ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Audiovisual Speech Processing in Relationship to Phonological and Vocabulary Skills in First Graders

Liesbeth Gijbels,[a,b] Jason D. Yeatman,[c,d] Kaylah Lalonde,[e] and Adrian K. C. Lee[a,b]

[a] Department of Speech & Hearing Sciences, University of Washington, Seattle [b] Institute for Learning & Brain Sciences, University of Washington, Seattle [c] Division of Developmental-Behavioral Pediatrics, School of Medicine, Stanford University, CA [d] Graduate School of Education, Stanford University, CA [e] Boys Town National Research Hospital, Center for Hearing Research, Omaha, NE

ARTICLE INFO

Article History:
Received April 2, 2021
Revision received July 6, 2021
Accepted August 11, 2021

Editor-in-Chief: Peggy B. Nelson
Editor: Rachael Frush Holt

https://doi.org/10.1044/2021_JSLHR-21-00196

ABSTRACT

Purpose: 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 (AV) gain in young children's speech understanding. We examined whether there is an 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.
Method: Thirty-seven first graders completed online psychophysical experiments. We used an online single-interval, four-alternative forced-choice picture-pointing task with age-appropriate consonant–vowel–consonant words to measure auditory-only, visual-only, and AV word recognition in noise at −2 and −8 dB SNR. We obtained standard measures of vocabulary and phonological awareness and included a general psychophysical test to examine correlations with AV benefits.
Results: 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.
Conclusions: 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.

Daily listening environments are often noisy, and the presence of background noise degrades or slows down our speech understanding (Mattys et al., 2012; Ross et al., 2006). Speech in noise (SIN) 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 et al., 2016, 2020). Second, they constantly interact in noisy backgrounds, for example, 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 improve 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 nonspeech (Hirst et al., 2020) and speech

Correspondence to Adrian K. C. Lee: akclee@uw.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

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).

AV Benefit in Speech Perception in Adults and Children

AV speech benefit is especially prominent when the auditory signal is degraded (Sumby & Pollack, 1954). AV enhancement in adults with normal-hearing thresholds can result in 6- to 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 et al., 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 elderly 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 with 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 that 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.

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 with 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 with 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 & 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 & 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 when 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 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). However, 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 setup but measure different intrinsic values would allow us to better focus on the experimental factors of interest.

Purpose of This Study

The purpose of this 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 that we asked was whether first graders show significant AV enhancement of SIN. 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 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 (R. F. Holt et al., 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. Furthermore, 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 & Holt, 2016; 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; R. F. Holt et al., 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. On the basis of 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 nonspeech task and, thus, better performance in psychophysical tasks would not be a main factor in explaining AV performance.

Method

Participants

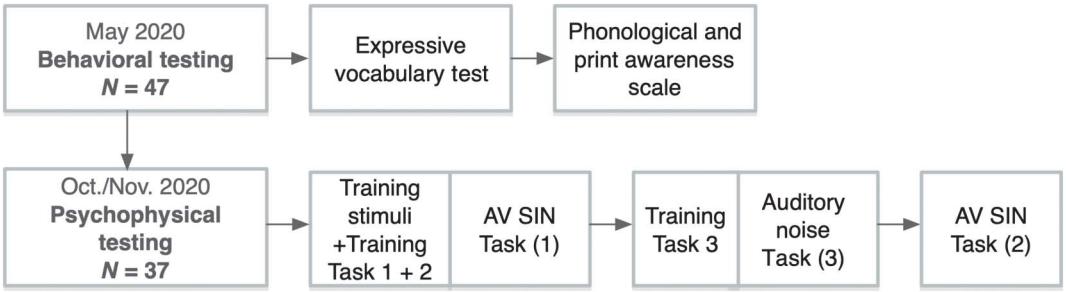
A group of 37 English-speaking children (M = 15, F = 22) participated in this study. Although this study was initially planned in person, it moved online due to COVID-19. These participants were a subset of the 48 children who participated in a 10-day camp in the summer of 2019 at University of Washington's 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 Dunn, 1997), Comprehensive Test of Phonological Processing (Wagner et al., 2013), and Test of Preschool Early Literacy (Lonigan et al., 2007). 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 this study. The other 10 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 3-alternative forced-choice (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.

