7610.2007.01766.x

Snowling, M. (1998). Dyslexia as a Phonological Deficit: Evidence and Implications. *Child Psychology & Psychiatry Review*, *3*(1), 4–11. https://doi.org/10.1017/S1360641797001366

Snowling, M. J. (2000) Dyslexia, 2nd Edition, Wiley-Blackwell. Available at: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0631205748.html.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, *32*(25), 8443–8453. https://doi.org/10.1523/JNEUROSCI.5069-11.2012

Sommers, M. S., & Phelps, D. (2016). Listening Effort in Younger and Older Adults: A Comparison of Auditory-Only and Auditory-Visual Presentations. *Ear and Hearing*, *37 Suppl 1*(1), 62S-68S. https://doi.org/10.1097/AUD.0000000000000322

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-Visual Speech Perception and Auditory-Visual Enhancement in Normal-Hearing Younger and Older Adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Spelke, E. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, *8*(4), 553–560. https://doi.org/10.1016/0010-0285(76)90018-9

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626–636. https://doi.org/10.1037/0012-1649.15.6.626

Spelke, E. S., Born, W. S., & Chu, F. (1983). Perception of Moving, Sounding Objects by Four-Month-Old Infants. *Perception (London)*, *12*(6), 719–732. https://doi.org/10.1068/p120719

Stasiak, K., Fleming, T., Lucassen, M. F. G., Shepherd, M. J., Whittaker, R., & Merry, S. N. (2016). Computer-Based and Online Therapy for Depression and Anxiety in Children and Adolescents. *Journal of Child and Adolescent Psychopharmacology*, *26*(3), 235–245. https://doi.org/10.1089/cap.2015.0029

Stecker, G. C., & Hafter, E. R. (2000). An effect of temporal asymmetry on loudness. *The Journal of the Acoustical Society of America*, *107*(6), 3358–3368. https://doi.org/10.1121/1.429407

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis. *Journal of Cognitive Neuroscience*, *8*(6), 497–506. https://doi.org/10.1162/jocn.1996.8.6.497

Stein, B., & Meredith, M. A. (1993). The merging of the senses (Cognitive neuroscience series). Cambridge, Mass.: MIT Press.

Stelmachowicz, P. G., Hoover, B. M., Lewis, D. E., Kortekaas, R. W. L., & Pittman, A. L. (2000). The Relation Between Stimulus Context, Speech Audibility, and Perception for Normal-Hearing and Hearing-Impaired Children. *Journal of Speech, Language, and Hearing Research*, *43*(4), 902–914. https://doi.org/10.1044/jslhr.4304.902

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage (Orlando, Fla.)*, *44*(3), 1210–1223. https://doi.org/10.1016/j.neuroimage.2008.09.034

Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Experimental Brain Research*, *227*(2), 249–261. https://doi.org/10.1007/s00221-013-3507-3

Stevenson, R. A., Baum, S. H., Krueger, J., Newhouse, P. A., & Wallace, M. T. (2018). Links Between Temporal Acuity and Multisensory Integration Across Life Span. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(1), 106–116. https://doi.org/10.1037/xhp0000424

Stevenson, R. A., Nelms, C. E., Baum, S. H., Zurkovsky, L., Barense, M. D., Newhouse, P. A., & Wallace, M. T. (2015). Deficits in audiovisual speech perception in normal aging emerge at the level of whole-word recognition. *Neurobiology of Aging*, *36*(1), 283–291. https://doi.org/10.1016/j.neurobiolaging.2014.08.003

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., & Wallace, M. T. (2014). Brief Report: Arrested Development of Audiovisual Speech Perception in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *44*(6), 1470–1477. https://doi.org/10.1007/s10803-013-1992-7

Stevenson, R. A., Wallace, M. T., & Altieri, N. (2014). The interaction between stimulus factors and cognitive factors during multisensory integration of audiovisual speech. *Frontiers in Psychology*, *5*, 352–352. https://doi.org/10.3389/fpsyg.2014.00352

