Reassessing the Benefits of Audio-Visual Integration to Speech Perception and Intelligibility

Brandon O'Hanlon[1], Professor Christopher J. Plack[1,2], Dr Helen E Nuttall[1]

1. Department of Psychology, Lancaster University, UK
2. Manchester Centre for Audiology & Deafness, The University of Manchester, UK

Corresponding Author: Dr Helen E Nuttall (h.nuttall1@lancaster.ac.uk)

Conflict of Interest: The authors declare no conflicts of interest.

Abstract

Purpose: In difficult listening conditions, the visual system assists with speech perception through lipreading. Stimulus onset asynchrony (SOA) is used to investigate the interaction between the two modalities in speech perception. Previous estimates of audiovisual benefit and SOA integration period differ widely. A limitation of previous research is a lack of consideration of visemes - categories of phonemes defined by similar lip movements when produced by a speaker - to ensure that selected phonemes are visually distinct. This study aimed to reassess the benefits of audiovisual lipreading to speech perception when different viseme categories are selected as stimuli and presented in noise. The study also aimed to investigate the effects of SOA on these stimuli.

Method: Sixty participants were tested online and presented with audio-only and audiovisual stimuli containing the speaker's lip movements. The speech was presented either with or without noise and had six different SOAs (0, 200, 216.6, 233.3, 250, and 266.6 ms). Participants discriminated between speech syllables with button presses.

Results: The benefit of visual information was weaker than that in previous studies. There was a significant increase in reaction times as SOA was introduced, but no significant effects of SOA on accuracy. Furthermore, exploratory analyses suggest that the effect was not equal across viseme categories: 'Ba' was more difficult to recognise than 'Ka' in noise.

Conclusion: In summary, the findings suggest that the contributions of audiovisual integration to speech processing are weaker when considering visemes but are not sufficient to identify a full integration period.

Keywords: audiovisual speech, speech perception, multisensory integration, visemes, vision

Reassessing the Benefits of Audio-Visual Integration to Speech Perception and Intelligibility

Intelligible speech is built up from speech phonemes. Phonemes are small linguistic units - such as the /b/ phoneme that begins ‘boy’ in the English language - and play a large role in the identification of speech (Ewen & Van der Hulst, 2001; Bowers *et al.*, 2016). Processing of speech can be made more difficult with the introduction of noise in the environment, which reduces the ability to discriminate successfully between phonemes (Summerfield, 1992). In many cases, information from the visual sense that is relevant to the speech – such as from lipreading – can be integrated into speech processing systems to improve comprehension. In background noise, this assisting sense is recruited further (Yuan *et al*., 2021). Viewing the lip movements when an individual is speaking can help to improve the intelligibility of speech-in-noise versus when the lips are not visible (Sumby & Pollack, 1954; Maier *et al.*, 2011). The inverse can also occur, wherein incongruent lip movements influence our ability to discriminate between speech sounds. An example of this is the McGurk Effect (McGurk & MacDonald, 1976), where presenting the speech phoneme ‘Ba’ with visual lip movements associated with ‘Ga’ leads to perceptions of the sound ‘Da’ instead. A more recent example comes from face mask-wearing due to the COVID-19 pandemic. Brown *et al.* (2021) found that if the speaker wore a facemask that either fully or partially covered lip movements, performance on speech discrimination tasks decreased dramatically. These data indicate that the visual and auditory systems interact to influence how we perceive speech.

However, estimates of audiovisual benefit vary widely in the literature, likely due to stimulus-dependent effects (Ma *et al.*, 2009), in that, how the stimuli are created for lab experimentation drastically affects how participants respond to speech discrimination tasks. For example, whilst it is important for research on audio-visual processing to consider how auditorily distinctive sounds are, visual distinctiveness is equally important. A way to examine the effect of visual distinctiveness is to select phonemes from separate viseme categories for testing. A viseme category is a group of phonemes from the English language that share the lip movements and visual information portrayed by each phoneme when spoken (Massaro *et al.*, 2012). Fisher (1968) identified five viseme categories based purely on visual distinguishability for English phonemes. Examples of phonemes that belong to the same viseme category are /b/, /p/, and /m/ which in syllable form can correspond to 'Ba', 'Pa', and 'Ma'. If two speech tokens share the same viseme, then it is impossible to discern which was spoken through lip-reading alone (Van Engen *et al*., 2022) and are only distinguishable through sound. This means that any measure of audiovisual benefit derived from discriminating within a viseme category will be lessened. It is therefore important to select stimuli from separate viseme categories when investigating how auditory and visual systems work together during speech syllable discrimination.

When speaking with others, we typically see lip movements before we hear the spoken words (Chandrasekaran *et al*., 2009) as a form of natural stimulus onset asynchrony (SOA). SOA is when two different modalities of information in cross-modal stimuli are presented at different onsets. The window of integration is the term given to the period in which visual information can lead or lag speech sounds before the visual information is no longer perceived as part of the same stimulus (Stein & Meredith, 1993). If the lip movements are desynchronised from the speech sounds within a specific period, then we still perceive the lip movements and the speech we hear to be congruent. If the SOA is large enough that the auditory and visual information do not fall within the same window of integration, we may perceive the two modalities as separate, and therefore not process the visual information as helpful extra information to discern and comprehend the speech. For speech signals, syllables have a window with an upper limit of about 240 ms and short words of about 300 ms (Navarra *et al*., 2005). Although it is important to note that this window of integration can be highly stimulus-dependent, and ranges in the literature between 150 and 800 ms (Colonius & Diederich, 2010; Schwartz & Savariaux, 2014, Ren *et al.,* 2017), and even differs between age groups (Ren *et al.,* 2017). This mixed range in the literature could also be due to a mismatch with reported display refresh rates (typically 60 Hz), video framerates (typically 30 – 60 frames per second) and levels of SOA used in research if reported at all (Ren *et al.*, 2017). For example, in 10 ms increments, a 60 Hz monitor can't display separate visual streams of information that refresh every 10 ms, as it is only capable of doing so every 16.6 ms, assuming the video plays at a full 60 frames per second as well.

The present study aimed to reassess the benefits of visual information to speech-in-noise perception using stimuli with visual distinctiveness. We also aimed to determine the effect of SOA on audiovisual speech perception. We tested the following hypotheses:

(i) purely audio speech discrimination accuracy will be decreased when speech is presented in noise compared to without noise.

(ii) reaction time to correctly discriminated purely audio speech will be increased when speech is presented in noise compared to without noise.

(iii) speech-in-noise discrimination accuracy will be increased when speech is presented with congruent visual information of the speaker's lip movements (audiovisual stimuli) compared to when no visual information is present (purely audio stimuli).

(iv) reaction time to correctly discriminated speech-in-noise will be decreased when speech is presented with congruent visual information of the speaker's lip movements (audiovisual stimuli) compared to when no visual information is present (purely audio stimuli).

(v) as the visual information precedes the auditory information by larger SOAs (0 - 266 ms), speech-in-noise discrimination accuracy will decrease.

(vi) as the visual information precedes the auditory information by larger SOAs, reaction time to correctly discriminated speech-in-noise will increase.

Further exploratory analysis also investigated the window of integration for these audiovisual stimuli, as well as differences in visual benefit between each syllable used.