Experimental Protocol

Figure 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. Because the PPA is a 3AFC task, participants were asked to name the word or the accompanying number (1, 2, or 3) for each

Figure 1. Experimental protocol of behavioral and psychophysical test sessions. Diagram describes the training and testing procedures. Task 1 + 2 represent the audiovisual (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 nonspeech auditory task that assesses general psychophysical testing performance. N denotes the number of participants in each test session.



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-hr online moderated session between October and November 2020. Each task took about 10–15 min when completed without breaks. While up to five 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 two breaks during a task.

Tasks 1 and 2: (AV) SIN Recognition

Stimuli and materials. All participants completed two 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 Child Language Data Exchange System 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 R. F. Holt et al. (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-min 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-s fragments of the noise were mixed with target words at two SNR levels: −2 or −8 dB. These noise fragments were added to a relatively constant speech signal. During the development of the stimuli, the speech signal was roved between 55 and 65 dB SPL and the noise was added, 2 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 or −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, SNRs of 0 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 is consisted of the target picture and three foil pictures: (a) a "minimal pair" foil picture—target and foil picture names only differed in the first or last consonant, (b) a "same-vowel" foil picture—target and foil picture names shared only the same vowel, and (c) 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 Appendix A. 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 the 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 (see Appendix A). 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 initials /g/ and /θ/ are typically not

confused during speechreading, so "gum" and "thumb" are low in visual similarity. In contrast, word initials /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., 2020).

Procedure. Parents were directed to load the tasks on their computer via a website and e-mail 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, and sending the output file back) and to make notes of any unforeseen circumstances (e.g., Internet issues and sibling intervening). 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 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 eight 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 (two AV, two audio-only, and one visual-only; see Figure 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-s 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 2 × 2 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.

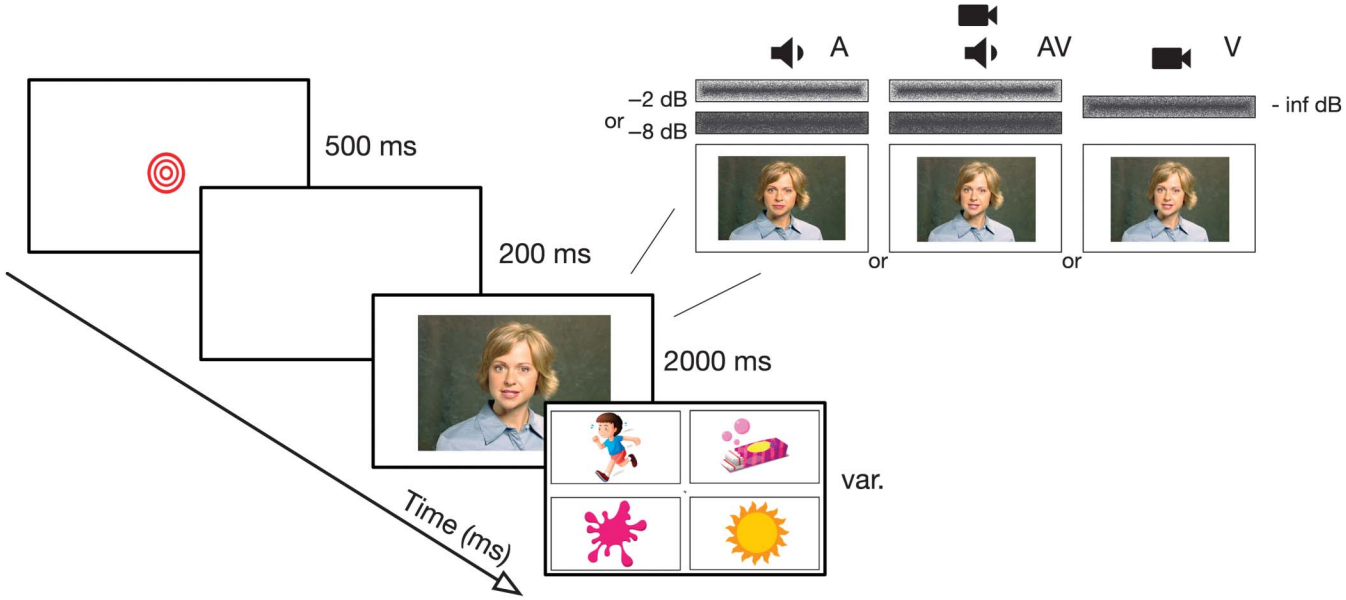
In order to monitor participants' visual attention, a picture of a cartoon character, which also narrated the instructions, was randomly shown on the screen. This character served as a catch-trial stimulus to measure cross-modal attention and was a fun motivator for the participants throughout the tasks. 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, 400, 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 × 1–4 "beeps" × 2 ISIs). These stimuli were presented in the presence of a LTASS noise at three different SNR levels (no noise, −2, and −8 dB). Similar to Tasks 1 and 2, the free online study builder, `lab.js` (Henninger et al., 2020), was used to build this task.

