VISUAL SPEECH CUES IMPROVE CHILDREN'S PROCESSING SPEED IN BOTH QUIET AND NOISE

Rebecca Holt[1], Laurence Bruggeman[1,2] & Katherine Demuth[1]

[1]Department of Linguistics, Macquarie University & ARC Centre of Excellence in Cognition and its Disorders
[2]MARCS Institute, Western Sydney University & ARC Centre of Excellence for the Dynamics of Language
rebecca.holt | laurence.bruggeman | katherine.demuth@mq.edu.au

ABSTRACT

The presence of visual cues can facilitate speech processing in adults, conferring an 'audiovisual (AV) benefit' in noisy listening conditions. However, it is unclear to what extent such benefits extend to quiet conditions and to children. A phoneme monitoring task was used to determine whether 7-11 year-old children show an AV benefit for accuracy and/or speed of processing in either quiet or noise, and whether the magnitude of this benefit differs between listening conditions. An AV benefit for processing speed was found, unaffected by listening conditions. This suggests that visual speech cues can be used by children to facilitate speech processing generally, not just in noise. We therefore believe that visual speech cues may be used to assist children to rapidly process speech in everyday situations.

Keywords: sentence processing, audiovisual, children, noise, phoneme monitoring

1. INTRODUCTION

Speech perception and processing often involve both auditory and visual speech cues. The integration of these cues can be advantageous to listeners in comparison to speech presented in the auditory modality alone. One advantage of audiovisual (AV) presentation is that it improves the *speed* and *accuracy* of speech perception/processing in noise for adult listeners [8, 11, 18, 19]. This processing advantage associated with AV presentation over auditory-only (AO) is known as the 'AV benefit'. While there is strong evidence for AV benefits arising in noisy listening conditions, some studies have concluded that AV benefits do not occur in quiet conditions, e.g., [19]. However, this may be due to ceiling performance in AO conditions, leaving no room for improvement with the addition of visual cues. Indeed, studies have found that adults do show an AV benefit in quiet if the task is sufficiently difficult [7, 17]. This suggests that the presence of an AV benefit is not attributable to the presence of noise *per se*, but to the degree of difficulty associated with processing the input, to which noise may contribute.

In comparison, AV benefits among children are relatively poorly understood. While it is known that AV presentation can improve *accuracy* of speech perception among children in both noise [9, 12] and vocoded speech [15], no study has examined whether this is accompanied by an AV benefit for processing *speed* (cf. [8, 11] for adults). Additionally, no study has yet examined whether children show any AV benefit in quiet. The current study therefore asked whether, for speech in quiet conditions, children show an AV benefit, in either accuracy or speed, and how this compares to any AV benefit found in noise.

It is important to address these questions for three reasons. Firstly, by better understanding the circumstances in which children show an AV benefit, those who interact with children may be better equipped to facilitate children's speech processing. If children show an AV benefit for processing speed, ensuring that visual cues are available to them in the classroom, for example, may improve their learning: Faster processing may give them more time to comprehend lesson content and assist children who struggle to 'keep up' during lessons. Secondly, it is necessary to understand the degree of AV benefit experienced by typically-developing children as a baseline for investigating special populations of children for whom AV benefit may be of extra importance, such as children with hearing impairment or learning difficulties. Finally, investigating children's AV benefit in quiet conditions may provide evidence in support of the suggestion that the presence of an AV benefit is determined not necessarily by the presence of noise, but by the difficulty associated with processing the speech input.

To investigate children's use of visual speech cues, we used a phoneme monitoring task [4, 6]: Participants listened to sentences presented in either AV or AO modality and made a button-press response as soon as they heard a pre-specified phoneme. Half of the participants completed the task in quiet and half in noise. The participants' monitoring accuracy and reaction times were measured to determine the presence of any AV accuracy and/or speed benefit. This method was chosen as phoneme monitoring is relatively difficult for children, making the presence of an AV benefit more likely, and, by measuring reaction times, avoids

the issue of ceiling performance as found in previous studies. We hypothesised that, in noise, children would show an AV benefit for accuracy, as in [9, 12], and a corresponding benefit in processing speed, as found for adults in [8, 11]. We also hypothesised that the same AV benefits would be found in quiet, due to the difficulty of the phoneme monitoring task for children. However, we expected that the AV benefits found in noise would be greater than in quiet due to the increased difficulty associated with processing in the noisy condition.