Method

Design

To address hypotheses (i) and (ii), a single within factor (noise type: speech without noise, and speech-in-noise) design was used for purely audio trials with no stimulus asynchrony. For hypotheses (iii) and (iv), a single within factor (stimulus type: audiovisual, and purely audio) design was used for speech-in-noise trials with no stimulus asynchrony. Finally, for hypotheses (v) and (vi), a single within factor (SOA; 0, 200, 216.6, 233.3, 250, and 266.6 ms) design was used for audiovisual, speech-in-noise trials only. In total, participants took part in all 14 unique conditions (see Table 1), and both the accuracy of speech discrimination and reaction time in the discrimination task were recorded.

Ethical approval was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (approval reference: FST-2022-2122-RECR-2, project ID: 2122). The study was pre-registered on AsPredicted.org before commencing data collection. The pre-registration can be found at https://aspredicted.org/aq98a.pdf. All deviations from this pre-registration are listed in the section below. The collected data have been archived on the Open Science Framework (OSF: https://osf.io/kcbzs).

Deviations from pre-registration

In the original study pre-registration, there were three set hypotheses listed:

- There will be a decrease in the accuracy of speech discrimination (measured by correct responses in trials) or an increase in response times in the auditory-only condition when the speech is in noise compared to speech without noise.
- When visual information is present (audiovisual), the accuracy of speech discrimination and response times for each trial will not be as obstructed in

speech-in-noise conditions compared to audio-only conditions (when no visual information is present).

- As the visual information precedes the auditory information by larger margins (200 ms, 216 ms, 233 ms, 250 ms, 266 ms), the accuracy of speech discrimination in the speech-in-noise conditions will decrease - or response times will increase - in audiovisual conditions compared to when the audiovisual information is congruent (0ms).

These were changed to the hypotheses listed in the introduction by splitting the dependent measures into separate hypotheses and improving readability. This was done to make interpretations of results more clearly defined when referring to the hypotheses. To accompany this, the models used to test the hypothesis were also adjusted, giving six separate models of analysis - one for each hypothesis - instead of four. Generalised linear mixed-effects models (GLMER) were used to test all six hypotheses, instead of the mixture of GLMER models for accuracy data and LMER models for reaction time data that was listed in the pre-registration. This was done as GLMER models are more appropriate than LMER models for reaction time data, which is generally positively skewed (Lo & Andrews, 2015). These GLMER models were preferable still over repeated measures generalised linear models for considering random effects that may be present on a participant-by-participant basis. Finally, in our sample size calculation using data simulation (see section 'Sample size calculation'), it was determined that 60 participants were needed to sufficiently power the study. In the pre-registration, we then added a further 10% after a priori calculations (another 6 participants) to make a sample size estimate of 66. Due to the availability of resources, this extra 10% was not collected, leaving the sample of the study at the original number of 60 participants.

Participants

All data were collected online, with 81 participants recruited for the study. Of these, a total of 60 participants completed the study (mean age = 25.66, 28 male, 30 female, two non-binary). The other 21 participants completed the eligibility questionnaire but were either not eligible or did not proceed to the study task and provide study data. Participants were recruited via online advertisements or through Prolific and were compensated for their time. All participants were monolingual, native speakers of British English to control any potential speech perception differences across languages and in bilingualism and multilingualism (Lotfi *et al.*, 2019). Participants reported no hearing disorders and had either normal or corrected-to-normal vision. Only those between the ages of 18 and 35 were tested, as the window of integration for audiovisual information increases significantly with age, which can make speech discrimination more difficult (Ganesh et al., 2018; Sekiyama *et al.*, 2014). Participants reported no developmental disorders, such as dyslexia, or history of developmental disorders. This was important as the window of integration for audiovisual stimuli is wider in individuals with learning difficulties such as autism spectrum disorder and developmental dyslexia (Smith & Bennetto, 2007; Megnin-Viggars & Goswami, 2013; Michalek et al., 2014; Noel et al., 2018). All participants were right-handed. Finally, participants had no musical expertise, as previous research suggests that individuals with continuous experience as musicians can detect smaller SOAs, even for speech syllables (Lee & Noppeney, 2014; Sorati & Behne, 2019). Musical expertise was defined as training with a single musical instrument or voice for more than 7 years (Varnet *et al.*, 2015; Lee *et al.*, 2020) and for at least 3.5 hours a week (Lee & Noppeney, 2014). Participants were screened for the experiment using Qualtrics (see section 'Procedure').

Sample size calculation

Before testing, data simulation was conducted using R studio for power and sample size analysis. Lme4 (vers. 1.1-27.1; Bates *et al.*, 2015), afex (vers. 1.0-1; Singmann *et al.*,

2024) and simr (vers. 1.0.5; Peter *et al.*, 2019) were the core packages utilised in this process. Firstly, means and standard deviations of accuracy were gathered from studies that used syllable or bi-syllable phonetic speech tokens to investigate visual integration in speech perception. These studies typically used either multiple signal-to-noise ratios (SNRs; between -12 and -18 dB: Altieri *et al.*, 2014; Grant & Seitz, 1998; Sekiyama *et al.*, 2014) or individualised ratios (Ten Oever *et al.* 2013). For those studies that used multiple speech-to-noise ratios, we took data from – or closest to – -16 dB SNR. -16 dB was selected for our speech-shaped noise as this was the average SNR at which there was a notable difference between perceiving speech with or without visual aid (Bernstein *et al.*, 2004). An average estimated mean and standard deviation were then calculated for each condition. A dataset was produced using the rtruncnorm function (truncnorm package; vers. 1.0-8; Mersmann *et al*., 2018) - to randomly generate data for each condition that had a mean and standard deviation close to the ones calculated. This was repeated for each speech token ('Ba', 'Fa', and 'Ka') and all trials of each condition, providing a full dataset of expected results.

The dataset was then analysed using our planned experimental analyses (see below) to generate predicted results. Simulations were repeated 1000 times. An aggregation of power was then calculated. If the power was insufficient (below .80 at an alpha level of .05), the sample size of the dataset was manually adjusted, and the data simulation was conducted again. This was done until a minimal sample size with sufficient power was found. A total of 60 participants were calculated to be needed for sufficient power. The code for data simulation is available on OSF (https://osf.io/kcbzs).

Materials

The experiment was created using PsychoPy 3's builder tools (vers. 2021.2.3; Peirce *et al.*, 2019) and hosted online through Pavlovia. A consent form and a screening form were created and hosted on Qualtrics (Qualtrics, 2005). Three single-syllable speech tokens were used: ‘Ba’, ‘Fa’, and ‘Ka’. These were chosen as they belong to three distinct viseme categories, did not rely on any tongue movements to distinguish that would have been obscured from sight (such as labiodental phonemes), and could be easily distinguished without visual aid when not in noise. These speech tokens were spoken by a native British English-speaking male speaker and were recorded using personal home equipment. An external USB 3.0 condenser microphone was used to record audio (HyperX Quadcast with default windshield, set to the cardioid position). The initial video footage was recorded at 1920 x 1080 resolution and 60 frames per second using a mobile device (OnePlus 7 Pro). Both devices were connected to a single desktop machine, which recorded the audio and video in tandem using open-source OBS Studio software (Open Broadcaster Software, version 29.1.3). After the initial recording, the speech tokens were edited in length and converted to mp4 files at a resolution of 1280 x 720 and a frame rate of 60 frames per second. As the study would be completed on participants’ laptops or desktop systems and using their internet connection, we could not ensure that all participants were using a device with a 1920 x 1080 resolution screen. By reducing the resolution of files to 1280 x 720, all likely participant resolution sizes could be accommodated whilst ensuring that all participants viewed the files at the same resolution. Sixty frames per second was chosen as the frame rate as home device monitors and laptop screens are typically to a standard 60 Hz or higher. By using the lower boundary and not a higher frame rate, we can be sure that all SOAs implemented in the stimuli were visually relayed to the participant. For audiovisual conditions, the video footage contained only the speaker’s lower face in view, containing mouth and lips. This meant that participants were only provided with visual information regarding the lip movements made when speaking, and not any other visual information relevant to other actions the speaker may have made during recordings. For audio-only conditions, the video of the lips was overlayed with a plain black PNG image file. This kept the audio-only stimuli in a consistent video format rather than exporting the file as an mp3. All video files were the same length of 2 s.