Procedure. The setup of this task was very similar to the previous two tasks. Participants were first trained in a 1-interval–4-alternative forced choice task with feedback. The participants were presented with the same 500-ms fixation marker, followed by a 200-ms blank screen. The participants heard one, two, three, or four "beeps" while seeing a fixed cartoon "listening" character. This was followed by four choices (arranged in a 2 × 2 block) with pictures showing one, two, three, or four dots (with the accompanying number). These choices stayed in the same order throughout

Figure 2. Visualization of setup for Tasks 1 and 2, which started with a 500-ms fixation marker 500 ms, followed by a 200-ms blank screen. Then the auditory, audiovisual (AV), or visual stimulus was presented (in different signal-to-noise ratios, 2,000 ms), followed by the four answer choices. There was no time limit in the response period.



Task 3, in contrast to the AV tasks. Answers were recorded via mouse clicks.

The training consisted of eight practice items with feedback, followed by 60 test trials without feedback, presented in three 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.

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, minimum = 16, maximum = 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.

AV 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 analysis of variance (ANOVA) function provided F statistics for the models generated. We used linear (mixed-effects) models to test two main hypotheses, whether children this age show AV gain and whether individual differences between participants could be explained by vocabulary knowledge, phonological awareness skills, or psychophysical testing performance.

First, we tested the hypothesis that the 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 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 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 1 summarizes how children benefited from visual salience of the phonemes (Model 1), and their corresponding task performance is shown in Figure 3. The random intercept had a standard deviation of 0.01, suggesting that performance across participants was quite consistent.

An 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 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 (see Figure 3). We further explored this relationship between audio-only and AV performance at the two SNR levels (see Figure 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 (see red in Figure 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 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 (see Figure 4). For the highest (> 90%) audio-only scores, we found that more than 60% of the children showed no 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% (see Figure 3, right). Importantly, children scored significantly above chance, $t(36) = 7.6754$, $p < .001$. Only five 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 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

Table 1. Summary of linear-mixed Model 1 with regression estimates, standard errors (*SEs*), *t* scores, and *p* values. The model predicts the test performance.