One possible drawback to using phoneme monitoring to examine AV processing is that an apparent AV speed benefit may actually be due to the temporal ordering of visual and auditory cues. For example, when monitoring for the phoneme /b/, the first visual cue (lip closure) becomes available before the first auditory cue (release burst). It would therefore be possible to observe shorter reaction times in the AV than the AO condition due to the availability of the different cues, rather than any change in processing speed. To mitigate this potential issue, we included phonemes at two places of articulation (PoAs): bilabial (/b,p/) and velar (/g, k/). The bilabial phonemes had a visual cue that preceded the auditory cue, as described above, but velars did not, as velar closure is not visually salient. Should the same apparent AV benefit be found across both PoAs, this would demonstrate that the benefit was due to an overall difference in processing speed, while an apparent benefit for bilabial, but not velar phonemes would implicate the temporal difference between the visual and auditory cues.

2. METHODS

2.1. Participants

Thirty-nine children aged 7-11 years participated in the study (M_{age} = 8 years, 10 months, *SD* = 1 year, 3 months**;** 18M, 21F). One further participant was excluded for failing to follow task instructions. Nineteen participants completed the task in quiet (M_{age} = 8 years, 11 months, *SD* = 1 year, 5 months**;** 10M, 9F) and 20 in noise (M_{age} = 8 years, 10 months, *SD* = 1 year, 1 month**;** 8M, 12F). All participants were monolingual native speakers of English, with native English-speaking parents. No participant had any reported hearing, language or cognitive impairment, or any vision impairment not corrected by glasses. Parental written informed consent and participant's verbal assent were obtained.

2.2. Stimuli

For the phoneme monitoring task, /b, p, g, k/ were chosen as target phonemes, each occurring in twelve target words. The target phoneme always occurred as a singleton in word-initial position, to control for singleton vs. cluster and word position effects on reaction time [6, 20]. All target words were familiar to children, with a log frequency of at least 3.00 in the CBBC section of the SUBTLEX database [21].

Each target word was embedded in a sentence of 9-12 syllables that did not contain any other occurrences of the target phoneme. The target phoneme always occurred in the fifth or sixth syllable, to control for sentence-position effects on reaction time [6]. Sentences were adapted from [10] (see Table 1 for examples).

Table 1: Sample sentences used in the phoneme monitoring task (target phoneme in **bold**).

Target phoneme	Sample sentence
/b/	The ladies put their **b**asket in the car.
/p/	The artist left his **p**aint on the floor.
/g/	The auntie bought a **g**ift for the little boy.
/k/	The girl left the **c**abbage on her plate.

Sixteen catch sentences were also created, which contained no instances of any target phonemes but followed the constraints of the test sentences. As the position of the target phoneme was consistent across the test sentences, catch sentences were included to prevent participants from responding at the same position in each sentence without paying attention.

All test and catch sentences were spoken by a female native speaker of Australian English and recorded using a Sony HXR-NX30P digital HD video camera with a Sony ECM-XMI electret condenser microphone. The speaker's whole face and shoulders were visible and centred in the recording. The speaker wore a black t-shirt and was shown in front of a solid grey background. Each sentence was produced with the target word prosodically focused to facilitate phoneme monitoring and to control for prosodic effects on reaction time [6].

Video recordings were segmented in iMovie and the audio tracks were extracted. The mean intensity of the sentence portion of each audio track was normalised in Praat [3] and a second version of each audio track was created with overlaid pink noise at a -2 dB signal-to-noise ratio (SNR) using Praat Vocal Toolkit [5], for use in the noise condition. For the AV stimuli, the audio and video tracks were recombined. For the AO stimuli, the audio tracks were matched with a static frame taken from an AV stimulus in which the speaker looked at the camera with her mouth closed and had a smiling expression. Each stimulus sentence appeared in the AV condition for half the participants and AO for the other half.