Stevenson, R. A., Wilson, M. M., Powers, A. R., & Wallace, M. T. (2013). The effects of visual training on multisensory temporal processing. *Experimental Brain Research*, *225*(4), 479–489. https://doi.org/10.1007/s00221-012-3387-y

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual Differences in the Multisensory Temporal Binding Window Predict Susceptibility to Audiovisual Illusions. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(6), 1517–1529. https://doi.org/10.1037/a0027339

STeVi Speech Test Video Corpus. (n.d.). Sensimetrics' Speech Videos. https://www.sens.com/products/stevi-speech-test-video-corpus/

Studdert-Kennedy, M. (1989). PART OF REVIEW SYMPOSIUM - Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry (see abstract of review in this issue) [Review of *PART OF REVIEW SYMPOSIUM - Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry (see abstract of review in this issue)*]. *Behavioral and Brain Sciences*, *12*(4), 774–775.

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

Summerfield, Q. (1992). Lipreading and Audio-Visual Speech Perception. *Philosophical Transactions of the Royal Society of London. Series B. Biological Sciences*, *335*(1273), 71–78. https://doi.org/10.1098/rstb.1992.0009

Tacca, M. C. (2011). Commonalities between Perception and Cognition. *Frontiers in Psychology*, *2*, 358–358. https://doi.org/10.3389/fpsyg.2011.00358

Taitelbaum-Swead, R., & Fostick, L. (2016). Auditory and visual information in speech perception: A developmental perspective. *Clinical Linguistics & Phonetics*, *30*(7), 531–545. https://doi.org/10.3109/02699206.2016.1151938

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850–855. https://doi.org/10.1016/j.cognition.2008.05.009

ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology*, *4*, 331–331. https://doi.org/10.3389/fpsyg.2013.00331

Thillay, A., Roux, S., Gissot, V., Carteau-Martin, I., Knight, R. T., Bonnet-Brilhault, F., & Bidet-Caulet, A. (2015). Sustained attention and prediction: distinct brain maturation trajectories during adolescence. *Frontiers in Human Neuroscience*, *9*, 519–519. https://doi.org/10.3389/fnhum.2015.00519

Thompson, L. A., & Malloy, D. (2004). Attention Resources and Visible Speech Encoding in Older and Younger Adults. *Experimental Aging Research*, *30*(3), 241–252. https://doi.org/10.1080/03610730490447877

Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611–624. https://doi.org/10.1080/17405629.2012.728076

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). Test of word reading efficiency–second edition (TOWRE-2). *Austin, TX: Pro-Ed.*

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition*, *96*(1), B13–B22. https://doi.org/10.1016/j.cognition.2004.10.004

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in School-Age Children: The Roles of Age, Hearing Status, and Cognitive Ability. *Journal of Speech, Language, and Hearing Research*, *57*(2), 556–565. https://doi.org/10.1044/2013_JSLHR-H-12-0273

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007a). Audiovisual Integration and Lipreading Abilities of Older Adults with Normal and Impaired Hearing. *Ear and Hearing*, *28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e31812f7185

Tye-Murray, N., Sommers, M., & Spehar, B. (2007b). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, *11*(4), 233–241. https://doi.org/10.1177/1084713807307409

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, Audiovisual Integration, and the Principle of Inverse Effectiveness. *Ear and Hearing*, *31*(5), 636–644. https://doi.org/10.1097/AUD.0b013e3181ddf7ff

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology*, *47*(S2), S31–S37. https://doi.org/10.1080/14992020802301662

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and Audiovisual Speech Recognition Across the Adult Lifespan: Implications for Audiovisual Integration. *Psychology and Aging*, *31*(4), 380–389. https://doi.org/10.1037/pag0000094