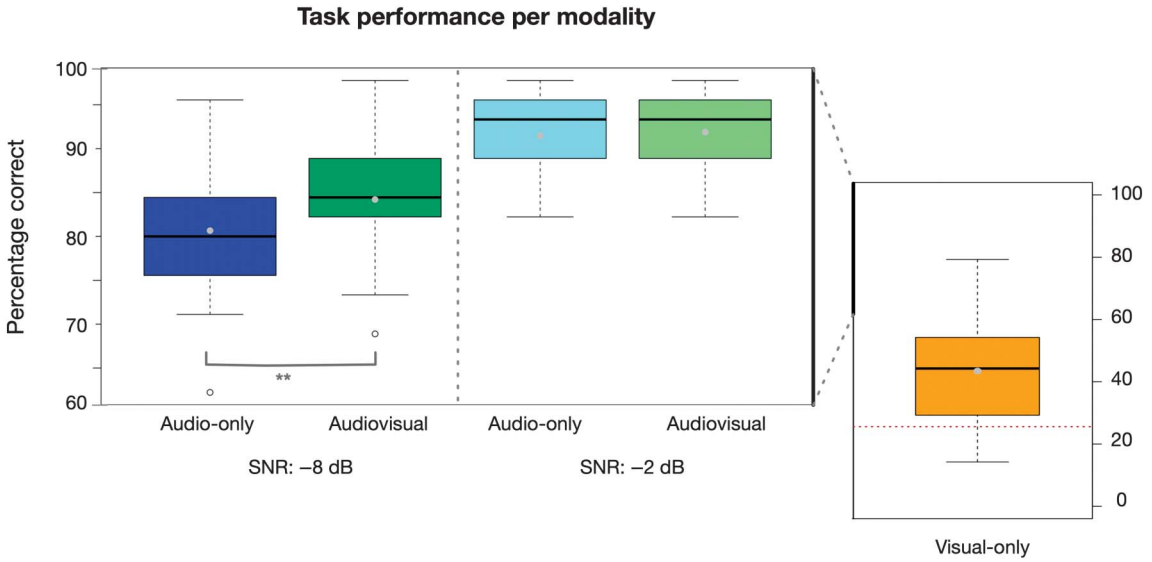
Model 1: Test score (%) ~ Modality × SNR + stimuli set + (1|participant) + (stimuli set|participant)				
Predictor	**Estimate**	***SE***	***t* score**	***p* value**
Intercept	0.797635	0.009551	83.517	**< 2e−16 *****
Modality~	0.03473	0.010644	3.263	**.00128 ****
SNR°	0.107973	0.010644	10.144	**< 2e−16 *****
Set^	0.020135	0.01028	1.959	.05794
Modality~: SNR°	−0.029459	0.015053	−1.957	.05162

Note. Bold values denote statistical significance at the p < .05 level. SNR = signal-to-noise ratio.

***p* < .01. ****p* < .001.

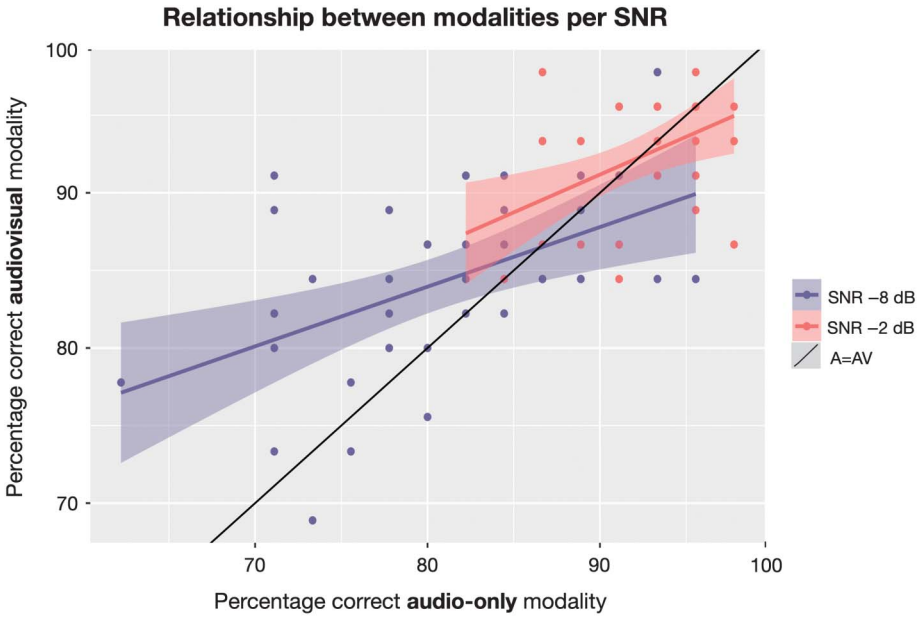
~Reference is audio-only. °Reference is −8 dB. ^Reference is large stimulus set.

