NeuroImage 256 (2022) 119217

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Seeing a talking face matters: The relationship between cortical tracking of continuous auditory-visual speech and gaze behaviour in infants, children and adults ☆

S.H. Jessica Tan [a,*], Marina Kalashnikova [b,c], Giovanni M. Di Liberto [d], Michael J. Crosse [e], Denis Burnham [a]

[a] The MARCS Institute of Brain, Behaviour and Development, Western Sydney University, Australia
[b] The Basque Center on Cognition, Brain and Language, Australia
[c] IKERBASQUE, Basque Foundation for Science, Australia
[d] School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[e] Department of Mechanical, Trinity Center for Biomedical Engineering, Manufacturing AND Biomedical Engineering, Trinity College Dublin, Dublin, Ireland

ARTICLE INFO

Keywords:
Auditory-visual speech benefit
Cortical tracking
Gaze behaviour
Auditory-visual speech perception
Infants
Children
Adults

ABSTRACT

An auditory-visual speech benefit, the benefit that visual speech cues bring to auditory speech perception, is experienced from early on in infancy and continues to be experienced to an increasing degree with age. While there is both behavioural and neurophysiological evidence for children and adults, only behavioural evidence exists for infants – as no neurophysiological study has provided a comprehensive examination of the auditory-visual speech benefit in infants. It is also surprising that most studies on auditory-visual speech benefit do not concurrently report looking behaviour especially since the auditory-visual speech benefit rests on the assumption that listeners attend to a speaker's talking face and that there are meaningful individual differences in looking behaviour. To address these gaps, we simultaneously recorded electroencephalographic (EEG) and eye-tracking data of 5-month-olds, 4-year-olds and adults as they were presented with a speaker in auditory-only (AO), visual-only (VO), and auditory-visual (AV) modes. Cortical tracking analyses that involved forward encoding models of the speech envelope revealed that there was an auditory-visual speech benefit [i.e., AV > (*A* + *V*)], evident in 5-month-olds and adults but not 4-year-olds. Examination of cortical tracking accuracy in relation to looking behaviour, showed that infants' relative attention to the speaker's mouth (vs. eyes) was positively correlated with cortical tracking accuracy of VO speech, whereas adults' attention to the display overall was negatively correlated with cortical tracking accuracy of VO speech. This study provides the first neurophysiological evidence of auditory-visual speech benefit in infants and our results suggest ways in which current models of speech processing can be fine-tuned.

1. Introduction

When listening to a speaker talk face-to-face, we process visual speech cues as well as the predominant auditory signal. These visual speech cues come from facial movements that occur in tandem with acoustic speech and can provide additional information that augments speech perception both in quiet (e.g., Fort et al., 2013; Navarra and Soto-Faraco, 2007) and in noise (e.g., Moradi et al., 2013; Rudmann et al., 2003; Schwartz et al., 2004; Sumby and Pollack, 1954). The augmentation of speech perception by visual speech cues, or the *auditory-visual speech benefit*, has been widely studied. Most of these studies have been conducted with adults, but findings from studies with children and infants suggest that they too benefit from visual speech information, even though the degree of auditory-visual speech benefit increases with age. The studies reported here concern the auditory-visual speech benefit in 5-month-old infants, 4-year-old children and adults.

Behavioural studies provide evidence of an auditory-visual speech benefit across ages. For instance, 7.5-month-olds successfully segmented words from a fluent speech stream that was blended with a back-

☆ This research was funded by a doctoral scholarship to the first author funded by the MARCS Institute at Western Sydney University and the HEARing Cooperative Research Centre (CRC), and by HEARingCRC funding to the last author. The second author's work is supported by the Basque Government through the BERC 2018–2021 program, and PIBA PI-2019–0054, and by the Spanish Ministry of Science and Innovation through the Ramon y Cajal Research Fellowship, PID2019–105528GA-I00.

* Corresponding author.
E-mail address: j.tan@westernsydney.edu.au (S.H. Jessica Tan).

https://doi.org/10.1016/j.neuroimage.2022.119217.
Received 7 November 2021; Received in revised form 9 April 2022; Accepted 14 April 2022
Available online 15 April 2022.

ground voice when the auditory stimuli were paired with videos of a speaker's talking face, but not when they were paired with a still image of the speaker's face (Hollich et al., 2005). Studies with children and adults found that children identified phonemes and words better in the auditory-visual modality compared to the auditory-only modality, and that this benefit is evident both in quiet (Lalonde and Holt, 2015) and in noise (Lalonde and Holt, 2016; Maidment et al., 2015; Ross et al., 2011). Additionally, comparisons between children and adults revealed that adults experienced greater auditory-visual speech benefit (Maidment et al., 2015; Ross et al., 2011).

The same developmental trend has been found in neurophysiological studies with children and adults. Knowland et al. (2014) presented 6- to 11-year-olds and adults with auditory-visual words and with auditory-only words. Both the children and the adults showed attenuated amplitude and shorter latencies of the auditory P2 event-related potential (ERP) component to auditory-visual compared to auditory-only words, but the adults additionally showed the attenuated amplitude and shorter latencies for N1 for auditory-visual compared to auditory-only stimuli. Together these results suggest that visual speech modulation of auditory ERP components is present, yet not fully developed in children (Knowland et al., 2014).

Other ERP studies have measured speech perception in auditory-visual vs auditory-only and visual-only speech in terms of integration rather than enhancement. The criterion for auditory-visual integration is based on the relative magnitude of neural responses to auditory-visual (AV) stimuli compared with the summation of neural responses to auditory-only (A) and visual-only (V) stimuli [i.e., by testing whether AV = (*A* + V) no integration, or whether AV > *A* + *V*, integration]. Using this method, Kaganovich and Schumaker (2014) revealed that peak amplitudes of N1 and P2, and the latency of P2 were attenuated in auditory-visual compared to the algebraic sum of ERP responses to auditory-only and visual-only /ba/, /da/, and /ga/ syllables in 7–8-year-olds, 10–11-year-olds, and adults, thereby indicating auditory-visual integration at all three ages. In a separate study, adult participants showed a significantly shorter latency of the auditory N1/P2 response peak when presented with /ka/, /pa/, and /ta/ in auditory-visual syllables than in auditory-only or visual-only syllables (van Wassenhove et al., 2005).

The same integration approach has not been used with infants; rather, the majority of the electrophysiological studies of auditory-visual speech perception in infants have involved the comparison of neural responses (in the form of ERPs) to congruent versus incongruent auditory-visual syllables (Bristow et al., 2009; Kushnerenko et al., 2008, 2013) and short phrases (Hyde et al., 2011; Reynolds et al., 2013). For example, Kushnerenko et al. (2008) examined 5-month-olds' neural processing of conflicting auditory-visual syllables that typically result in the McGurk effect. Congruent stimuli consisted of auditory-visual /ba/ and auditory-visual /ga/ while incongruent stimuli consisted of the McGurk effect stimuli (auditory /ba/ dubbed onto a visual /ga/ which usually results in a "da" or "δ_a" response) and a conflicting stimulus (auditory /ga/ dubbed onto a visual /ba/ which usually results in a combination, "bga", response). The ERPs in response to the conflicting stimulus were more positive over frontal areas and more negative over temporal areas compared to ERPs in response to the other stimulus types, suggesting that 5-month-olds detected the mismatch between the auditory /ga/ and visual /ba/ but integrated the auditory /ba/ and visual /ga/, treating it the same as they did for the integration of congruent auditory-visual stimuli. Similar findings were reported in a study that used short phrases. Hyde et al. (2011) presented 5-month-olds with an auditory recording of the phrase, "Oh, hi baby", that was either paired with a *matched* video of a face saying the same phrase or a *mismatched* video of a face saying a different phrase. Mean amplitude of visual N1 and attentional Nc components were more negative in the asynchronous than the synchronous condition, while mean amplitude of auditory P2 component was more positive in the synchronous than the asynchronous condition. Although these infant ERP studies provide some neural level evidence for auditory-visual integration by comparing neural responses to congruent versus incongruent auditory-visual stimuli, they did not include auditory-only and visual-only conditions and so do not truly quantify auditory-visual integration and, in addition, do not afford comparison with the modulating effect of visual information found in children and adults.

Beyond electrophysiological studies, the hemodynamic (fNIRS) approach has been used to investigate infants' processing of auditory-visual speech (Altvater-Mackensen and Grossman, 2016; 2018). The neural responses of six-month-old German-learning infants were enhanced in the left inferior frontal regions when they were presented with matched auditory-visual speech as compared to when they were presented with mismatched auditory-visual speech (Alvater-Mackensen and Grossman, 2016). A separate study compared infants' processing of unimodal auditory, visual, and multimodal auditory-visual speech at the neural level by presenting six-month-old German-learning infants with unimodal and multimodal speech stimuli /a/, /e/, and /o/ (Altvater-Mackensen and Grossman, 2018). This study revealed that the infant participants did not show differential responses to unimodal and multimodal speech within the frontal regions and between hemispheres.