Audacity software (Audacity Team, 2021) was then used to rip the audio from the MKV files to be edited as WAV files in Praat software (Boersma & Weenink, 2021) for the creation of speech-shaped noise. First, a sentence using English words – '*His plan meant taking a big risk'* - was recorded to provide a base for the speech-shaped noise. White noise was then produced using Praat's white noise generator. The noise was brought down to an intensity tier, then an amplitude tier. This was then multiplied with the sentence above to create speech-shaped noise (Van Engen *et al.*, 2017). Praat was then used to combine the speech-shaped noise with the speech-in-noise conditions at a speech-to-noise ratio of –16 dB. This was done using a Praat script developed by McCloy (2021). Finally, Audacity was used again to ramp up the start and ramp down the ends of all audio files for every condition. The audio was then stitched back onto the MP4 files.

For the conditions where the stimuli were asynchronous, Lightworks was again used to desynchronise the onset of the audio ahead of the onset of the lip movements using exact frames of the video footage (12, 13, 14, 15, and 16 frames per second) which corresponded with the SOAs of the relevant conditions (audio starting after the visual lip information by 200, 216.6, 233.3, 250, and 266.6 ms). The result was 42 stimuli in MP4 format, representing three speech tokens ('Ba', 'Fa', and 'Ka') for each of the 14 condition levels presented to the participant.

Procedure

Participants were linked to Qualtrics once they had consented to the study. Participants were also reminded at this stage to ensure that they were in a quiet room with no background noise, as well as to load the experiment on either Microsoft Edge, Google Chrome, or Mozilla Firefox internet browsers on a laptop or desktop computer. They were explicitly told not to open the experiment on any other browser, such as Safari, nor a mobile or tablet device as these were incompatible. Participants were also instructed to use headphones for the experiment, rather than to play the stimuli through their device's speakers.

A volume check began, in which a constant pure tone played (440 Hz frequency), and participants were asked to adjust the volume of their device as necessary for a comfortable auditory experience and to ensure that the audio was playing correctly at a sufficient volume level. This tone would play for as long as the participant wished to alter the volume levels of their device. Once complete, the spacebar would be pressed, and the tone stopped. Participants were informed that a video would play either showing no visual information or visual information of lips moving. Meanwhile, speech would be played. Participants were told to listen carefully to the speech sound spoken, and after hearing the sound to press one of three buttons on their keyboards that corresponded with the three available speech tokens. They were instructed to respond to each trial as quickly as possible. They were reminded before and after each trial to press 'z' on their keyboard if they heard 'Ba', 'x' for 'Fa', or 'c' for 'Ka'. Participants were told to answer as quickly as possible. If they were unsure, they were told to make a guess.

Participants were given six practice trials before data were collected. This was using the speech without noise, 0 ms, and audiovisual condition stimuli, with two trials for each of the three speech tokens (Ba, Fa, and Ka). A white crosshair would be displayed on the screen for 1000 ms before the trial began to bring attention to the centre of the screen where the video trials would be displayed. Stimuli were shown for 2500 ms, then the response screen would display. On this screen, the participants were reminded of the buttons to press for each of the three speech sounds. Only the three buttons could be pressed and pressing the buttons whilst the stimuli were still playing would not record a response or stop the trial. A total of 546 trials (not including the practice trials) were completed. The order of the trials and conditions was completely random to avoid any potential order bias. After every 42 trials, a break screen would appear. This screen told the participant to take a short break before continuing with a press of the spacebar. If the participant did not wish to take a break, they were permitted to continue with a spacebar press immediately. There was a total of 12 breaks in the experiment, each with a short attention check question to ensure participants remained attentive to the experiment. Upon completing the study, participants could close the browser tab or window down and all data would remain recorded on the Pavlovia system.

Analysis

Descriptive statistics were first gathered from each condition for both the accuracy ratings and the reaction times. Reaction times were taken from the offset of the stimuli to the participant response. The average accuracy and reaction time of accurately responded trials for each condition and each participant was calculated, with reaction times winsorised over the 95th percentile only. This was done to replace any large, outlying reaction times to trials that may be due to a distraction at home during testing or the participant taking a short break before the break period. The assumptions of linear and generalised linear mixed-effects models were tested, including residual plots to check for linearity, quantile-quantile plots for normality, assessing the levels of multicollinearity between stimulus type, noise, and SOA using variance inflation factors, and ensuring the assumption of homoscedasticity was met. All the above tests were conducted on the dataset and all assumptions were met. As we were testing six separate hypotheses, the experiment-wise error rate was controlled using the Bonferroni-Holm method (Holm, 1979).

With further regards to stimulus variability, previous studies often employ analyses such as repeated measures analysis of variance (ANOVA) tests which do not consider random effects (Bates *et al.*, 2015). Including random effects is important for ensuring that any effects found in the model are not influenced by differences in participant ability or by the stimuli themselves, as some stimuli may be easier to recognise and comprehend in noise than others. To counter this issue, mixed-effects models can be used that consider the random effects, such as participant number and stimuli number, across intercepts and slopes within the model to provide a more valid interpretation of the integration between visual and auditory systems in speech perception.

Using the lme4 package (Bates *et al*., 2015), generalized linear mixed-effects regression model (GLMER) analyses were conducted for the accuracy scores to test hypotheses (i), (iii), and (v) and for reaction time scores to test hypotheses (ii), (iv), and (vi). GLMERs were chosen instead of repeated measures generalised linear models such as ANOVA tests because they consider random effects that may be present across all 546 trials on a participant-by-participant basis. GLMER was chosen over LMER for analysis with reaction times as these scores are typically positively skewed. As noted by Lo and Andrews (2015), generalised linear mixed models are more appropriate for skewed datasets in this context. Furthermore, accuracy in a trial is a binary outcome variable that can either be correct (1) or incorrect (0). Therefore, GLMERs were used to ensure that assumptions of categorical dependent variables in mixed-effects models were met. GLMERs were conducted using the lme4 package still, as this package supported a generalised approach. Due to the generalised nature of the model and package restrictions, no suitable p-values were provided with the GLMER analyses. Instead, significance was interpreted using 99.2% confidence intervals (CIs), chosen to reflect our lowest criterion of significance in the Bonferroni-Holm approach being $p < .008$ for six comparisons. If the resulting confidence intervals showed insignificance, the next boundary of Bonferroni-Holm ($p < .01$) was checked using 99% confidence intervals. This kept going until either significance was found or no significance was found at a significance level of $p < .05$. Once detected or classed as insignificant, the test was ranked with the other p-values in our analyses as the lowest boundary of significance and Bonferroni-Holm was conducted as normal on our six ranked comparisons.