2.3. Procedure

First, all participants completed a simple reaction time task to train them to respond to stimuli as quickly as possible, so as to obtain valid reaction times. Participants viewed a black screen on which a white cross appeared at irregular intervals and were instructed to press a button on the response pad as quickly as possible whenever they saw the cross.

The phoneme monitoring task then began with a block of practice trials to familiarise participants with the task procedure, followed by eight test blocks (four AV blocks and four AO, one target phoneme per block). Half of the participants received the AV blocks first and the others the AO blocks first. Each block began with an introductory sentence (presented in the same modality as the rest of the block) to inform participants which phoneme they were required to monitor, e.g., "Listen for the /b/, as in band and beetles." This was followed by six test trials and two catch trials, presented in randomised order. At the end of each block a video of the speaker making a funny face was shown to maintain participants' attention.

The task was presented in E-Prime 3 on an ASUS X550J laptop computer in a quiet room, using a Genelec 8020C external speaker (at a constant volume across participants) and a Cedrus RB-840 response pad. Participants were instructed to listen to the sentences while watching the screen and press a designated button on the response pad with their dominant hand as quickly as possible when they heard the target phoneme, but to avoid pressing the button when the target phoneme was not present.

2.4. Analysis

Before data analysis, all catch trials and misses (test trials for which no button-press response was made) were removed from the dataset, leaving only test trials for which participants made a response. Then, trials with unfeasibly short (less than 100 ms) and long (greater than 3000 ms) reaction times were also removed, following, e.g., [10]. In total, 106 of 1872 test trials (5.7%) were removed: 58 were misses (3.1%) and 48 had extreme reaction times (2.6%).

All statistical analyses were carried out in R [16]. For the accuracy analysis, percent correct was calculated by dividing the number of hits (i.e., test trials remaining after misses and trials with extreme reaction times were removed) by the total number of trials for each modality and PoA per participant. Participants who scored 100% correct across all conditions (n = 17) were excluded from the accuracy analysis due to lack of variance in their responses. Percent correct values per participant per condition from the remaining participants (n = 22) were used as the dependent variable in a generalised linear mixed-effects model with family 'inverse Gaussian' and identity link function [14] using the lme4 package [2]. The fixed factors were Modality (AV vs. AO), Listening condition (quiet vs. noise) and PoA (bilabial vs. velar; all contrasts were deviation coded). A maximal random effects structure was used, with a random intercept for Participant and random slopes for Modality and PoA by Participant. The model syntax was: *glmer(Accuracy ~ Modality * ListeningCondition * PoA + (1 + Modality + PoA | Participant), data = AccuracyData, family = inverse.gaussian(link = "identity"))*.

After model fitting, outliers were removed from the data. Outliers were defined as data points for which the standardised residuals were greater than 2.5 standard deviations from the mean [1]. Two outliers were removed (2.3% of the data), then the model was re-fit and tested for statistical significance. Post-hoc pairwise comparisons on significant interactions were performed using the emmeans package [13].

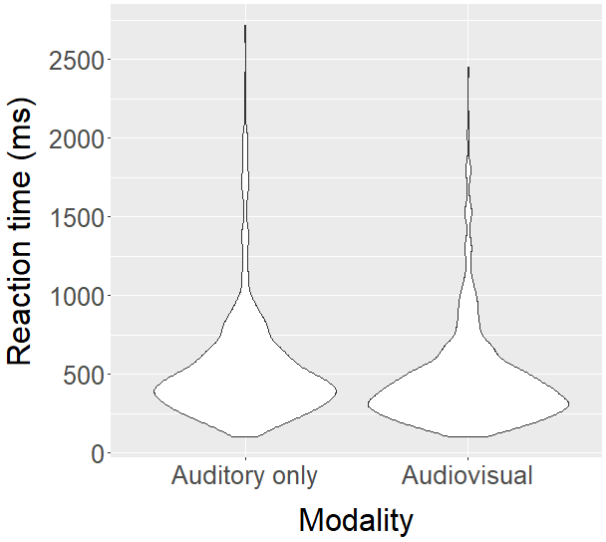
For the reaction time analysis, untransformed reaction time was used as the dependent variable in another generalised linear mixed-effects model, again with family 'inverse Gaussian' and identity link function [14]. The fixed factors were the same as for the accuracy analysis and the random effects structure was again maximal, including random intercepts for Participant and Item and random slopes for Modality and PoA by Participant and Modality and Listening condition by Item. To facilitate model convergence, the nAGQ argument was set to zero. The model syntax was: *glmer(RT ~ Modality * ListeningCondition * PoA + (1 + Modality + PoA | Participant) + (1 + Modality + ListeningCondition | Item), data = RTData, family = inverse.gaussian(link = "identity"), nAGQ = 0)*. Outlier identification was conducted as previously and 49 data points were removed (2.8% of the data). The model was then re-fit and tested for statistical significance.