Figure 3. Boxplots of percent correct test scores per modality (audio-only, audiovisual, and visual-only) and signal-to-noise ratio (SNR; –8 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× below or above the 25th and 75th percentiles, respectively, and are shown as circles. Significance based on Model 1b: *$p < .05$, **$p < .01$, and ***$p < .001$. The red dotted line in the visual-only modality represents chance level.



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

Figure 4. The relationship between percent correct scores on the audio-only (A) and audiovisual (AV) modalities per participant at each of the signal-to-noise ratio (SNR) levels. Regression lines are plotted with 95% confidence intervals. The black line shows where audio-only and AV performances are equal. Thus, all dots above the black line represent AV benefit.



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 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 2) exploring the relationship between AV processing, vocabulary, phonological awareness, general psychophysical test performance, and gender is presented in Table 2. Variability of both dependent and independent variables are described in Table 3.

The ANOVA 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.

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, nonsignificant, and 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.

Table 2. Output of linear Model 2 with regression estimates, standard errors (*SE*s), *t* scores, and *p* values.

Model 2: AV gain ~ stimuli set × vocabulary + PA + psychophysical + gender

Predictor	Estimate	*SE*	*t* value	*p* value
Intercept	−1.295493	1.05849	−1.224	.2255
Set^	0.14172	0.974434	0.145	.8848
Vocabulary	0.012885	0.007588	1.698	*.0944*
Gender+	0.319262	0.134055	2.382	**.0203***
PA	−0.033742	0.01754	−1.924	*.0589*
Psychophysical task	0.763777	0.765357	0.998	.3221
Set ^ * vocabulary	−0.001892	0.00994	−0.19	.8496

Note. The bold value denotes statistical significance at the $p < .05$ level. ~ is used to define the relationship between dependent variable and independent variables in a statistical model formula. Audiovisual (AV) gain: adjusted (AV-A/1-A), for signal-to-noise ratio (SNR) −8 dB; vocabulary = vocabulary score; psychophysical = psychophysical testing score; PA = phonological awareness score.
^Reference is large stimulus set. +Reference is female. *$p < .05$.

Table 3. Descriptive statistics, showing mean, median, standard deviation, and minimum (Min) and maximum (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

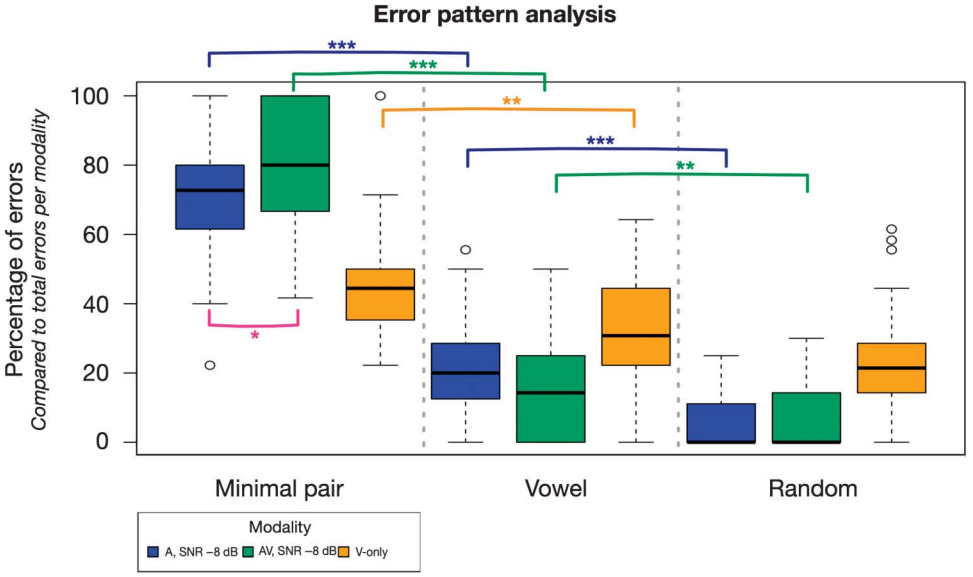
Note. Audio-only, audiovisual, audiovisual (AV) gain, and relative AV gain statistics are calculated of the −8 dB signal-to-noise ratio (SNR) stimuli. Vocabulary and Phonological awareness scores are expressed in raw scores. Audio-only, AV, AV gain, and the control psychophysical task are expressed in percentages.