To test hypothesis (i), a GLMER analysis was conducted using the accuracy of responses on the speech discrimination task as the dependent variable and using noise type (no noise or speech-shaped noise) as the independent variable in the model. As we hypothesised that presenting speech in noise would significantly decrease accuracy compared to without noise, we expected to find a significant effect of noise type from this GLMER analysis. Hypothesis (ii) was the same as the first but looked at reaction times to correctly discriminated speech-in-noise on the same task. A GLMER was used to test this hypothesis, using reaction times as the dependent variable and noise type as the independent variable. Similarly, we expected to find a significant effect of noise type, increasing reaction times.

To test hypothesis (iii), a GLMER analysis was conducted using the accuracy of responses on the speech discrimination task as the dependent variable and using stimulus type (purely audio or audiovisual) as the independent variable in the model. As we hypothesised that presenting audiovisual stimuli in noise would significantly increase accuracy compared to purely audio stimuli in noise, we expected to find a significant effect of stimulus type from this GLMER analysis. Hypothesis (iv) was the same as the third but looked at reaction times to correctly discriminated speech-in-noise on the same task. A GLMER was used to test this hypothesis, using reaction times as the dependent variable and stimulus type as the independent variable. Similarly, we expected to find a significant effect of stimulus type, decreasing reaction times.

To test hypothesis (v), we conducted a GLMER analysis using accuracy as a dependent variable and SOA as the independent variable. SOA was treated as a categorical variable in this model and the model for hypothesis (vi) below. We expected to find a significant effect of SOA, with accuracy decreasing when more asynchrony was introduced to the stimuli. This would reflect that the window of integration for audiovisual speech is important for visual information to be beneficial to understanding speech in noise. Finally, in a similar manner, hypothesis (vi) was tested using a GLMER analysis with reaction times as the dependent variable and with SOA levels as the independent variable in the model. Again, we expected a significant effect of SOA on reaction times, with reaction times increasing with the introduction of asynchrony.

For all six GLMER models listed above, the speech sound token used (Ba, Fa, or Ka), participant age and the participant ID were all included as random effects. No further model selection of these random and fixed effects was undergone, as we wanted a conservative model that included a full random effects structure to account for the expected larger individual differences of an online experiment. All model equations and structures can be found in the supplementary materials (Table 2).

Furthermore, we also conducted exploratory analyses to assess the effect of noise on speech discrimination accuracy between the three visually distinct, chosen phonemes (‘Ba’, ‘Fa’, and ‘Ka’). To do this, a GLMER analysis was conducted using accuracy as the dependent variable and speech token as the independent variable. Purely audio trials in noise were used for this analysis. Furthermore, we also conducted pairwise comparisons within the GLMER models used to test hypotheses (v) and (vi) as another exploratory analysis, comparing between each level of our SOA independent variable. We expect that not all the SOA interactions will show significance. As we expected the benefits of visual stimuli to only be present during the window of integration, there would only be a significant decrease in accuracy and an increase in reaction times at SOAs outside this window. Therefore, this exploratory analysis can be used to better understand the window of integration for our stimuli. All exploratory analyses will use an inference criterion of $p < .008$ as this was the strictest threshold for significance included in our Bonferroni-Holm correction.

Results

Descriptive statistics

The means and standard deviations of the accuracy of responses and reaction times of responses can be seen in Table 1. Descriptive statistics were also calculated for each speech token (Ba, Fa, and Ka). Figure 1 shows the mean reaction times and mean accuracy rates for both audio-only and audiovisual stimuli when no SOA is considered (0 ms SOA), whilst Figure 2 shows these data for all SOAs when audiovisual stimuli are used for speech-in-noise conditions. Figure 3 shows the mean reaction times and accuracy rates for all SOAs when audiovisual stimuli are presented without noise. Furthermore, Figure 4 shows accuracy rates and reaction times in purely audio and audiovisual stimuli in noise between each of the three speech tokens. Violin plots were used for all figures to highlight the distribution of accuracies and reaction times across participants for each condition, as individual differences were large in this dataset likely due to online experimentation.

Effect of noise on speech perception

The first planned GLMER analysis was conducted to test hypothesis (i). There was a significant effect of noise type (with or without noise), showing a decrease in accuracy in speech-in-noise discrimination when noise was introduced versus clear speech ($\beta = -.29$, $t = -12.95$, 99.2% $CI = [-.35, -.23]$, $p < .008$). This model supports hypothesis (i), as we expected to find that the introduction of noise to speech would decrease performance. For testing hypothesis (ii), the planned GLMER analysis was conducted. There was a significant effect of noise type on reaction times ($\beta = .06$, $t = 3.10$, 99.2% CI = [.01, .11], $p < .008$). This model supports hypothesis (ii), as we expected to find that introducing noise would increase reaction times to correctly discriminated speech.

Effect of congruent, distinguishable visual information on speech perception

Our next planned GLMER analysis was conducted to test hypothesis (iii). There was a significant effect of stimulus type (purely audio or audiovisual), as there was an increase in accuracy in speech-in-noise discrimination when stimulus type was audiovisual versus purely audio ($\beta = .26$, $t = 11.36$, 99.2% *CI* = [.20, .32], $p < .008$). This model supports hypothesis (iii), as we expected to find that introducing relevant visual information would improve speech perception in noise. For testing hypothesis (iv), the planned GLMER analysis was conducted. There was a significant effect of stimulus type on reaction times ($\beta = -.08$, $t = -4.15$, 99.2% CI = [-.13, -.03], $p < .008$). This model supports hypothesis (iv), as we expected to find that introducing relevant visual information would decrease reaction times and improve speech perception in noise.

Effect of stimulus onset asynchrony on audiovisual speech perception

When testing hypothesis (v), the planned GLMER analysis was done for data across all SOA levels for audiovisual speech-in-noise stimuli only. There was no significant effect of SOA on accuracy at any interval, even at a 95% confidence interval, showing no support for hypothesis (v). Finally, our planned GLMER analysis was run to test hypothesis (vi). There was a significant main effect of SOA ($\beta = .04$, $t = 3.31$, $p < .008$) on reaction times, indicating reaction times increased with SOA. This supports hypothesis (vi).

Exploratory analyses

As a further, exploratory analysis, a GLMER model was used to investigate phoneme differences in speech-in-noise discrimination. Looking at pairwise comparisons, there was a significant difference between accuracy rates of the ‘Ba’ and ‘Fa’ tokens ($\beta = -.17$, $t = -5.03$, $p < .008$), ‘Ba’ and ‘Ka’ tokens ($\beta = -.53$, $t = -15.50$, $p < .001$), and ‘Fa’ and ‘Ka’ tokens ($\beta = -.36$, $t = -10.47$, $p < .008$) for purely audio stimuli. For audiovisual stimuli, however, there was no significant change in accuracy rate between the three tokens. A GLMER model for reaction times showed similar patterns, although only ‘Ba’ and ‘Ka’ were significantly different for purely audio stimuli, with ‘Ba’ having increased reaction times in comparison to ‘Ka’ ($\beta = .14$, $t = 4.21$, $p < .008$).

Finally, to explore differences between SOA intervals to see if a window of integration could be determined, pairwise comparisons were made on the GLMER analyses used to test hypothesis (vi). Pairwise comparisons were not made on the GLMER used to test hypothesis (v) as no significant effect of SOA on accuracy was observed. Pairwise comparisons made on the GLMER to test hypothesis (vi) indicated that reaction times were significantly reduced compared to 0 ms at 250 ($\beta = -.05$, $t = -3.94$, $p = .001$) and 266.6 ms ($\beta = -.05$, $t = -3.88$, $p = .002$). However, no other comparisons between levels of SOA were significantly different. Whilst this implies that a minimal end of the window of integration could lie above 233.3 ms (as SOAs between 233.3 and 250 ms were not tested), no accurate window of integration can be determined from the data.