3. RESULTS

Participants were highly accurate at monitoring the target phonemes: Mean accuracy per participant was 97% (SD = 7%). Modelling revealed no significant main effects or two-way interactions of the fixed factors (Modality, Listening condition and PoA; all ps ≥ .59), but a significant three-way interaction was present (β = -1.10, SE = 0.35, p = .002). Post-hoc tests compared percent correct in the AV and AO modalities for all combinations of Listening condition and PoA, but none were significant (all ps ≥ .38).

For the reaction time analysis, a significant main effect of Modality was found (β = -34.31, SE = 6.74, p < .001), indicating that participants responded 34 ms faster to AV than AO stimuli on average (Figure 1). No other main effects or interactions of the fixed factors were significant (all ps ≥ .05).

Figure 1: Violin plots of phoneme monitoring reaction times across AO and AV conditions.



4. DISCUSSION

Using a phoneme monitoring task, this study investigated whether an AV benefit in accuracy or processing speed occurs among children listening in either quiet or noisy conditions. Participants' percentage of correct responses (i.e., successful monitoring) was used to investigate accuracy benefit, and reaction times were used to examine processing speed benefit.

Results of the accuracy analysis did not reveal any significant AV benefit in either quiet or noise. However, this was likely due to the ceiling or near-ceiling accuracy of most participants across both modalities. It seems that, like in previous studies, this ceiling performance has prevented any potential AV benefit from appearing, cf. [19]. Unlike in [19], however, we observe this ceiling performance across both quiet and noisy conditions. While a significant interaction between Modality, Listening condition and PoA was found, post-hoc tests did not elucidate where the significant effect arose. This may be due to the limited number of data points per cell available for the pairwise comparisons. The accuracy data also suffered from a lack of variance, as only a limited number of values were possible (each data point was calculated from 12 responses, so one error equated to an 8.3% reduction in accuracy), which may also have contributed to the lack of significant effects. Our accuracy results are therefore largely uninformative.

However, as expected, the reaction time data was not subject to the same ceiling effects. In the reaction time analysis, a main effect of Modality was found: Participants responded to AV stimuli 34 ms faster than to AO stimuli, regardless of PoA or Listening condition. This demonstrates that children do indeed show an AV benefit in processing speed in both quiet and noisy conditions, confirming our hypothesis. Furthermore, as this effect of Modality did not interact with PoA, we can be confident that this AV benefit was not driven by the earlier availability of visual than auditory cues, as the same benefit was found for both visually salient bilabial phonemes and non-visually-salient velar phonemes. This suggests that the difference in reaction times is indeed due to facilitated processing in the AV condition.

No interaction was observed between Modality and Listening condition: The same degree of AV benefit was found in both the quiet and noisy listening conditions. This contrasted with our hypothesis that the AV benefit found in noise would be greater than in quiet, and may be due to the relatively favourable SNR of -2 dB in the noise condition. Previous studies using phoneme monitoring to investigate AV benefit have used more negative SNRs: -9 dB for children [9] and -9 to -18 dB for adults [8]. However, these studies involved single-word stimuli, motivating our choice of a less-difficult SNR for the more complex sentence stimuli in our task. Had a more challenging SNR been used in our study, a Listening condition by Modality interaction may have become apparent. Interestingly, no overall difference in reaction time was found between the quiet and noisy conditions, further suggesting that the SNR chosen for the noise condition may not have been sufficiently difficult to disrupt processing. An alternative explanation is that manipulating Listening condition between-subjects may have resulted in insufficient power to detect an effect, due to extensive individual variability.