Taken together, ERP studies with adults and children illustrate that auditory-visual integration occurs at a neural level and suggest that visual speech information is beneficial for speech perception. In contrast, infant ERP studies demonstrate only the detection of a mismatch between auditory and visual stimuli, and do not show whether visual speech information augments infants' speech perception, i.e., whether there is an auditory-visual speech benefit. The fNIRS approach used with infants did not find any difference in neural responses to unimodal or multimodal speech within frontal regions In addition to the paucity of studies investigating auditory-visual speech benefit in infants, a major drawback of these studies in general is that in order to evoke brain responses they require presenting participants with multiple repetitions of identical short stimuli which are averaged and then compared between conditions. In the case of auditory-visual speech perception, this comprises the use of syllables or short phrases, stimuli that are not entirely representative of natural, conversational speech.

A recent approach addresses this drawback by assessing cortical tracking, or the mathematical relationship between the speech dynamics and the corresponding brain responses (e.g., Ding and Simon, 2012; Fiedler et al., 2019; Golumbic et al., 2013; Gross et al., 2013; J. O'Sullivan et al., 2014). This approach has greater ecological validity than ERP approaches, as it allows the use of continuous stimuli rather than discrete, repeated stimuli, e.g., rather than single words, passages that more closely resemble natural speech, such as audiobooks or podcasts. Accordingly, this method has been increasingly used to examine auditory-only speech perception in adults (e.g., Ding and Simon, 2013; Ding et al., 2016), children (Di Liberto, Peter, et al., 2018; Vander Ghinst et al., 2019), and infants (e.g., Jessen et al., 2019; Kalashnikova et al., 2018). Even so, the few studies conducted with adults so far suggest that cortical tracking is augmented when visual speech information from a speaker's talking face is provided (e.g., Crosse et al., 2015; Crosse et al., 2016; O'Sullivan et al., 2019). Importantly, although there is evidence that cortical tracking of speech can be reliably measured in children and infants, whether cortical tracking of auditory-visual speech is enhanced in children and infants remains an open question, one that this paper will address.

The auditory-visual speech benefit effect rests upon the assumption that listeners attend to a speaker's *facial* movements. It is thus somewhat surprising that most auditory-visual speech perception studies do not concurrently examine participants' looking behaviour to the speaker's face (although see Foxe et al., 2015). It has been shown that while the eyes convey emotional and social information, the mouth translates information closely related to the temporal and acoustic properties of speech (Yehia et al., 1998). Face viewing studies indicate that humans are cognisant of the various types of information that different facial features provide and will shift their gaze from one facial

region to another accordingly (e.g., Buchan et al., 2008; Lansing and McConkie, 1999). This attentional shift is observed even in infants as young as 6 months (Tenenbaum et al., 2013). In addition, idiosyncratic differences between individuals in facial scanning patterns of the eye and mouth regions are related to perceptual performance (Gurler et al., 2015; Mehoudar et al., 2014; Peterson and Eckstein, 2012). For instance, Gurler et al. (2015) found that individuals who report experiencing the McGurk effect more frequently also spend a larger proportion of time fixating on the speaker's mouth. This finding points toward the strong likelihood that individuals' idiosyncratic preferences in their fixation of the speaker's mouth or eyes will influence the extent to which visual speech information augments their speech perception.

Interindividual variations in looking behaviour to the speaker's face may result in subtle but significant differences in speech perception. For example, the opening and closing of the mouth corresponds to the syllabic timescale of auditory speech (Chandrasekaran et al., 2009), thus providing the richness of redundant cues relating to the start and end points of syllables that may augment speech perception, especially for listeners who fixate on the speaker's mouth region. This pertains particularly to young infants in normal listening conditions because they are just beginning to acquire a language system. In this regard, Lewkowicz and Hansen-Tift (2012) provided evidence of a developmental trend in looking behaviour: infants move away from preferential attention to the speaker's eye region to attending more to the speaker's mouth region sometime between 4 and 8 months, and then back to attending more to the speaker's eye region by 12 months of age. As this pattern coincides with the developmental timeline of speech production (Imafuku et al., 2019), the researchers propose that the initial eye-to-mouth attentional shift reflects infants' attempt to extract the redundant cues present in auditory-visual speech while the second attentional shift converges with adults' looking behaviour to a talking face and suggests some level of language expertise that reduces the need to focus specifically on the speaker's mouth (Lewkowicz and Hansen-Tift, 2012). Notably, relative attention to a talker's mouth at 6 months is positively related to expressive language skills both then (Tsang et al., 2018) and at 18 months (Young et al., 2009), and to receptive vocabulary at 12 months (Imafuku and Myowa, 2016). Failure to attend to the speaker's mouth is associated with later language learning disorders (Pons et al., 2019). Adults, by comparison, are proficient language users and instead focus more on the talker's eye region under optimal listening conditions but will increasingly direct their attention to the talker's mouth as listening situations become more challenging, such as when there is background noise (e.g., Buchan et al., 2008; Stacey et al., 2020; Vatikiotis-Bateson et al., 1998). These findings raise the possibility that individuals' idiosyncratic differences in looking patterns to a talking face will influence the degree of auditory-visual speech benefit experienced. Investigating whether this is indeed the case forms the second aim of this study.

1.1. This study and the hypotheses

To examine whether cortical tracking of auditory-visual speech is enhanced in infants and children, and whether gaze behaviour modulates the extent of auditory-visual speech benefit, EEG and gaze data were simultaneously recorded as 5-month-old and 4-year-old participants watched short clips of a speaker in auditory-only (AO), visual-only (VO), and auditory-visual (AV) presentation modes. AO presentations consisted of still photos of the speaker's face paired with auditory recordings, VO presentations consisted of silent videos of the speaker talking, and AV presentations consisted of both the videos and the auditory recordings. As this paradigm has been used previously with adult participants (Crosse et al., 2015; Crosse et al., 2016a, 2016b), a group of adults was tested as a control.

Behavioural studies illustrate that the auditory-visual speech benefit is evident across development. Neurophysiological studies show the same for children (using ERPs) and adults (using ERPs and cortical tracking), while none have yet directly examined the auditory-visual speech benefit in infants. Even so, ERP studies with infants that investigated their detection of auditory-visual asynchrony coupled with behavioural findings suggest that the auditory-visual speech benefit may also be evident at the neurophysiological level in infants.

With these considerations in mind, we hypothesise that, across the three age groups, (1) cortical tracking of the speech envelope will be most accurate during AV presentations, followed by AO then VO presentations, and (2) auditory-visual speech benefit will be evident as indexed by the additive criterion [i.e., AV > (A + V)]. Next, facial scanning and speech perception findings suggest that gaze behaviour may modulate cortical tracking accuracy differently for infants compared to children and adults. At five months, infants are likely to be in the process of shifting their attentional focus from the speaker's eyes to the speaker's mouth region (Lewkowicz and Hansen-Tift, 2012; Pons et al., 2015). Furthermore, 5-month-olds are in the process of acquiring language and may benefit from any additional information that can be extracted from visual speech cues. Accordingly, we hypothesise that the proportion of time that infants spend attending to the speaker's mouth will be positively correlated with cortical tracking accuracy when visual speech information is available, i.e., during VO and AV presentations. On the other hand, the same positive correlation is not expected for 4-year-olds and adults, given previous findings that older children and adults focus more on the speaker's eyes when the auditory speech signal is clear (e.g., Lewkowicz and Hansen-Tift, 2012), presumably because the acoustic properties from the auditory signal are sufficient for speech perception and they turn to the eyes to seek out emotional and social information that may not be conveyed as clearly by auditory speech.

2. Methods

2.1. Participants

Five-month-olds: A final sample of eighteen 5-month-old infants from Australian English monolingual backgrounds were included (M age = 5.49 months, SD = 0.30 months, 8 females). This sample size was decided upon by drawing on previous neurophysiological studies that investigated infant neural processing of AV asynchrony (e.g., Hyde et al., 2011; Kushnerenko et al., 2008; Reynolds et al., 2013) and compared children's and adults' neural processing of AV speech (e.g., Kaganovich and Schumaker, 2014; Knowland et al., 2014). An additional 20 babies were tested but excluded because of fussiness (n = 6), excessively noisy EEG recordings (n = 11), or insufficient gaze data (n = 3). The attrition rate in this study is not uncommon for infant EEG studies (e.g., deBoer et al., 2007; Hyde et al., 2011; Reynolds et al., 2013). All infants came from a monolingual Australian English-speaking background.