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M. S., & Hale, S. (2011). Cross-Modal Enhancement of Speech Detection in Young and Older Adults: Does Signal Content Matter? *Ear and Hearing*, *32*(5), 650–655. https://doi.org/10.1097/AUD.0b013e31821a4578

van Atteveldt, N. M., Peterson, B. S., & Schroeder, C. E. (2014). Contextual control of audiovisual integration in low-level sensory cortices. *Human Brain Mapping*, *35*(5), 2394–2411. https://doi.org/10.1002/hbm.22336

van Bergen, E., de Jong, P. F., Maassen, B., & van der Leij, A. (2014). The Effect of Parents' Literacy Skills and Children's Preliteracy Skills on the Risk of Dyslexia. *Journal of Abnormal Child Psychology*, *42*(7), 1187–1200. https://doi.org/10.1007/s10802-014-9858-9

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience*, *13*, 335–335. https://doi.org/10.3389/fnhum.2019.00335

Van der Burg, E., Cass, J., Alais, D., & Theeuwes, J. (2011). The Temporal Window of Multisensory Integration under Competing Circumstances. *I-Perception (London)*, *2*(8), 962–962. https://doi.org/10.1068/ic962

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

van Eijk, R. L. J., Kohlrausch, A., Juola, J. F., & van de Par, S. (2008). Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. *Perception & Psychophysics*, *70*(6), 955–968. https://doi.org/10.3758/PP.70.6.955

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America*, *152*(6), 3216–3225. https://doi.org/10.1121/10.0015262

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing Speech Intelligibility: Interactions among Context, Modality, Speech Style, and Masker. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1908–1918. https://doi.org/10.1044/JSLHR-H-13-0076

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics*, *79*(2), 396–403. https://doi.org/10.3758/s13414-016-1238-9

van Laarhoven, T., Keetels, M., Schakel, L., & Vroomen, J. (2018). Audio-visual speech in noise perception in dyslexia. *Developmental Science*, *21*(1). https://doi.org/10.1111/desc.12504

van Viersen, S., de Bree, E. H., Verdam, M., Krikhaar, E., Maassen, B., van der Leij, A., & de Jong, P. F. (2017). Delayed Early Vocabulary Development in Children at Family Risk of Dyslexia. *Journal of Speech, Language, and Hearing Research*, *60*(4), 937–949. https://doi.org/10.1044/2016_JSLHR-L-16-0031

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

van Wassenhove, V., Grant, K. W., Poeppel, D., & Halle, M. (2005). Visual Speech Speeds up the Neural Processing of Auditory Speech. *Proceedings of the National Academy of Sciences - PNAS*, *102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Vance, M., Stackhouse, J., & Wells, B. (2005). Speech-production skills in children aged 3-7 years. *International Journal of Language & Communication Disorders*, *40*(1), 29–48. https://doi.org/10.1080/13682820410001716172

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, *1111*(1), 134–142. https://doi.org/10.1016/j.brainres.2006.05.078

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Attention, Perception & Psychophysics*, *69*(5), 744-.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision (Charlottesville, Va.)*, *8*(9), 14.1-14. https://doi.org/10.1167/8.9.14

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. *Experimental Brain Research*, *185*(3), 521–529. https://doi.org/10.1007/s00221-007-1168-9

Venezia, J. H., Fillmore, P., Matchin, W., Lisette Isenberg, A., Hickok, G., & Fridriksson, J. (2016). Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech. *NeuroImage (Orlando, Fla.)*, *126*, 196–207. https://doi.org/10.1016/j.neuroimage.2015.11.038

Vestergaard, M. D., Fyson, N. R. C., & Patterson, R. D. (2009). The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *The Journal of the Acoustical Society of America*, *125*(2), 1114–1124. https://doi.org/10.1121/1.3050321

Vidyasagar, T. R., & Pammer, K. (2010). Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences*, *14*(2), 57–63. https://doi.org/10.1016/j.tics.2009.12.003

Voss, A. H. (2016). Within-subjects changes in lipreading and visual enhancement among older adults.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*(1), 78–86. https://doi.org/10.1016/j.cognition.2010.10.002

Wagner R.K.,Torgesen J.K., Rashotte C.A. & Pearson N.A.. (2013). Comprehensive Test of Phonological Processing – Second Edition (CTOPP-2). Austin, TX: PRO-ED.