Our findings have two main implications. Theoretically, our results are compatible with viewing AV benefits as an intrinsic part of speech perception and processing which manifest when speech processing performance is not at ceiling, rather than a phenomenon which occurs only under certain circumstances (i.e., in noise). More practically, the overall AV benefit for processing speed suggests that ensuring that children's auditory input is supplemented by visual speech cues may facilitate speech processing in both quiet and noise. Provision of visual cues may therefore facilitate, for example, learning in classroom settings, following conversation, and other situations where rapid understanding of language input is required.

5. ACKNOWLEDGEMENTS

We thank Elise Tobin and Julien Millasseau for assistance with stimuli recording, Craig Richardson for technical assistance, Peter Humburg for statistical advice, and the Child Language Lab, Macquarie University, for their feedback.

6. REFERENCES

[1] Baayen, R. H., Milin, P. 2010. Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12-28.

[2] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw*, 67, 1-48.

[3] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer. http://www.praat.org/.

[4] Connine, C. M., Titone, D. 1996. Phoneme monitoring. *Lang Cognitive Proc*, 11, 635-646.

[5] Corretge, R. 2012. Praat vocal toolkit. http://www.praatvocaltoolkit.com/.

[6] Cutler, A., Norris, D. 1979. Monitoring sentence comprehension. In: Garrett, M. F., Cooper, W. E., Walker, E. C. T. (eds), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale, NJ: L. Erlbaum Associates, 113-134.

[7] Davis, C., Kim, J. 2004. Audio-visual interactions with intact clearly audible speech. *Q J Exp Psychol-A*, 57, 1103-1121.

[8] Fort, M., Spinelli, E., Savariaux, C, Kandel, S. 2010. The word superiority effect in audiovisual speech perception. *Speech Commun*, 52, 525-532.

[9] Fort, M., Spinelli, E., Savariaux, C, Kandel, S. 2012. Audiovisual vowel monitoring and the word superiority effect in children. *Int J Behav Dev*, 36, 457-467.

[10] Holt, C. M., Demuth, K., Yuen, I. 2016. The use of prosodic cues in sentence processing by prelingually deaf users of cochlear implants. *Ear Hearing*, 37, e256-e262.

[11] Jesse, A., Janse, E. 2012. Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners. *Lang Cognitive Proc*, 27, 1167-1191.

[12] Lalonde, K., Holt, R. F. 2015. Preschoolers benefit from visually salient speech cues. *J Speech Lang Hear R*, 58, 135-150.

[13] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. 2018. emmeans: Estimated marginal means, aka least-squares means. https://cran.r-project.org/package=emmeans.

[14] Lo, S., Andrews, S. 2015. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Front Psychol*, 6, 1171.

[15] Maidment, D. W., Kang, H. J., Stewart, H. J., Amitay, S. 2015. Audiovisual integration in children listening to spectrally degraded speech. *J Speech Lang Hear R*, 58, 61-68.

[16] R Core Team. 2018. R: A language and environment for statistical computing. https://www.r-project.org/.

[17] Reisberg, D., McLean, J., Goldfield, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Dodd, B., Campbell, R. (eds), *Hearing by eye: The psychology of lip-reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, 97-114.

[18] 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. *Cereb Cortex*, 17, 1147-1153.

[19] Sumby, W. H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am*, 26, 212-215.

[20] Treiman, R., Salasoo, A., Slowiaczek, L. M., Pisoni, D. B. 1982. Effects of syllable structure on adults' phoneme monitoring performance. In: Pisoni, D. B. (ed), *Research on speech perception progress report no. 8*. Bloomington, IN: Indiana University Speech Research Laboratory, 63-81.

[21] Van Heuven, W. J. B., Mandera, P., Keuleers, E., Brysbaert, M. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Q J Exp Psychol*, 67, 1176-1190.