Four-year-olds: A final sample of 19 Australian English monolingual 4-year-olds were included (M age = 4.16 years, SD = 0.14 years, 12 females). An additional 14 children were tested but excluded because they were very fidgety and did not complete the experiment (n = 5), had excessively noisy EEG recordings (n = 3), or had insufficient gaze data (n = 7).

Adults: A final sample of 18 Australian English monolingual adults aged between 18 and 56 years were included (M age = 23.42 years, SD = 8.75 years, 15 females). An additional eight adults were tested but excluded because seven had insufficient gaze data, and one experienced technical failure.

All infants and children were born full-term, not at-risk for any cognitive or language delay, with normal hearing and vision, and no history of ear infections. Prior to the study, their parents provided written informed consent, were briefed about the procedure and told that the session would terminate immediately if they wished so, or if their child showed any signs of distress during the session. All adult participants had self-reported normal hearing and normal or corrected-to-normal vision, were free of neurological diseases, and provided written informed

consent. Adult participants took part in this study as part of a Psychology course requirement and received research participation points. This study was approved by the Human Research Ethics Committee at Western Sydney University (approval number H11517). The approved protocol regarding participant recruitment, data collection and data management was adhered to.

For all groups of participants, noisy EEG recordings were defined as datasets that contain more than 20 bad channels as in previous infant studies (e.g., Kalashnikova et al., 2018). Additionally, for analysis purposes, participants were required to have at least 10 out of 30 common trials across the three conditions (auditory-only, visual-only, and auditory-visual) with a minimum of 15% attention (as calculated by $attention = \frac{total\ fixation\ duration\ to\ screen\ during\ trial}{trial\ duration}$) to be included in the final sample. The exclusion criterion for attention (at least 15% attention in a minimum of 10 common trials) was decided upon because previous eye-tracking studies with young infants have used similar exclusion criterion (e.g., 15% in LoBue et al., 2016; 20% in Taylor & Herbert, 2013). As infant EEG studies have a typical attrition rate of 50–75% (deBoer et al., 2007), the lower bound of 15% attention was chosen to reduce further data loss. The mean number of trials (per condition) included in the analyses are 15.83 for infants, 21.26 for 4-year-olds, and 25.61 for adults. The mean levels of attention across conditions are 56.24% for infants, 62.66% for 4-year-olds, and 79.95% for adults.

2.2. Stimuli

Audiovisual recordings of 30 short speech passages were made by a female native speaker of Australian English experienced in producing infant-directed speech (see Appendix A for transcripts). To allow for infants' limited attention span these passages were relatively short, but long enough to ensure an amount of EEG recording that was sufficient for analyses (Crosse et al., 2021). These speech passages were adapted from Richoz et al. (2017) or from recordings of infant-directed speech between mothers and their babies and varied in durations from 8.44 s to 16.35 s (M = 11.35 s, SD = 1.76 s). The recordings consisted of a close-up of the speaker's face and shoulders against a white background. There were three presentation modes, auditory-only (AO), visual-only (VO) and auditory-visual (AV) with the unimodal auditory and visual recordings extracted separately from the auditory-visual recordings. In the auditory-only condition, a still image of the speaker's resting face was shown on the screen as the auditory track was played. In the visual-only condition, the dynamic video of the speaker's talking face was presented in silence. In the auditory-visual condition, both the dynamic video and its soundtrack were played together. The auditory recordings have a sampling rate of 44.1 kHz and a 16-bit resolution. The 30 speech passages were presented in three blocks. Each block consisted of 10 speech passages that were presented once in each modality. Presentation order was randomised across modalities in such a manner that the same sentence did not appear in two modalities on consecutive trials.

Attention-getter stimuli were used throughout the experiment to maintain participants' attention. The type and frequency differed between age groups. For 5-month-olds, attention-getters consisted of 2-s animations (often used in the infant calibration routine in Tobii Studio) that appeared after each trial. For 4-year-olds and adults, attention-getters consisted of different pictures of 'Minions' that appeared in a random order after either two or three trials, with their frequency randomly determined. In addition, a different 3-s cartoon animation was played to mark the end of the block and to re-engage participants.

2.3. Procedure

2.3.1. Five-month-olds

Infants sat on their caregiver' laps approximately 70 cm away from the centre of an LCD screen. Continuous EEG data were recorded with a 128-channel Hydrocel Geodesic Sensor Net (HCGSN), NetAmps 300 amplifier, and NetStation 4.5.7 software (EGI Inc) at a sampling rate of 1000 Hz, with the reference electrode placed at Cz. Electrode impedances were kept below 50 kΩ. The EEG recordings were saved for offline analyses.

Stimulus presentation was controlled using Presentation software (Neurobehavioural Systems). Triggers indicating the start and end of each trial were recorded along with the EEG. Eye-tracking recordings were co-registered with EEG recordings for two purposes: (i) to ensure that infants were attending to the visual stimuli and (ii) to examine whether gaze behaviour to the mouth region modulates cortical tracking of the speech envelope. To this end, a Tobii X120 eye tracker was placed below the screen to gather gaze fixation data.

As the entire duration of the session was quite long for an infant study (approximately 25 min), the stimuli continued to play until infants showed signs of fussiness or until completion, whichever came first.

2.3.2. Four-year-olds and adults

The procedure for 4-year-olds was identical to that for 5-month-olds with two exceptions. First, 4-year-olds were seated on their own. Second, the session was framed as a game; in order to motivate children to focus on the screen, children were required to press a button on a response pad whenever a picture of a *Minion* appeared on the screen (Kaganovich and Schumaker, 2014).

Adult participants were informed prior to the start of the experiment that they are part of a control group for an infant and child study. The procedure for adults was similar to 4-year-olds, except that adults also participated in a second EEG task which used similar stimuli but in adult-directed speech (ADS). Its order of presentation (immediately before or after the first task) was counterbalanced between participants (the results of this ADS session are not reported here).

2.3.3. EEG measure

2.3.3.1. Pre-processing. EEG data were pre-processed using EEGLAB (Delorme and Makeig, 2004), FieldTrip (Oostenveld et al., 2011), NoiseTools (http://audition.ens.fr/adc/NoiseTools/), the mTRF Toolbox (Crosse et al., 2016) and custom scripts in MATLAB R2019a (The Mathworks, Inc). First, EEG data from the three outer rings of the net were removed because these channels have been found to be very noisy in infants and children (Di Liberto et al., 2018; Folland et al., 2015; Kalashnikova et al., 2018). EEG data from the remaining 92 channels were high-pass filtered at 0.1 Hz, low-pass filtered at 12 Hz with Butterworth 8th order filters. As infant and child EEG recordings are noisy due to movements, artefact subspace reconstruction (ASR; Kothe and Jung, 2014) was applied to remove noise. ASR uses a sliding window technique whereby each EEG window is decomposed via principal component analysis. Each EEG window is then statistically compared with reference EEG data obtained from clean portions of the EEG recording. Within each window, the ASR algorithm searches for principal subspaces that significantly deviate from the reference EEG data. These subspaces are rejected and then reconstructed using a mixing matrix computed from the reference EEG data (Chang et al., 2019). As in Kalashnikova et al. (2018), this study used a sliding window of 500 ms and a threshold of 20 standard deviations to identify corrupted subspaces. Noisy channels that were removed during ASR were replaced with an estimate of neighbouring clean channels using spherical interpolation. Finally, EEG data were re-referenced to the average of all channels (e.g., Kalashnikova et al., 2018) and later downsampled to 100 Hz to reduce processing time.

To investigate the impact of visual speech cues on the cortical tracking of auditory speech, the speech stimuli were pre-processed in a manner following Jessen et al. (2019). The auditory soundtracks of each video were extracted, downsampled to 100 Hz to match the sampling rate of the EEG data and characterised using the broadband speech envelope of the acoustic signal through the NSL toolbox that models the auditory peripherical and subcortical processing stages (Ru, 2001). A spectrogram representation of each stimulus contained band-specific envelopes of 128 logarithmically-spaced frequency bands between 0.1 and

4 kHz. The broadband temporal envelope of each soundtrack was obtained by summing up the band-specific envelopes across all frequencies.