Discussion

This study aimed to reassess the contribution of audiovisual integration to speech perception in noise when stimuli belonged to different viseme categories. As speech perception can differ wildly with stimuli sets, it was important to first reassess the detriment of noise on speech discrimination, as well as the benefits of speech-relevant visual integration. The study incorporated the visual distinguishability of each speech phoneme used in the speech discrimination task by selecting phonemes from separate viseme categories. Furthermore, the study also aimed to examine the effects of stimulus onset asynchrony (SOA) on audiovisual speech perception. This may assist in determining a window of integration for these stimuli, which was explored in further analyses.

Reassessing the detriment of noise on speech perception

GLMERs were used to investigate the influence of the predictor variables on accuracy ratings on the speech discrimination task. The first model, using noise type as the predictor, supported our first hypothesis, showing that there was a decrease in accuracy for purely audio stimuli when the speech was presented in noise and not without. Additionally, the introduction of noise to the speech signal increased reaction times significantly. These results support our second hypothesis. As both the accuracy and reaction time to trials with noise differed significantly from those without, it can be said that the detriment of noise on speech perception was present with our created stimuli and chosen SNR ratio of -16 dB using speech-shaped white noise.

Reassessing the contribution of audiovisual information on speech processing in noise

There was a significant increase in accuracy when relevant, congruent visual information was present with the stimuli versus purely audio stimuli in noise. This supports hypothesis (iii) and confirms previous findings regarding the contribution of audiovisual information to speech-in-noise processing. However, it should be noted that whilst the effect is prominent, it is not as great as previous literature findings which used a similar speech-to-noise ratio (Van de Rijt *et al.*, 2019). Here, the effectiveness of audiovisual enhancement of speech recognition was assessed with SNR ratios as low as -21 dB SNR, where the introduction of relevant visual cues provided an increase in accuracy of up to 50% for some stimuli, with greater enhancements for words like 'Pieter'. Even at -16 dB SNR, Van de Rijt et al.'s data suggests that greater audiovisual enhancement should have been seen, though reaction time data was not reported in the study.

This could also be explained using results from our exploratory analysis. When the speech was in noise and the stimuli contained auditory information only, the token 'Ba' displayed much lower mean accuracy scores than the other tokens. This suggests that there are specific differences in the acoustic properties of the tokens used that are influencing the perception of speech-in-noise. In previous literature, 'Ba' and other tokens within the same viseme (such as 'Pa') are frequently used, which could suggest why results in previous literature show a larger speech discrimination effect in noise. It is therefore important for future research to determine if there are differences in speech perception between other viseme categories that were not used in this study (Fisher, 1968). In our LMER model for hypothesis (iii), the token used was loaded as a random factor. This variance between tokens was removed from the variance found in fixed effects in the outputs of the model. This mixed effect modelling also considered participant differences and age, unlike previous literature that did not investigate speech discrimination effects using more complex models (Bernstein *et al.*, 2004; Sekiyama *et al.*, 2014). As the tokens appear to be largely variant, this could further account for the weaker overall patterns of change seen between fixed effects.

Next, there was a significant decrease in reaction times when audiovisual stimuli were used over purely audio, supporting hypothesis (iv). Interestingly, there was a decrease in reaction time in audiovisual conditions with noise over without noise as well. When processing multisensory stimuli that are not beneficial to us, reaction times likely increase due to extra unnecessary processing (Brown & Strand, 2019). In this case, the audiovisual information is only beneficial to us in noise. Therefore, in this model where no comparisons to clear speech are made, reaction times significantly decrease with the introduction of noise as the extra processing of visual information becomes beneficial. Comparatively, when audiovisual information is present without noise, reaction times increase as the added visual information is no longer beneficial to speech recognition as it is already clear to understand.

Investigating the effects of stimulus onset asynchrony on the speech processing benefits of audiovisual information

Our GLMER model testing hypothesis (v) uncovered no meaningful change in accuracy between any SOA value. In previous research, the maximal window of integration was around 250 to 260 ms for syllables (Dixon & Spitz, 1980). Here, SOAs up to 266.6 ms did not affect speech discrimination accuracy, implying that the stimuli were still inside the window of integration and that the maximal end of the window lies beyond 266.6 ms. Our final LMER model testing hypothesis (vi) found significant increases in reaction time when an SOA was introduced. Alternatively, this implies that the range of SOAs used does cover the maximal end of the window concerning processing speed, as there was a gradual increase in reaction times as SOA was further increased reducing the benefit of audiovisual information. When looking at exploratory pairwise comparisons between SOA levels, there was a distinct decrease in reaction times at 250 and 266.6 ms compared to no SOA. This implies that the ability to discriminate the speech was made less taxing past 250 ms asynchrony. It could be, based on these findings, that the minimal end of the window of integration for our created stimuli lies between 233.3 and 250 ms. Given that the stimuli were simple syllables, an alternative interpretation may be that the processing of the auditory and the visual information was completed before integration had finished, although this would not explain the differences in reaction times between the SOA levels. Furthermore, as participants could only respond after the stimuli had played in full with visual cues preceding the auditory cues, we would expect integration to have occurred as long as the SOA remained within the window of integration. As these comparisons are exploratory, however, and there is no account of accuracy changing with SOAs, further research would be needed to determine the full window of integration.

Limitations of the study and future directions

One explanation for the audiovisual benefit in our data not being as large as in previous studies could be the lack of ecological validity and the artificial nature of the online experimentation. Speech-shaped white noise was utilised for speech-in-noise conditions. Despite this noise modulating speech, it is still unlike that in a real environment. This may mean that the speech-shaped noise was too distinct from the speech itself, especially considering that we used syllables for recognition rather than words or sentences. Speaker babble or background noise such as light vocal music would be much more akin to that in everyday life, making it perhaps more suitable and valid for investigating audiovisual speech perception when speech is in noise (Krishnamurthy & Hansen, 2009). Furthermore, the stimuli used were single syllable speech tokens, which do not reflect typical communicative speech in a real-world environment. Given their simplicity, other aspects of speech perception, such as prediction of oncoming words in larger sentences, would not be used as a method of speech processing here (Solberg Økland *et al.*, 2019). The overall simplicity and artificial design of these stimuli may be obscuring other benefits of audiovisual integration in speech perception when applied to realistic speech settings. To better reassess audiovisual integration in speech, further research with more ecologically valid speech stimuli (e.g., full sentences) would be of benefit.

The SNR used for our study was -16 dB. This was selected based on previous research investigating audiovisual syllable perception in noise, for which there was a notable difference between perceiving speech with or without visual aid (Bernstein *et al.*, 2004). However, whilst this may have been true for speech token 'Ba', this did not seem to translate to 'Ka', indicating that different speech viseme categories were affected by speech-shaped noise at the SNR -16 dB. Furthermore, initial data collection for this study was conducted from 2021 to 2022 after multiple lockdowns in the UK due to the COVID-19 pandemic. Many adults in the UK during this time had been socially distancing and wearing facemasks to prevent contamination. These facemasks would obscure the lip and mouth area of the wearer, meaning that social interactions between many people in this period would have lacked visual information to assist with speech perception. In many cases, the facemasks obscured sound, making it more difficult to understand speech and imitating difficult listening conditions (Yi *et al.*, 2021; Smiljanic *et al.*, 2021). It is possible that due to facemask wearing for a year, participants had adapted to listening to speech in difficult conditions without visual aids. Furthermore, only three phonemes from three viseme categories were used in this study. As there was an apparent difference between these phonemes selected, with 'Ba' being more impacted by added noise than 'Ka', future studies may wish to investigate the differences between more viseme categories and the phonemes within them. It may also be beneficial to further apply this to more than single-syllable units of speech. This would provide a broader view of the contributions of visual information to speech processing.