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, and 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).

Figure 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 with 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

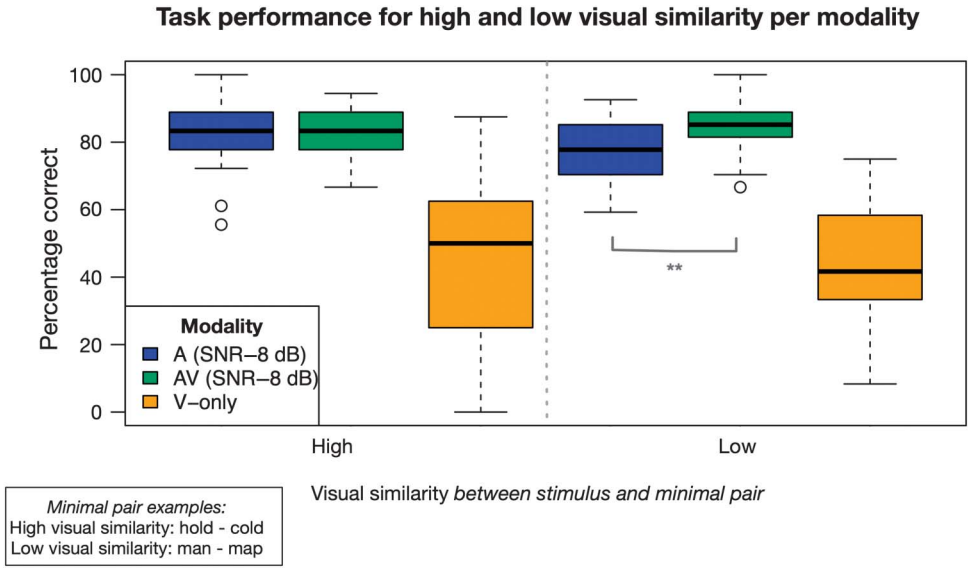
Figure 5. Error pattern analysis; percentage of errors by error type (minimal pair, vowel, and random) and modality (audio-only [A], audiovisual [AV], and visual-only [V-only]). The percentage on the *y*-axis is the percentage of errors in relation to total errors per modality. Significance: **p* < .05, ***p* < .01, and ****p* < .001.



speechreading without auditory input is much more difficult than audio-only or AV judgments, 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 and 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 Appendix B.

All our minimal pairs were auditory minimal pairs. We therefore classified the minimal pairs by visual similarity (see Appendix A; e.g., high visual similarity: "hold"–"cold"; low visual similarity: "man"–"map") to further analyze the utility of the visual information. Visual similarity analyses

Figure 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 (A), visual-only (V-only), or audiovisual [AV]), significance: **p* < .05, ***p* < .01, and ****p* < .001. Examples of a minimal pair with high visual similarity: "hold"–"cold" and low visual similarity: "man"–"map."



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 (see Figure 6). 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 ($p = .811$).

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 that 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, 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 nonsignificant trend toward 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 (R. Holt et al., 2020; Lewis et al., 2016) 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 speech reading 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 ($M = 44\%$ correct), they might not have developed speech reading skills required by the use of an open-set task. R. F. Holt et 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 this 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 SIN 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 4AFC 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 increase intelligibility.

Relationship Between AV Gain and Other Factors

Exploratory analysis showed, in both −2 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 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 setup but no visual component to confirm that any individual differences that we observed were due to more than task performance. Although there was, as expected, a small but nonsignificant, 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 1, it was a significant predictor for the relative AV gain, $(AV - A)/(1 - A)$, in Model 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 as 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 show 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 setup (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 setup, and 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 3–5 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 setup 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.