2.3.3.2. Data analysis. Cortical tracking of the speech envelope was measured by mathematically modelling response functions that describe the linear mapping between the stimulus speech envelopes and the corresponding neural responses. For this study, the stimulus-response mapping function is modelled in the forward direction (see Crosse et al. (2016) for details), i.e., the resulting model describes an optimal linear transformation from the stimulus domain to the neural-signal domain. Such a model is fit by conducting a lagged ridge regression between the envelope and the EEG data while accounting for likely time-delays between the acoustic input and the corresponding EEG response. The regression weights obtained with this procedure estimate the temporal response function (TRF) between envelope and EEG at each EEG channel. Significant non-zero weights reflect EEG channels where cortical activity is related to stimulus encoding (Haufe et al., 2014). TRFs are similar to event-related potentials (ERPs) in that they allow for an examination of the amplitude, latency, and scalp topography of the stimulus-EEG relationship. Specifically, the distribution of TRF weights can be examined across the scalp at different latencies, or different relative time lags between the ongoing speech and EEG signals. For example, a time lag of 100 ms refers to the impact that a change in the speech stimulus at time t has on the EEG at time $t + 100$ ms.

To investigate neural tracking of continuous stimuli, adult studies commonly compute response functions based on a subset (e.g., $n - 1$ trials) of the available data from each participant (e.g., Crosse et al., 2015), resulting in TRFs that are then used to model responses for the nth trial for each participant. This approach—*subject-dependant* modelling—requires lengthy datasets for each participant that are typically unattainable for the infant population. To account for the limited amount of available data from the infant sample, the *subject-independent* approach (Di Liberto and Lalor, 2017) was used for this study. Instead of computing an individual response function for each participant, this approach involves computing an average response function over $n - 1$ participants that is then used to predict the EEG signal of the nth participant via leave-one-out cross-validation. The subject-dependant modelling approach has been shown to yield better results than the subject-dependant modelling approach when used with 5-min EEG recordings from 7-month-olds and adults (Jessen et al., 2019). Subject-dependant modelling was used for each age group. In other words, an average response function was computed for each age group to predict the EEG signal of the nth participant from that age group.

Initially, TRFs were calculated for each stimulus at time lags between −200 and 1000 ms before selecting a temporal region of the TRF (0–600 ms) that included all relevant components to map the stimulus to the EEG signal with no visible response outside of this range. Leave-one-out cross-validation using Tikhonov regularization was conducted to assess how well the unseen EEG data could be predicted based on the TRF. The regularisation parameter of the ridge regression was set to $\lambda = 100$ for all participants. The lambda parameter value was chosen to mitigate the potential failure of lambda tuning due to the limited amount of data available (for a discussion, see Crosse et al., 2021). Prediction accuracy was quantified by calculating the Pearson correlation coefficient between the predicted and original EEG responses at each electrode. If EEG data is indeed reflecting the encoding of the speech envelope, then the correlation values would be significantly greater than zero. To investigate auditory-visual speech benefit, $(A + V)$ TRFs were computed and compared to AV TRFs in accordance with the additive criterion. The additive criterion was chosen to investigate auditory-visual speech benefit because this was used in previous studies with similar paradigms (e.g., Crosse et al., 2015, 2016). The AV speech benefit was quantified as the difference in prediction accuracy for AV TRFs relative to $A + V$ TRFs.

Table 1
Means (and Standard Deviations) of spatial offsets (Measured in Pixels) in gaze data for each age group.

	5-month-olds	4-year-olds	Adults
X-coordinate	39.91 (519.75)	72.85 (278.33)	33.26 (159.45)
y-coordinate	25.37 (225.46)	98.80 (315.86)	164.78 (130.44)

Fig. 1. Areas of interest (AOIs) defined for the speaker's eye and mouth regions.

2.3.4. Gaze measures

Means and standard deviations of the spatial offsets (x- and y-coordinates) for each age group are reported in Table 1. As 5-month-olds and 4-year-olds were more fidgety than adults during the study, there was a considerable amount of data loss from the eye-tracker for those groups. To circumvent the cumulative effect of data loss due to gaze as measured by the eye-tracker and to noisy EEG data, videos of participants who met the EEG data inclusion criterion ($\leq$ 20 noisy channels) but had eye-tracking issues (i.e., participants were looking at the screen but their gaze was not detected by the eye-tracker) were coded frame-by-frame manually using ELAN software (version 5.9) for whether or not they were looking at the screen. This resulted in hand-coded videos for 11 four-year-olds, and 3 five-month-olds.

Areas of interest (AOIs) covering the top half and bottom half of the speaker's face demarcated the speaker's eye and mouth regions (Fig. 1). These AOIs were of equal dimensions (640×340 pixels) and were adjusted using the derived mean spatial offsets of each age group. The proportion of total looks (PTLs) to these AOIs, in addition to attention, were computed for each trial:

1 Attention $= \left[\frac{total\ fixation\ duration}{trial\ duration}\right]$, (hereafter referred to as Attention) and
2 Proportion looking to the speaker's mouth region (hereafter referred to as PTL Mouth) $= \left[\frac{total\ fixation\ duration\ to\ mouth}{total\ fixation\ duration\ to\ mouth + total\ fixation\ duration\ to\ eyes}\right]$.

Note that PTL Mouth is a relative measure of attention to the mouth compared to eyes, so chance is 0.5, scores > 0.5 show greater fixation to mouth than eyes and scores < 0.5 show greater fixation to eyes than mouth. All statistical analyses on these two gaze measures were conducted using custom scripts in MATLAB R2019a (The MathWorks, Inc). The 11 four-year-olds and 3 five-month-olds whose gaze data were manually coded were only included for analyses that examined attention to screen—they were excluded from analyses that involved PTL Mouth.

2.4. Statistical analyses

Estimates of global field power were computed and topographic maps of TRF weights plotted to inspect the scalp regions where responses

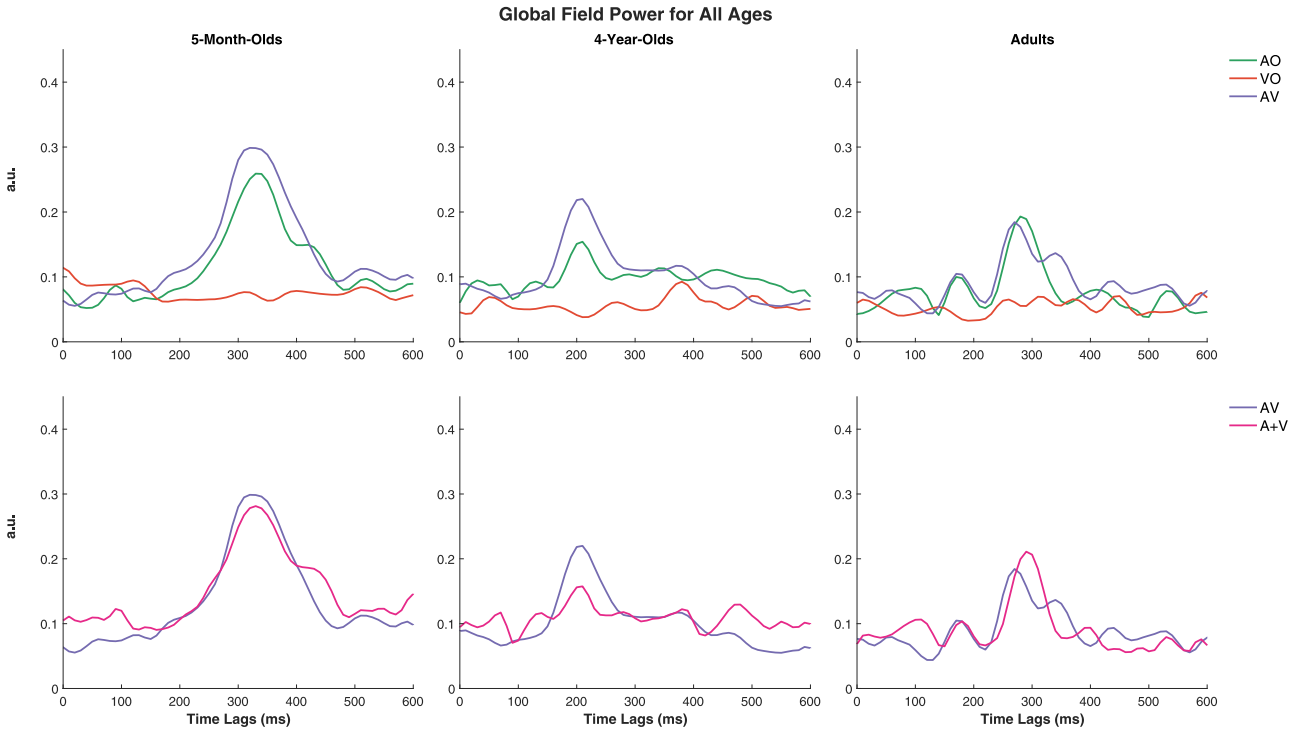

Fig. 2. Global field power measured at each time lag for all ages.

to the speech envelope were greatest. Mean TRFs were then computed for those scalp locations identified as regions of interest (ROIs) for each condition.