Finally, this experiment did use home equipment to record stimuli as well as the home equipment of participants to play the stimuli through online experimentation. Whilst the recording equipment was of laboratory standard and the recording procedure rigorous, there will still be discrepancies between these stimuli and other lab-created stimuli which might make replications difficult. Furthermore, the environments that participants were in whilst taking part in the study may be different between participants. We do not have measures of how well participants understood the task, how noisy their environment was during listening, the hardware they used to run the study, and if they followed pre-experiment instructions such as to wear headphones. These are likely to contribute to the large individual differences seen in the dataset. Whilst GLMER models can consider the participant differences, further in-person lab testing with similar methodologies may be needed to fully control these confounds.

Conclusion

A set of purely audio and viseme-controlled audiovisual stimuli was created to investigate the contributions of audiovisual information to speech-in-noise processing. Introducing visual information increased accuracy and decreased reaction times in speech-in-noise conditions relative to audio-only stimuli. When looking at accuracy and reaction times at varying SOA intervals in our audiovisual stimuli, introducing SOAs influenced reaction times, but not accuracy. In the future, more syllables from more viseme categories could be tested to investigate a full range of speech sounds in audio-only and audio-visual contexts, as well as with further SOA intervals to ensure that a window of integration can be determined with accuracy.

Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC) Training Grant (O'Hanlon, ES/P000665/1), the Manchester Biomedical Research Centre and the National Institute for Health and Care Research (NIHR) (Plack, NIHR203308), and the Biotechnology and Biological Sciences Research Council (BBSRC) New Investigator Grant (Nuttall, BB/S008527/1). We would like to thank all participants who expressed interest and participated in this research. We also thank Kyle Stonehouse for their contributions to study material creation.

Data Availability Statement

Upon publication, all collected data are available to view online through the Open Science Framework (OSF: https://osf.io/kcbzs), as well as all stimuli used in the experiment code relevant to data analysis.

References

Altieri, N., Townsend, J. T., & Wenger, M. J. (2014). A measure for assessing the effects of audiovisual speech integration. *Behavior Research Methods, 46*(2), 406-415.

Audacity Team. (2021). *Audacity(R): Free Audio Editor and Recorder* [Computer application]. Version 3.0.0 retrieved March 17th, 2021, from https://audacityteam.org/ . *Copyright statement: Audacity® software is copyright © 1999-2021 Audacity Team. The name Audacity® is a registered trademark.*

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

Bernstein, L. E., Auer Jr, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1-4), 5-18.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.53, retrieved August 2021 from http://www.praat.org/

Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language, 87*, 71-83.

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition, 2*(1).

Brown, V. A., Van Engen, K. J., & Peelle, J. E. (2021). Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults. *Cognitive Research: Principles and Implications, 6*(1), 1-12.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology, 5*(7), e1000436.

Colonius, H., & Diederich, A. (2010). The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in integrative neuroscience, 4*, 1316.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719-721.

Ewen, C. J., & Van der Hulst, H. (2001). *The phonological structure of words: an introduction*. Cambridge University Press.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of speech and hearing research, 11*(4), 796-804.

Ganesh, A. C., Berthommier, F., & Schwartz, J. L. (2018). Audiovisual binding for speech perception in noise and in aging. *Language Learning, 68*, 193-220.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Krishnamurthy, N., & Hansen, J. H. (2009). Babble noise: modeling, analysis, and applications. *IEEE transactions on audio, speech, and language processing, 17*(7), 1394-1407.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in linear mixed effects models* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/lmerTest/index.html

Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in psychology, 5*, 868.

Lee, J., Han, J., & Lee, H. (2020). Long-Term Musical Training Alters Auditory Cortical Activity to the Frequency Change. *Frontiers in Human Neuroscience, 14*, 329.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in psychology, 6*, 148545.

Lotfi, Y., Chupani, J., Javanbakht, M., & Bakhshi, E. (2019). Evaluation of speech perception in noise in Kurd-Persian bilinguals. *Auditory and Vestibular Research, 28*(1), 36-41.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior research methods, 49*(4), 1494-1502.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PloS one, 4*(3), e4638.

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. Journal of Experimental Psychology: *Human Perception and Performance, 37*(1), 245.

Massaro, D. W., Cohen, M. M., Tabain, M., & Beskow, J. (2012). Animated speech: Research progress and applications In Clark RB, Perrier J, P, & Vatikiotis-Bateson E (Eds.), Audiovisual Speech Processing (pp. 246–272). *Cambridge: Cambridge University*.

McCloy, D. (2021). *Praat Script: 'Mix speech with noise'* [Praat script]. LICENSED UNDER THE GNU GENERAL PUBLIC LICENSE v3.0: http://www.gnu.org/licenses/gpl.html

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

Megnin-Viggars, O., & Goswami, U. (2013). Audiovisual perception of noise vocoded speech in dyslexic and non-dyslexic adults: the role of low-frequency visual modulations. *Brain and language, 124*(2), 165-173.

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *Truncated normal distribution* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/truncnorm/

Michalek, A. M., Watson, S. M., Ash, I., Ringleb, S., & Raymer, A. (2014). Effects of noise and audiovisual cues on speech processing in adults with and without ADHD. *International journal of audiology, 53*(3), 145-152.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507.

Noel, J. P., Stevenson, R. A., & Wallace, M. T. (2018). Atypical audiovisual temporal function in autism and schizophrenia: similar phenotype, different cause. European *Journal of Neuroscience, 47*(10), 1230-1241.

Open Broadcaster Software. (2024). OBS Studio (Version 29.1.3) [Computer software]. https://obsproject.com/

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Peter, G., Catriona, M., & Phillip, A. (2019). *Power analysis for generalised linear mixed models by simulation* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/simr/index.html

Qualtrics. (2005). *Qualtrics software*, Provo, Utah, USA. Copyright@2021, Current version: 09-21. Retrieved from: https://www.qualtrics.com

Ren, Y., Yang, W., Nakahashi, K., Takahashi, S., & Wu, J. (2017). Audiovisual integration delayed by stimulus onset asynchrony between auditory and visual stimuli in older adults. *Perception, 46*(2), 205-218.

Satterthwaite, F. E. (1941). *Synthesis of variance. Psychometrika, 6*(5), 309–316.

Schwartz, J. L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10*(7), e1003743.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology, 5*, 323.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2024). *Analysis of factorial experiments* [computer manual]. Retrieved from: https://cran.r-project.org/web/packages/afex/index.html

Smiljanic, R., Keerstock, S., Meemann, K., & Ransom, S. M. (2021). Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *The Journal of the Acoustical Society of America, 149*(6), 4013-4023.

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

Solberg Økland, H., Todorović, A., Lüttke, C. S., McQueen, J. M., & De Lange, F. P. (2019). Combined predictive effects of sentential and visual constraints in early audiovisual speech processing. *Scientific Reports, 9*(1), 7870.

Sorati, M., & Behne, D. M. (2019). Musical Expertise Affects Audiovisual Speech Perception: Findings From Event-Related Potentials and Inter-trial Phase Coherence. *Frontiers in Psychology, 10*, 2562.