Wagner, R. K., & Torgesen, J. K. (1987). The Nature of Phonological Processing and Its Causal Role in the Acquisition of Reading Skills. *Psychological Bulletin*, *101*(2), 192–212. https://doi.org/10.1037/0033-2909.101.2.192

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). CTOPP examiner's manual. *Austin (TX): PROED*

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of Training on the Visual Recognition of Consonants. *Journal of Speech and Hearing Research*, *20*(1), 130–145. https://doi.org/10.1044/jshr.2001.130

Walden, B. E., Surr, R. K., Cord, M. T., & Dyrlund, O. (2004). Predicting Hearing Aid Microphone Preference in Everyday Listening. *Journal of the American Academy of Audiology*, *15*(5), 365–396. https://doi.org/10.3766/jaaa.15.5.4

Walker-Andrews, A. S., & Lennon, E. M. (1985). Auditory-Visual Perception of Changing Distance by Human Infants. *Child Development*, *56*(3), 544–548. https://doi.org/10.2307/1129743

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, *64*(Nov), 105–123. https://doi.org/10.1016/j.neuropsychologia.2014.08.005

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*(2), 252–258. https://doi.org/10.1007/s00221-004-1899-9

Wallace, M. T., Woynaroski, T. G., & Stevenson, R. A. (2020). Multisensory Integration as a Window into Orderly and Disrupted Cognition and Communication. *Annual Review of Psychology*, *71*(1), 193–219. https://doi.org/10.1146/annurev-psych-010419-051112

Weatherhead, D., & White, K. S. (2017). Read my lips: Visual speech influences word processing in infants. *Cognition*, *160*, 103–109. https://doi.org/10.1016/j.cognition.2017.01.002

Weinstein, N. D. (1978). Individual differences in reactions to noise: A longitudinal study in a college dormitory. *Journal of Applied Psychology*, *63*(4), 458–466. https://doi.org/10.1037/0021-9010.63.4.458

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, *7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Whitehead, L. C. (2007). Methodological and ethical issues in Internet-mediated research in the field of health: An integrated review of the literature. *Social Science & Medicine (1982)*, *65*(4), 782–791. https://doi.org/10.1016/j.socscimed.2007.03.005

Whitley, M. S. (2002). *Spanish-English contrasts: A course in Spanish linguistics*. Georgetown University Press.

Whitton, J. P., Hancock, K. E., Shannon, J. M., & Polley, D. B. (2017). Audiomotor Perceptual Training Enhances Speech Intelligibility in Background Noise. *Current Biology*, *27*(21), 3237-3247.e6. https://doi.org/10.1016/j.cub.2017.09.014

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *The Journal of the Acoustical Society of America*, *119*(6), 3940–3949. https://doi.org/10.1121/1.2195121

Wilbiks, J. M. P., & Beatteay, A. (2020). Individual differences in multiple object tracking, attentional cueing, and age account for variability in the capacity of audiovisual integration. *Attention, Perception & Psychophysics*, *82*(7), 3521–3543. https://doi.org/10.3758/s13414-020-02062-7

Wilbiks, J. M. P., & Dyson, B. J. (2018). The Contribution of Perceptual Factors and Training on Varying Audiovisual Integration Capacity. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(6), 871–884. https://doi.org/10.1037/xhp0000503

Williams, K. (2014). *Phonological and Print Awareness Scale.* Torrance, CA: WPS Publishing.

Williams, K. T. (2018). *Expressive Vocabulary Test* (3rd ed.). NCS Pearson.

Winneke, A. H., & Phillips, N. A. (2011). Does Audiovisual Speech Offer a Fountain of Youth for Old Ears? An Event-Related Brain Potential Study of Age Differences in Audiovisual Speech Perception. *Psychology and Aging*, *26*(2), 427–438. https://doi.org/10.1037/a0021683

Witton, C., Talcott, J. B., & Henning, G. B. (2017). Psychophysical measurements in children: challenges, pitfalls, and considerations. *PeerJ (San Francisco, CA)*, *5*, e3231–e3231. https://doi.org/10.7717/peerj.3231

Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming-Speed Processes, Timing, and Reading: A Conceptual Review. *Journal of Learning Disabilities*, *33*(4), 387–407. https://doi.org/10.1177/002221940003300409

Woodhouse, L., Hickson, L., & Dodd, B. (2009). Review of visual speech perception by hearing and hearing-impaired people: clinical implications. *International Journal of Language & Communication Disorders*, *44*(3), 253–270. https://doi.org/10.1080/13682820802090281

Woynaroski, T. G., Kwakye, L. D., Foss-Feig, J. H., Stevenson, R. A., Stone, W. L., & Wallace, M. T. (2013). Multisensory Speech Perception in Children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, *43*(12), 2891–2902. https://doi.org/10.1007/s10803-013-1836-5

Yantis, S. (2005). How visual salience wins the battle for awareness. *Nature Neuroscience*, *8*(8), 975–977. https://doi.org/10.1038/nn0805-975

Yeatman, J. D., Caffarra, S., Clarke, M. D., Ender, S., Gijbels, L., Joo, S. J., ... & Taulu, S. (2022). Reading instruction causes changes in category-selective visual cortex. BioRxiv.

Yeung, H. H., & Werker, J. F. (2013). Lip Movements Affect Infants' Audiovisual Speech Perception. *Psychological Science*, *24*(5), 603–612. https://doi.org/10.1177/0956797612458802

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception & Psychophysics*, *81*(2), 558–570. https://doi.org/10.3758/s13414-018-1635-3

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science*, *12*(5), 798–814. https://doi.org/10.1111/j.1467-7687.2009.00833.x

Yu, T.-L. J., & Schlauch, R. S. (2019). Diagnostic Precision of Open-Set versus Closed-Set Word Recognition Testing. *Journal of Speech, Language, and Hearing Research*, *62*(6), 2035–2047. https://doi.org/10.1044/2019_JSLHR-H-18-0317

Yuan, J., He, Y., Qinglin, Z., Chen, A., & Li, H. (2008). Gender differences in behavioral inhibitory control: ERP evidence from a two-choice oddball task. *Psychophysiology*, *45*(6), 986–993. https://doi.org/10.1111/j.1469-8986.2008.00693.x

Yuan, Y., Wayland, R., & Oh, Y. (2020). Visual analog of the acoustic amplitude envelope benefits speech perception in noise. *The Journal of the Acoustical Society of America*, *147*(3), EL246–EL251. https://doi.org/10.1121/10.0000737

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics*, *67*(3), 531–544. https://doi.org/10.3758/BF03193329

Zerr, M., Freihorst, C., Schütz, H., Sinke, C., Müller, A., Bleich, S., Münte, T. F., & Szycik, G. R. (2019). Brief Sensory Training Narrows the Temporal Binding Window and Enhances Long-Term Multimodal Speech Perception. *Frontiers in Psychology*, *10*, 2489–2489. https://doi.org/10.3389/fpsyg.2019.02489

Zhou, H., Cheung, E. F. C., & Chan, R. C. K. (2020). Audiovisual temporal integration: Cognitive processing, neural mechanisms, developmental trajectory and potential interventions. *Neuropsychologia*, *140*, 107396–107396. https://doi.org/10.1016/j.neuropsychologia.2020.107396

Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

Ziegler, J. C., Pech-Georgel, C., George, F., & Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Developmental Science*, *12*(5), 732–745. https://doi.org/10.1111/j.1467-7687.2009.00817.x

Zion Golumbic, E., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *The Journal of Neuroscience*, *33*(4), 1417–1426. https://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zou, Z., Chau, B. K. H., Ting, K.-H., & Chan, C. C. H. (2017). Aging Effect on Audiovisual Integrative Processing in Spatial Discrimination Task. *Frontiers in Aging Neuroscience*, *9*, 374–374. https://doi.org/10.3389/fnagi.2017.00374

Zurif, E. B., & Carson, G. (1970). Dyslexia in relation to cerebral dominance and temporal analysis. *Neuropsychologia*, *8*(3), 351–361. https://doi.org/10.1016/0028-3932(70)90079-5