To evaluate model performance, mean prediction accuracies were obtained by averaging across all electrodes belonging to the ROIs and then tested against zero. Additionally, these mean prediction accuracies were compared between conditions to investigate the auditory-visual speech benefit and any age differences in model performance. As age-related anatomical differences may influence cortical tracking between groups independent of effects due to speech modality, TRF components and their respective prediction accuracy were not directly compared statistically between age groups.

To examine gaze behaviour, ANOVAs were conducted for each age group to examine the differences in attention and proportion looking at speaker's mouth between conditions (see Eqs. (1) and (2)). To examine the relationship between gaze behaviour and cortical tracking, Pearson's correlations were conducted for each condition between (1) cortical tracking and attention, and (2) cortical tracking and looking preference for each age group, where cortical tracking is quantified by TRF prediction accuracy.

3. Results

3.1. Prediction accuracies

First, as a preliminary step, global field power (GFP) — a reference-independent measure of response strength across the entire scalp at each time lag (Murray et al., 2008) — was estimated by calculating the TRF variance across all channels. The temporal profile of GFP for each age group showed clear TRF components at ~200–400 ms for AO, AV and (*A* + *V*), but not VO (Fig. 2). Topographies of TRF weights (Figs. 3–5) revealed that the observed components were mainly located over the frontal, occipital and temporal scalp regions. To avoid diluting the effects of interest, subsequent analyses of TRFs were therefore focused on the frontal, occipital, and temporal groups of electrodes. These groupings were used in previous infant (e.g., Folland et al., 2015; Peter et al., 2016) and child (e.g., Corrigall and Trainor, 2014) EEG studies to examine the average responses across scalp regions (Fig. 6).

Table 2
Mean prediction accuracies (and Standard Deviations), quantified by pearson's r, of TRFs from frontal, temporal and occipital scalp ROIs for each condition and age group.

	AO	VO	AV	*A* + *V*
5-month-olds	.021 (0.018)	.001 (0.008)	.035 (0.019)	.032 (0.018)
4-year-olds	.020 (0.018)	−0.005 (0.011)	.018 (0.020)	.014(0.015)
Adults	.009 (0.011)	.0004 (0.011)	.022 (0.015)	.007 (0.012)

To examine the *presence* of envelope tracking, TRF prediction accuracies at the three scalp ROIs were tested against zero. To assess the difference in the *extent* of envelope tracking, these prediction accuracies were then compared between conditions. Of interest are (1) the differences between cortical tracking of AO, VO and AV speech, and (2) the presence of an auditory-visual speech benefit as quantified by the additive criterion [i.e., AV vs. (*A* + *V*)]. One-sample *t*-tests were first conducted to test prediction accuracies against zero. Next, one-way ANOVAs were conducted for each age group with their respective prediction accuracies as the dependant variable to examine whether prediction accuracies differed between conditions. Subsequent post-hoc comparisons were conducted using two-tailed paired-sample *t*-tests with Bonferroni-adjusted alpha levels where multiple comparisons were made. The same analyses were conducted with 15 randomly selected trials per condition for 4-year-olds and adults to examine whether different amounts of data from each age group influenced the results. Fifteen trials were chosen because infant data had the least number of trials included with approximately 15 trials per condition.

3.1.1. Evidence of cortical tracking

All means and standard deviations of prediction accuracy for each condition and age group are set out in Table 2.

Five-month-olds: One-sample *t*-tests indicated that prediction accuracy of AO, AV, and (*A* + *V*) TRFs were significantly greater than zero

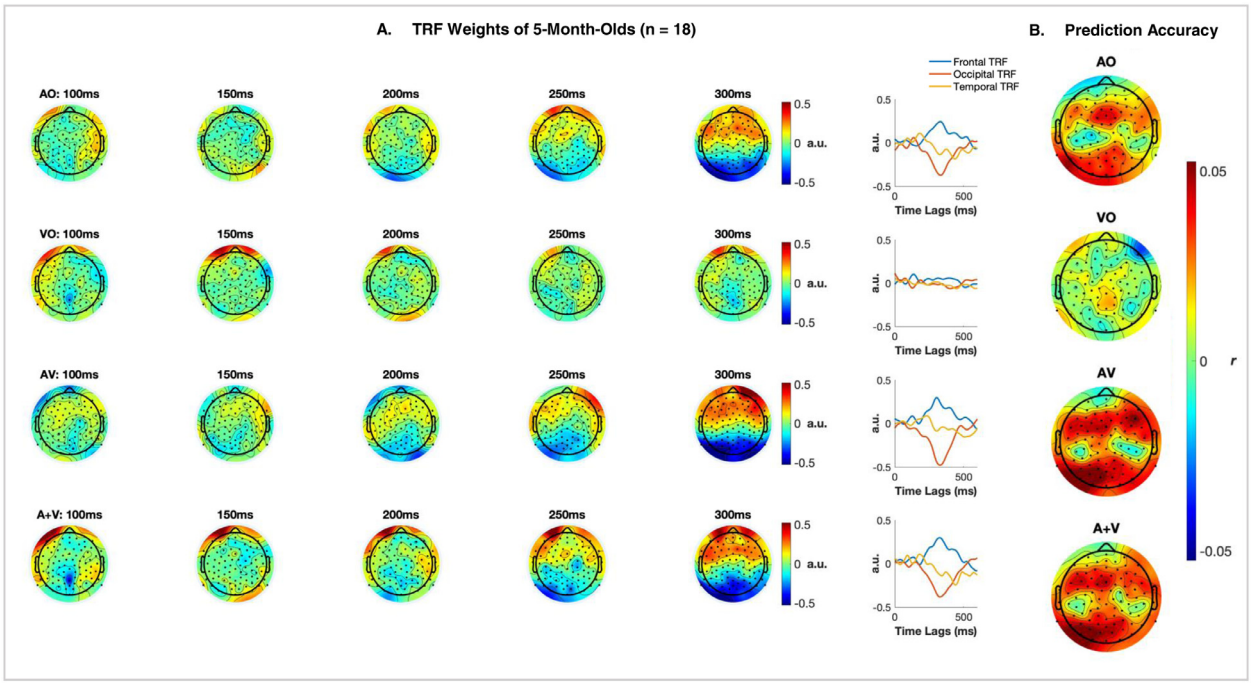

Fig. 3. (A) Topographies and TRFs of frontal, occipital and temporal locations, and (b) prediction accuracy of TRFs from 5-month-olds' data.

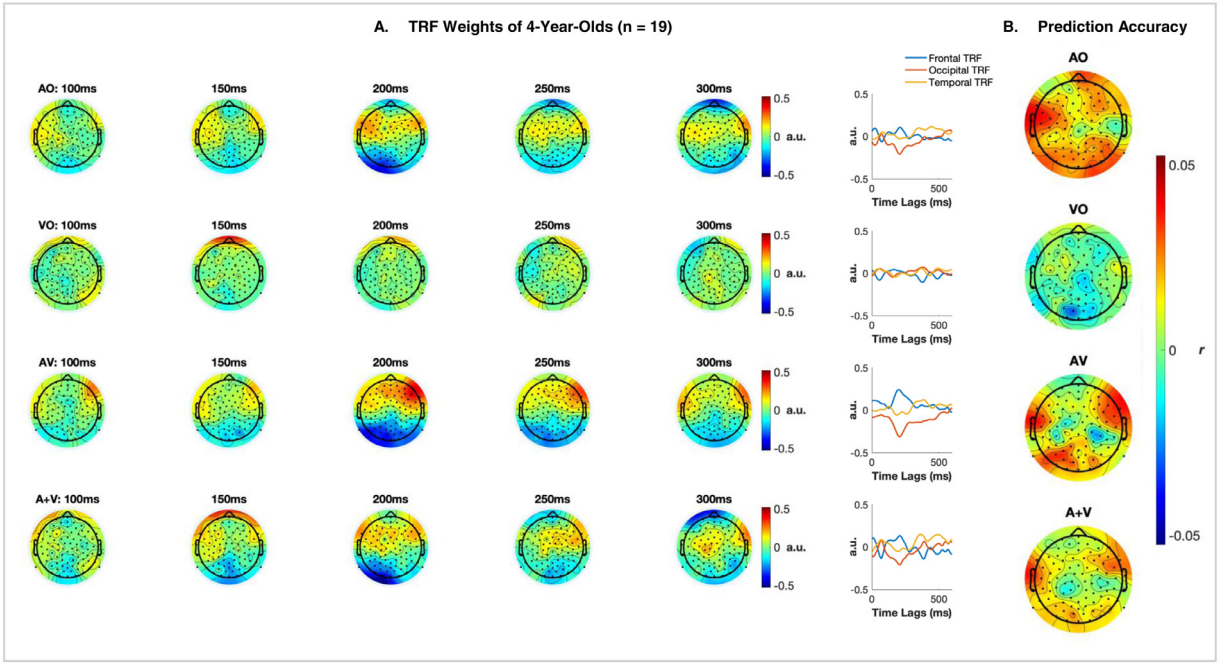

Fig. 4. (A) Topographies and TRFs of frontal, occipital and temporal locations, and (b) prediction accuracy of TRFs from 4-year-olds' data.