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

Sumby, W., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America, 26*, 212-215.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London. *Series B: Biological Sciences, 335*(1273), 71-78.

Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & Van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in psychology, 4*, 331.

Van de Rijt, L. P., Roye, A., Mylanus, E. A., Van Opstal, A. J., & Van Wanrooij, M. M. (2019). The principle of inverse effectiveness in audiovisual speech perception. *Frontiers in human neuroscience, 13*, 335.

Van Engen, K. J., Dey, A., Sommers, M. S., & Peelle, J. E. (2022). Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America, 152*(6), 3216-3225.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics, 79*, 396-403.

Varnet, L., Wang, T., Peter, C., Meunier, F., & Hoen, M. (2015). How musical expertise shapes speech perception: Evidence from auditory classification images. *Scientific Reports, 5*(1), 14489.

Yi, H., Pingsterhaus, A., & Song, W. (2021). Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic. *Frontiers in psychology,* 12.

Yuan, Y., Lleo, Y., Daniel, R., White, A., & Oh, Y. (2021). The impact of temporally coherent visual cues on speech perception in complex auditory environments. *Frontiers in neuroscience, 15*, 678029.

Supplementary Materials

The supplementary materials contain the following file:

A caption and display of Table 2, showing the model equations used for each of the six GLMERs.

Tables and Figures

			Accuracy Rate (%)		Reaction Time (ms)	
Speech	Stimuli	SOA (ms)	Mean	Std. Dev	Mean	Std. Dev
Clear	AO	0	96.11	10.24	538	216
Clear	AV	0	96.60	13.55	564	257
Clear	AV	200	95.95	14.09	551	241
Clear	AV	216.6	96.56	13.82	573	232
Clear	AV	233.3	96.58	12.61	575	249
Clear	AV	250	96.54	13.90	568	236
Clear	AV	266.6	96.84	13.23	575	255
Noise	AO	0	67.33	21.91	597	285
Noise	AV	0	93.10	15.21	518	239
Noise	AV	200	92.87	16.08	553	223
Noise	AV	216.6	93.62	15.75	554	218
Noise	AV	233.3	93.35	17.52	562	227
Noise	AV	250	93.11	16.30	570	224
Noise	AV	266.6	93.26	15.01	569	237

Table 1:

Means and Standard Deviations (Std. Dev) of accuracy rates and reaction times for speech with and without noise, audio-only (AO) or audiovisual (AV) stimuli, and different stimulus onset asynchronies (SOAs), with each speech token and participant aggregated into a single mean.

Figure 1: Violin plots showing the accuracy rates and reaction times of participants when speech was presented either with or without noise, for both audio-only (AO) and audiovisual (AV) stimuli. Boxplots show the median and interquartile ranges for each condition.

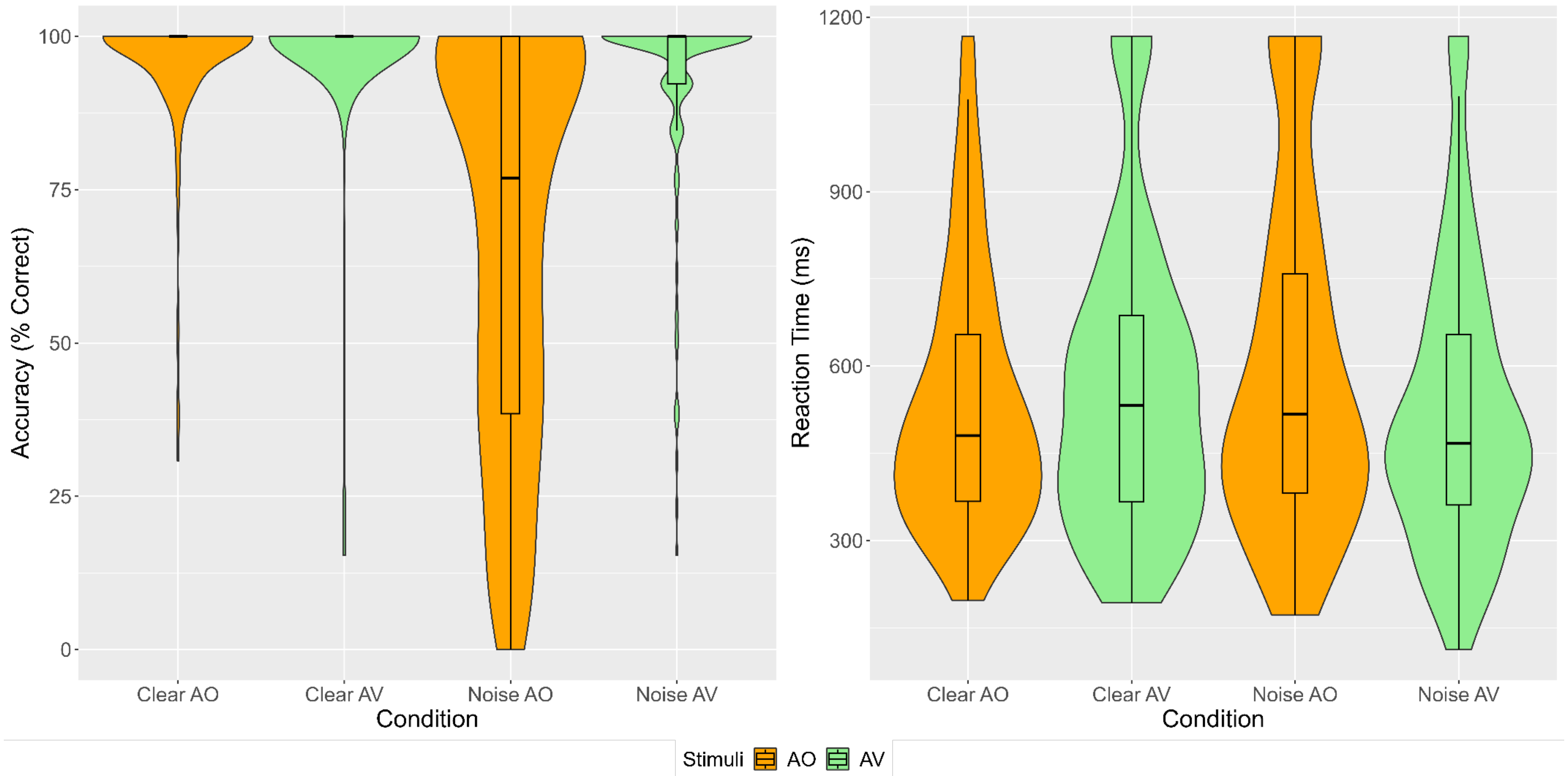

Figure 2. Violin plots showing the accuracy rates and reaction times of participants when audiovisual stimuli were presented in noise at different SOAs. Boxplots show the median and interquartile ranges for each condition.