Conclusions

This study showed that children in first grade can use visual speech to improve accuracy of SIN perception at low SNRs. Their performance gain was dependent on the salience of the visual cues in SIN tasks. This was evident from their speechreading performance and 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.

Acknowledgments

This work was supported by a National Institute of Child Health and Human Development Grant (R21HD092771) to J. D. Y. We would like to thank Rachael F. Holt for the use of the audiovisual stimulus materials and Eric Larson and Hans Reyserhove for coding support.

References

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

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

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*(1–2), 38–50. https://doi.org/10.1016/j.jecp.2009.08.005

Bates, D., Maechler, 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

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

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

Bishop, C. W., & Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PLOS ONE, 6*(8), Article e24016. https://doi.org/10.1371/journal.pone.0024016

Bjorklund, D. (2005). *Children's thinking: Cognitive development and individual differences* (4th ed.). Thomson/Wadsworth.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* [Computer program]. Version 6.1.41. Retrieved March 25, 2021, from http://www.praat.org/

Buss, E., Hall, J. W., & Grose, J. H. (2011). Development of auditory coding as reflected in psychophysical performance. In L. Werner, R. Fay, & A. Popper (Eds.), *Human auditory development (Springer Handbook of Auditory Research)* (Vol. 42,

pp. 107–136). Springer. https://doi.org/10.1007/978-1-4614-1421-6_4

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

Campbell, C. S., & Massaro, D. W. (1997). Perception of visible speech: Influence of spatial quantization. *Perception, 26*(5), 627–644. https://doi.org/10.1068/p260627

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

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

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

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

Dunn, L. (1997). *PPVT-III: Peabody Picture Vocabulary Test–Third Edition*. American Guidance Service.

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

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

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

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

Grant, K. W., & Bernstein, J. G. W. (2019). Toward a model of auditory-visual speech intelligibility. In A. Lee, M. Wallace, A. Coffin, A. Popper, & R. Fay (Eds.). *Multisensory processes (Springer Handbook of Auditory Research)* (Vol. 68, pp. 33–57). Springer. https://doi.org/10.1007/978-3-030-10461-0_3

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., & 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

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

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B E. (2020). *lab.js: A free, open, online study builder*. https://doi.org/10.31234/osf.io/fqr49

Hillock-Dunn, A., & Wallace, M. T. (2012). Developmental changes in the multisensory temporal binding window persist into adolescence. *Developmental Science, 15*(5), 688–696. https://doi.org/10.1111/j.1467-7687.2012.01171.x

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

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., Bruggeman, L., & Demuth, K. (2020). Visual speech cues speed processing and reduce effort for children listening in quiet and noise. *Applied Psycholinguistics, 41*(1), 933–961. https://doi.org/10.1017/S0142716420000302

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)

Jerger, J., Speaks, C., & Trammell, J. L. (1968). A new approach to speech audiometry. *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

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

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

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)

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., & 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., & 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. (2021). Development of the mechanisms underlying audiovisual speech perception benefit. *Brain Sciences, 11*(1), 49. https://doi.org/10.3390/brainsci11010049

Leibold, L. J., & Buss, E. (2019). Masked speech recognition in school-age children. *Frontiers in Psychology, 10,* 1981. https://doi.org/10.3389/fpsyg.2019.01981

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

Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). Test of Preschool Early Literacy. Pro-Ed.