(AO: $t(17) = 5.15$, $p = < 0.001$, Hedges' $g = 1.16$; AV: $t(17) = 7.47$, $p < .001$, Hedges' $g = 1.68$,; *A + V:* $t(17) = 7.42$, $p < .001$, Hedges' $g = 1.67$), but prediction accuracy of VO TRFs was not significantly greater than zero, $t(17) = 0.75$, $p = .23$, Hedges' $g = 0.17$.

Four-year-olds: Prediction accuracies of AO, AV, and (*A + V*) TRFs were significantly greater than zero (AO: $t(18) = 4.93$, $p = < 0.001$, Hedges' $g = 1.08$; AV: $t(18) = 3.86$, $p < .001$, Hedges' $g = 0.85$; *A + V:* $t(18) = 3.96$, $p < .001$, Hedges' $g = 0.87$), whereas prediction accuracy of VO TRFs was not significantly greater than zero ($t(18) = -2.13$, $p = .98$, Hedges' $g = -0.47$). The analyses with 15 trials revealed only one difference: prediction accuracy of VO TRFs was significantly lower than zero ($t(18) = -2.38$, $p = .03$, Hedges' $g = 0.48$.

Adults: Prediction accuracies of AO, AV, and (*A + V*) TRFs were significantly greater than zero (AO: $t(17) = 3.49$, $p = .001$, Hedges' $g = 0.79$; AV: $t(17) = 6.11$, $p < .001$, Hedges' $g = 1.38$; *A + V:* $t(17) = 2.48$, $p = .012$, Hedges' $g = 0.56$), whereas prediction accuracy of VO TRFs was not significantly greater than zero, $t(17) = 0.17$, $p = .44$, Hedges' $g = 0.04$. Results from the analyses with 15 trials were not different.

3.1.2. *Difference in strength of cortical tracking between conditions*

The one-way ANOVAs testing between conditions (AO, VO, AV, *A + V*) were significant for all age groups (5-month-olds: $F(3, 68) = 14.95$, $p < .001$, $\eta_p^2 = 0.40$; 4-year-olds: $F(3, 72) = 9.63$, $p < .001$, $\eta_p^2 = 0.29$; adults: $F(3, 68) = 9.22$, $p < .001$, $\eta_p^2 = 0.29$). To in-

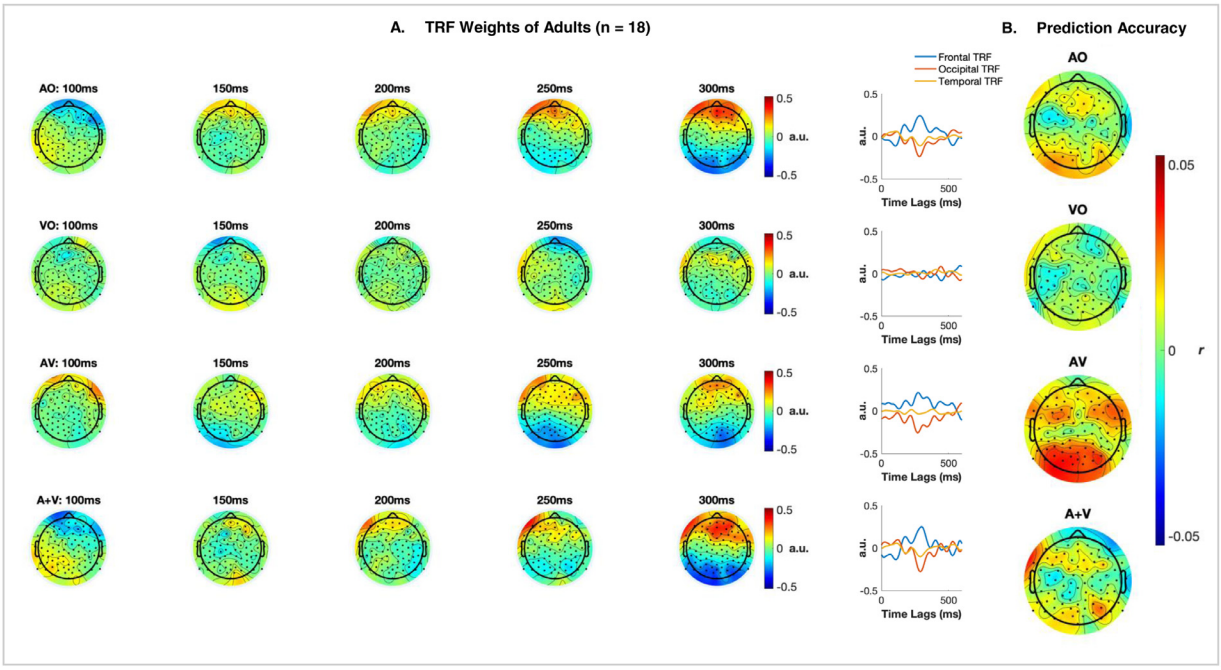

Fig. 5. (A) Topographies and TRFs of frontal, occipital and temporal locations, and (b) prediction accuracy of TRFs from adults' data.

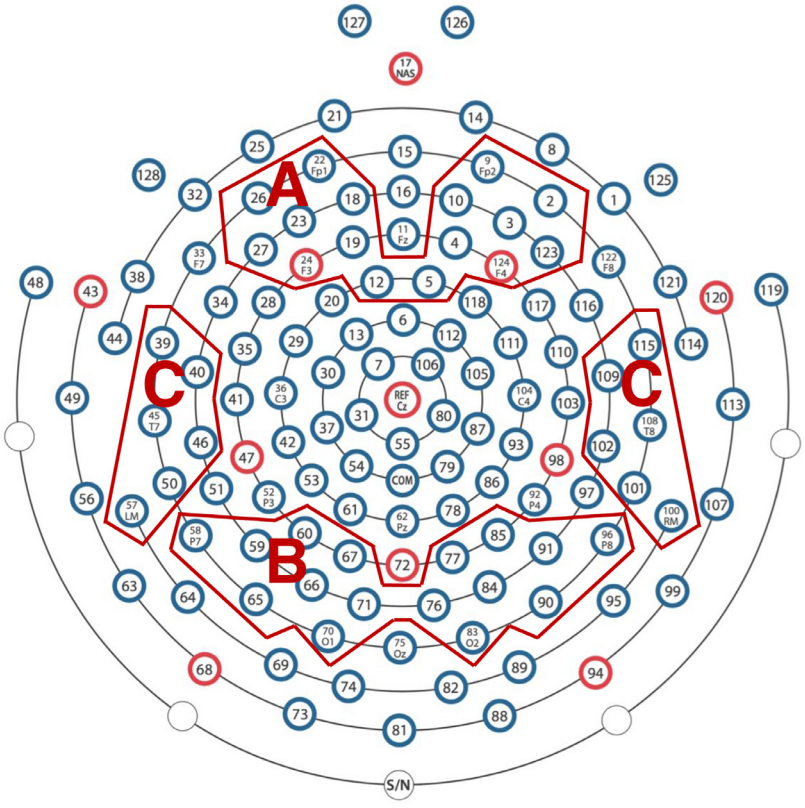

Fig. 6. Electrode groupings used for analyses. (A) frontal electrodes, (B) occipital electrodes, (C) temporal electrodes.

spect the differences between conditions and to identify whether there was auditory-visual speech benefit [i.e., AV > (*A* + *V*)], post hoc comparisons were subsequently performed using paired-sample *t*-tests with Bonferroni-adjusted alpha level of 0.013 (0.05/4).