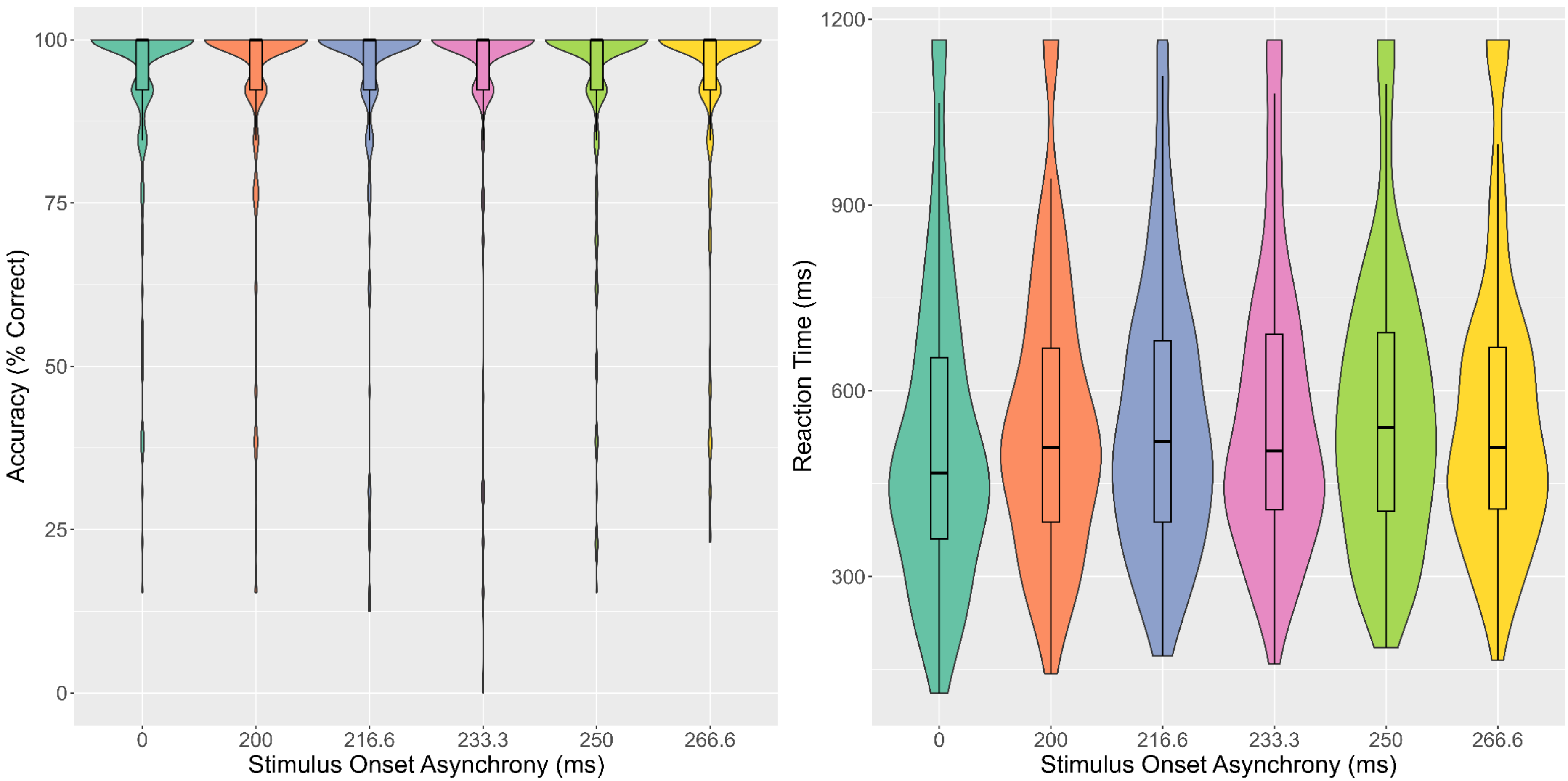

Figure 3. Violin plots showing the accuracy rates and reaction times of participants when audiovisual stimuli were presented without noise at different SOAs. Boxplots show the median and interquartile ranges for each condition.

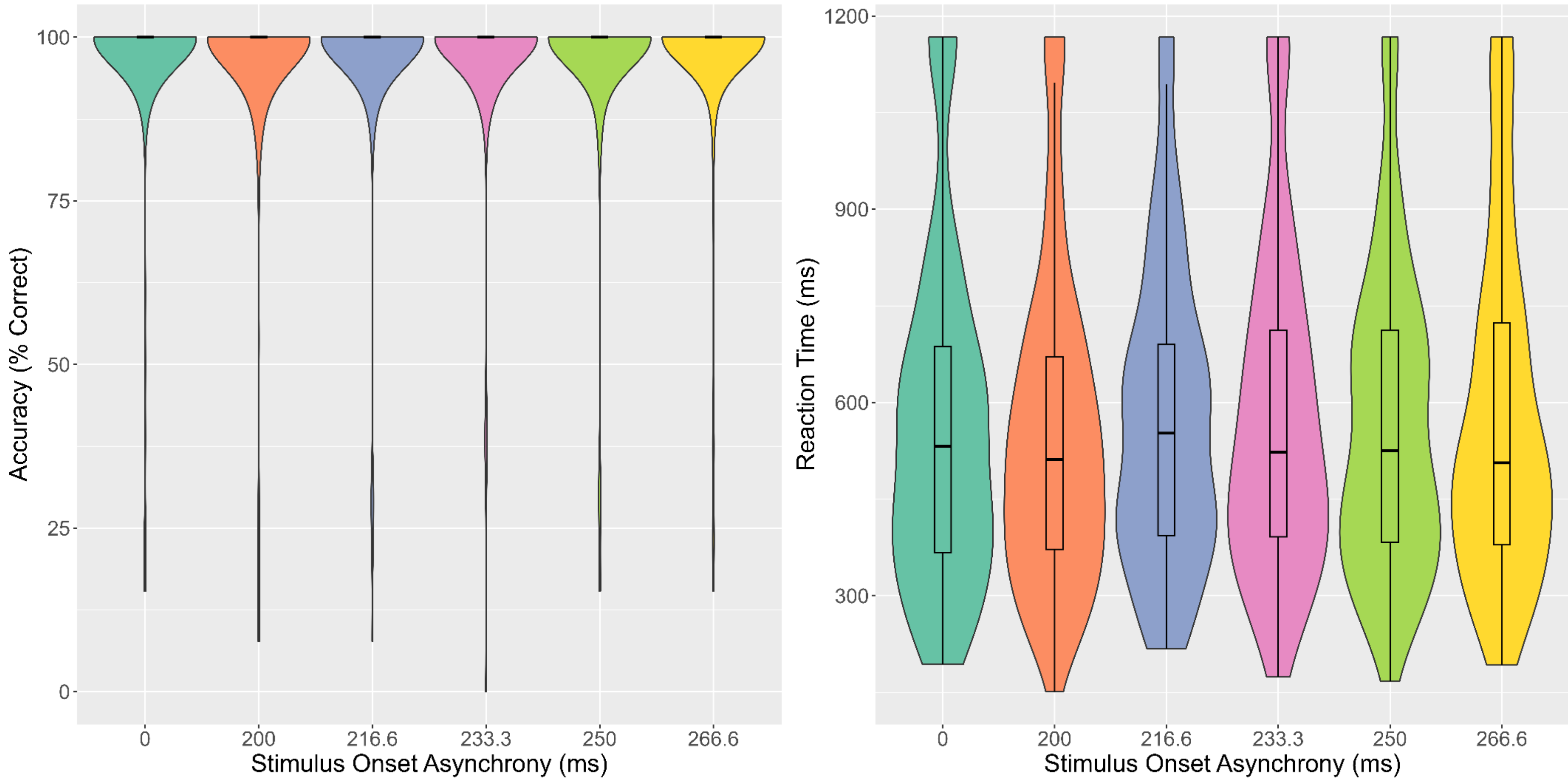

Figure 4. Violin plots showing the accuracy rates and reaction times of participants when speech tokens were investigated individually in noise for both Audio-Only (AO) and Audiovisual (AV) stimuli. Boxplots show the median and interquartile ranges for each condition.