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, 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), Article e4638. https://doi.org/10.1371/journal.pone.0004638

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,* Article e04995. https://doi.org/10.7554/eLife.04995

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

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

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. (2016). Individual differences in language and working memory affect children's speech recognition in noise. *International Journal of Audiology, 56*(5), 306–315. https://doi.org/10.1080/14992027.2016.1266703

McFadden, D. (1998). Sex differences in the auditory system. *Developmental Neuropsychology, 14*(2–3), 261–298. https://doi.org/10.1080/87565649809540712

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature (London), 264*(5588), 746–748. https://doi.org/10.1038/264746a0

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

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. https://doi.org/10.1016/j.heares.2019.107773

Nelson, P. B., & Soli, S. (2000). Acoustical barriers to learning. *Language, Speech, and 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), 133–139. https://doi.org/10.1515/JBCPP.1997.8.3.133

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

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *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., & Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *The European Journal of Neuroscience, 33*(12), 2329–2337. https://doi.org/10.1111/j.1460-9568.2011.07685.x

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex (New York, NY 1991), 17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

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. https://doi.org/10.3389/fnins.2015.00185

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

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

Stein, B., & Meredith, M. A. (1993). *The merging of the senses (cognitive neuroscience series)*. MIT Press.

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., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Human Perception and Performance, 38*(6), 1517–1529. https://doi.org/10.1037/a0027339

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

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

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

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

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing–Second Edition*. Pro-Ed.

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

Williams, K. (2014). *Phonological and Print Awareness Scale*. 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, 5,* E3231. https://doi.org/10.7717/peerj.3231

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2018). 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

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

Appendix A

Large and Small Stimulus Set With Accompanying Minimal Pair, Vowel Alternative, Random Alternative, and Assigned Visual Similarity Scores

Large stimulus set

	Goal stimulus	Minimal pair	Vowel	Random	Visual similarity
1	cold	hold	goat	fish	high
2	hold	cold	goat	fish	high
3	hand	stand	bag	thumb	high
4	stand	hand	bag	thumb	high
5	sun	run	gum	pink	low
6	run	sun	gum	pink	low
7	snake	cake	whale	nine	low
8	cake	snake	whale	nine	low
9	thumb	gum	sun	sit	low
10	gum	thumb	sun	sit	low
11	kick	stick	sing	man	high
12	ball	wall	log	goat	high
13	ten	men	bed	mouth	low
14	pig	pink	sit	bath	high
15	pink	pig	sit	bath	high
16	sit	sing	pig	run	high
17	sing	sit	pig	run	high
18	bag	bath	hand	sing	low
19	bath	bag	hand	sing	low
20	bed	bell	ten	snake	low
21	goat	goal	hold	bed	low
22	fish	fin	kick	cold	low
23	mouth	mouse	couch	ball	low
24	man	map	stand	gum	low
25	nine	knife	fight	cake	low

Small stimulus set

	Goal stimulus	Minimal pair	Vowel	Random	Visual similarity
1	hand	stand	bag	thumb	high
2	stand	hand	bag	thumb	high
3	sun	run	gum	bath	low
4	run	sun	gum	bath	low
5	gum	thumb	sun	sit	low
6	thumb	gum	sun	sit	low
7	sit	sing	pig	run	high
8	sing	sit	pig	run	high
9	bag	bath	hand	sing	low
10	bath	bag	hand	sing	low

blue = change of initial consonant
green = change of final consonant
orange = word not presented as stimulus word in this set

Appendix B

Pairwise Comparisons Using Wilcoxon Rank Sum Test (False Discovery Rate Corrected Using Benjamini–Hochberg Method) for Errors in Comparison to Total Errors per Modality

Modality	Error type	*p* value
Audio-only	Minimal pair vs. vowel	**1.5e–06*** **
	Vowel vs. random	**1.7e–05*** **
Audiovisual	Minimal pair vs. vowel	**6.1e–07*** **
	Vowel vs. random	**.00209****
Visual-only	Minimal pair vs. vowel	**.0085****
	Vowel vs. random	.05203
Audio-only vs. audiovisual	Minimal pair	**.04731***
	Vowel	.07376
	Random	.67674
Audio-only vs. video-only	Minimal pair	**2.7e–06*** **
	Vowel	**.00759****
	Random	**5.0e–06*** **
Audiovisual vs. video-only	Minimal pair	**2.0e–06*** **
	Vowel	**6.1e–05*** **
	Random	**1.1e–05*** **

Note. Bold values are significant.

*$p < .05$. **$p < .01$. ***$p < .001$.