Five-month-olds: When prediction accuracies of AO, VO, and AV TRFs were compared, paired-sample *t*-tests indicated that prediction accuracy of AV TRFs was greatest, followed by AO, then VO TRFs (AO vs. VO: $t(17) = 5.13$, $p < .001$, Hedges' $g = 1.42$; AO vs. AV: $t(17) = -4.07$, $p < .001$, Hedges' $g = -0.69$; AV vs. VO: $t(17) = 7.73$, $p < .001$, Hedges' $g = 2.15$). Prediction accuracy of AV TRFs was also significantly greater than (*A* + *V*) TRFs, $t(17) = 2.82$, $p = 0.001$, Hedges' $g = 0.16$, suggesting that auditory-visual speech benefit was present at the scalp ROIs.

Four-year-olds: Paired-sample *t*-tests revealed that the prediction accuracy of AO TRFs was significantly greater than that of VO TRFs

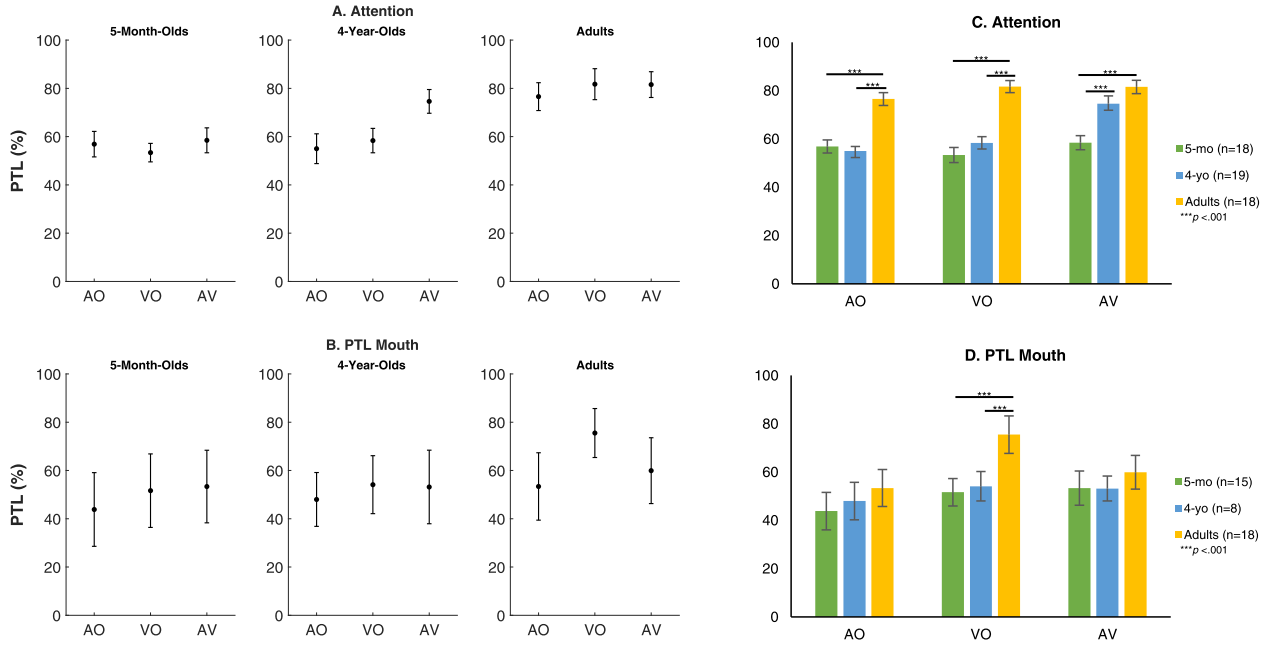

Fig. 7. Scatterplots of Attention (A) and proportion of total looking time to the mouth vs. Eyes (PTL Mouth) (B) for all conditions and age groups and their corresponding bar graphs (C: Attention; D: PTL Mouth). error bars represent standard errors of mean (SEM). With respect to attention, across age groups, greater attention was captured in the AV condition. With respect to the speaker's mouth, adults fixated the speaker's mouth to a greater extent in the AV condition than in AO and VO.

($t(18) = 5.66$, $p < .001$, Hedges' $g = 1.68$) but not significantly different from the prediction accuracy of AV TRFs ($t(18) = 0.58$, $p = 0.57$, Hedges' $g = 0.14$). The prediction accuracy of AV TRFs was significantly greater than that of VO TRFs ($t(18) = 4.75$, $p < .001$, Hedges' $g = 1.39$), but was not significantly greater than that of ($A + V$) TRFs ($t(18) = 1.06$, $p = 0.30$, Hedges' $g = 0.21$). The analyses with 15 trials had similar findings.

Adults: Paired-sample *t*-tests showed that the prediction accuracy of AV TRFs was greatest, followed by AO, then VO TRFs (AO vs. VO: $t(17) = 4.10$, $p < .001$, Hedges' $g = 0.78$; AO vs. AV: $t(17) = -3.85$, $p = .001$, Hedges' $g = -0.88$; AV vs. VO: $t(17) = 7.36$, $p < .001$, Hedges' $g = 1.57$). Prediction accuracy of AV TRFs was also significantly greater than ($A + V$) TRFs ($t(17) = 5.01$, $p < .001$, Hedges' $g = 1.06$), suggesting that auditory-visual speech benefit was present at the scalp ROIs. The analyses with 15 trials revealed only one difference: prediction accuracy of AO TRFs is not significantly different from that of VO TRFs, $t(17) = 1.97$, $p = .07$, Hedges' $g = 0.54$.

3.2. Gaze behaviour

3.2.1. Attention

Separate one-way within-subjects ANOVAs were conducted for each age group with Attention as the dependant variable (see Eq. (1) in Statistical Analyses) and Condition as the independent variable. The ANOVAs revealed a significant main effect of Condition for all age groups (5-month-olds: $F(2, 34) = 3.58$, $p = .04$, $\eta_p^2 = 0.17$; 4-year-olds: $F(1.44, 25.89) = 26.67$ with Greenhouse-Geisser correction, $p < .001$, $\eta_p^2 = 0.60$; adults: $F(2, 34) = 7.16$, $p = .002$, $\eta_p^2 = 0.30$). Subsequent post-hoc comparisons between conditions were made using paired-sample *t*-tests with Bonferroni-adjusted alpha level of 0.017 (0.05/3). Fig. 7 contains scatterplots and bar graphs of Attention and PTL Mouth for all conditions and age groups.

Five-month-olds: Attention was significantly greater in the AV than the VO condition ($t(17) = 2.93$, $p = .009$, Hedges' $g = 0.50$), but the differences between AO and VO and between AO and AV conditions were not significant (AO vs. VO: $t(17) = 1.49$, $p = .15$, Hedges' $g = 0.34$; AO vs. AV: $t(17) = -0.94$, $p = .36$, Hedges' $g = 0.14$).

Four-year-olds: Attention was significantly greater in the AV than in the AO condition ($t(18) = 6.10$, $p < .001$, Hedges' $g = 1.54$) and in the VO condition ($t(18) = 9.19$, $p < .001$, Hedges' $g = 1.43$), whereas the difference in attention between AO and VO conditions was not significant ($t(18) = -1.00$, $p = .33$, Hedges' $g = -0.26$).

Adults: Attention was significantly greater in the VO than the AO condition ($t(17) = 3.58$, $p = .002$, Hedges' $g = 0.38$) and in the AV than the AO condition ($t(17) = 3.06$, $p = .007$, Hedges' $g = 0.40$). The difference in attention between VO and AV conditions was not significant ($t(17) = 0.11$, $p = .91$, Hedges' $g = 0.01$).

Age comparisons: An Age x Condition mixed-design ANOVA was conducted with Attention as the dependant variable. The main effects of Condition and Age, and the Age x Condition interaction were significant (Condition: $F(1.68, 87.50) = 26.00$ with Greenhouse-Geisser correction, $p < .001$, $\eta_p^2 = 0.33$; Age: $F(2, 52) = 25.21$, $p < .001$, $\eta_p^2 = 0.49$; Age x Condition: $F(3.37, 87.50) = 12.47$ with Greenhouse-Geisser correction, $p < .001$, $\eta_p^2 = 0.32$). To examine the Age x Condition interaction, we conducted independent-samples *t*-tests for each condition. Five-month-olds attended less to the screen than 4-year-olds only in the AV condition ($t(35) = -4.45$, $p<.001$), whereas they attended to the screen similarly during AO and VO presentations (AO: $t(35) = 0.45$, $p = .65$; VO: $t(35) = -1.53$, $p = .13$). Five-month-olds attended less to the screen than adults in all conditions (AO: $t(34) = -4.94$, $p < .001$; VO: $t(34) = -7.43$, $p < .001$: AV: $t(34) = -6.10$, $p < .001$). Four-year-olds attended less to the screen in AO and VO conditions than adults but not during AV presentations (AO: $t(35) = -4.99$, $p < .001$; VO: $t(35) = -5.64$, $p < .001$; AV: $t(35) = -1.88$, $p = .07$).

3.2.2. PTL to the speaker's mouth

Separate one-way within-subjects ANOVAs were conducted for each age group (DV: PTL Mouth, IV: Condition). The ANOVAs were significant for 5-month-olds and adults (5-month-olds: $F(2, 26) = 4.98$, $p = .01$, $\eta_p^2 = 0.28$; adults: $F(1.35, 23.00) = 13.40$ with Greenhouse-Geisser correction, $p < .001$, $\eta_p^2 = 0.44$), but not for 4-year-olds ($F(2, 14) = 1.82$, $p = .20$, $\eta_p^2 = 0.21$). Subsequent analyses involved one-sample *t*-tests to assess whether PTL Mouth was significantly greater than chance and paired-sample *t*-tests with Bonferroni-adjusted alpha level of 0.017

(0.05/3) to examine whether looking preference differed between conditions.

Five-month-olds: One-sample *t*-tests indicated that infants' relative attention to the speaker's mouth region was not significantly greater than chance across conditions (AO: $t(14) = -0.72$, $p = .48$, Hedges' $g = -0.18$; VO: $t(13) = 0.19$, $p = .85$, Hedges' $g = 0.05$; AV: $t(13) = 0.14$, $p = .89$, Hedges' $g = 0.10$). Next, paired-sample *t*-tests indicated that infants' looking preference for the speaker's mouth was greater in the VO than the AO condition ($t(13) = 3.44$, $p = .004$, Hedges' $g = 0.16$), and in the AV than the AO condition ($t(13) = 2.82$, $p = .015$, Hedges' $g = 0.28$), but the difference between VO and AV conditions was not significant, $t(13) = 0.17$, $p = .87$, Hedges' $g = 0.01$.

Four-year-olds: One-sample *t*-tests indicated that PTL Mouth was not significantly greater than chance in any condition (AO: $t(7) = -0.23$, $p = .83$, Hedges' $g = -0.07$; VO: $t(8) = 0.49$, $p = .63$, Hedges' $g = 0.14$; AV: $t(8) = -0.09$, $p = .93$, Hedges' $g = 0.09$). Paired-sample *t*-tests indicated that the difference in proportion of time spent fixating on the speaker's mouth region did not differ between conditions (AO vs. VO: $t(7) = -1.80$, $p = .11$, Hedges' $g = -0.35$; AO vs. AV: $t(7) = -0.65$, $p = .53$, Hedges' $g = -0.14$; VO vs. AV: $t(8) = 2.28$, $p = .05$, Hedges' $g = 0.13$).

Adults: One-sample *t*-tests indicated that PTL Mouth was significantly greater than chance in the VO condition ($t(17) = 4.93$, $p < .001$, Hedges' $g = 1.11$), but not in the AO or AV conditions (AO: $t(17) = 0.47$, $p = .64$, Hedges' $g = 0.11$; AV: $t(17) = 1.43$, $p = .17$, Hedges' $g = 0.32$). Paired-sample *t*-tests indicated that adults spent the greatest proportion of time attending to the speaker's mouth in the VO, followed by the AV then the AO condition (AO vs. VO: $t(17) = -4.12$, $p < .001$, Hedges' $g = -0.81$; AO vs. AV: $t(17) = -2.57$, $p = .02$, Hedges' $g = -0.21$; VO vs. AV: $t(17) = 3.28$, $p = .004$, Hedges' $g = 0.58$).

Age comparisons: A mixed-design Age x Condition ANOVA was conducted with PTL Mouth as the dependant variable. The main effect of Condition and the Age x Condition interaction were significant, but not the main effect of Age (Condition: $F(1.71, 60.66) = 13.01$ with Greenhouse-Geisser correction, $p < .001$, $\eta_p^2 = 0.27$; Age: $F(2, 37) = 25.21$ $p = .40$, $\eta_p^2 = 0.58$; Age x Condition: $F(3.42, 60.66) = 12.47$ with Greenhouse-Geisser correction, $p = .03$, $\eta_p^2 = 0.41$). To investigate the Age x Condition interaction, independent-samples *t*-tests were conducted. Five-month-olds and four-year-olds did not differ in their mouth preference across conditions (AO: $t(21) = -0.31$, $p = .76$; VO: $t(24) = -0.21$, $p = .84$; AV: $t(23) = -0.01$, $p = .99$). Five-month-olds showed a lower mouth preference than adults in VO but not in AO and AV conditions (AO: $t(31) = -0.87$, $p = .39$; VO: $t(30) = -2.46$, $p = 0.02$; AV: $t(31) = -0.61$, $p = .55$). Likewise, four-year-olds showed a lower mouth preference than adults in VO but not in AO and AV conditions (AO: $t(24) = -0.44$, $p = 0.66$; VO: $t(28) = -2.40$, $p = .023$; AV: $t(26) = -0.55$, $p = .59$).

3.3. Relationship between prediction accuracy and gaze measures

To investigate the relationship between attention and cortical tracking, Pearson's correlations between attention and TRF prediction accuracy were conducted for each condition and age group. Attention was significantly negatively correlated with prediction accuracy only in VO condition for adults ($r(17) = -0.49$, $p = .038$). All other correlations were not significant (all rs < 0.30, all ps > 0.09; see Fig. 8 for details).

Pearson's correlations between PTL Mouth and prediction accuracy were conducted for each condition and age group to examine whether individual differences in proportion looking time to the speaker's mouth region (vs. eye region) is associated with the strength of cortical tracking. The correlation in 5-month-olds between relative attention to the speaker's mouth region and prediction accuracy was significant for VO condition ($r(14) = 0.54$, $p = .048$) but not for AO or AV and not for any condition for the 4-year-olds or adults (all rs $\langle$ 0.24, all ps $\rangle$ 0.07; see Fig. 9 for details).

4. Discussion

This study examined the auditory-visual speech benefit in infants, children and adults at the neurophysiological level. Brain responses from 5-month-olds, 4-year-olds, and adults to continuous auditory-only, visual-only, and auditory-visual speech were analysed via forward modelling of temporal response functions and evaluating their predictive power. auditory-visual speech benefit was inferred using the additive criterion, i.e., the difference in predictive power between auditory-visual TRFs versus the summation of unimodal auditory-only and visual-only TRFs [AV vs. $(A + V)$].

Cortical tracking results for the 5-month-olds and the adults were similar and in line with our hypotheses. First, for both 5-month-olds and adults, comparison of predictions of AO, VO and AV TRFs (averaged across all electrodes at the three scalp ROIs) showed that the AV TRFs had the greatest accuracy followed by the prediction accuracy of AO, then VO TRFs. Second, for both 5-month-olds and adults, when auditory-visual speech benefit was assessed by comparing the prediction accuracy of AV TRFs with the prediction accuracy of the sum of unimodal TRFs $(A + V)$, there was a significant auditory-visual speech benefit as reflected by significantly greater prediction accuracy of AV TRFs compared to $(A + V)$ TRFs. Compared to this, the results from the 4-year-olds were unexpected: the prediction accuracy of AV TRFs was not significantly greater than AO TRFs, and there was no auditory-visual speech benefit. In this regard, further studies with both infants and young children would be of interest.

Analyses of gaze behaviour showed subtle differences in looking patterns between age groups. First, fixation durations to the screen, the measure of Attention (see Eq. (1)), during presentations of AV speech were greater than during AO presentations, i.e., AV > AO, for 4-year-olds and adults, and during VO presentations, i.e., AV > VO, for 5-month-olds and 4-year-olds. Adults attended similarly to the screen during VO and AV speech, but this is to be expected because adults are better than infants and children at self-motivation and understanding the task. Five-month-olds' attention was similar for AO and AV speech, and for AO and VO speech. These findings suggest that children's and adults' attention to the screen was generally greater when visual speech information was available, but 5-month-olds attended to the screen as long as auditory information was available. Next, although 5-month-olds and adults both spent a larger proportion of time looking to the speaker's mouth region relative to the eye region in the two conditions when visual speech information was available (VO and AV) than when it was not (AO), only the adults showed greater relative attention to the speaker's mouth in the VO than in the AV condition. Four-year-olds' looking patterns to the speaker's mouth were not significantly different across conditions.

When we measured the correlation between gaze measures and cortical tracking accuracy, two significant relationships were found, both regarding cortical tracking of VO speech. First, 5-month-olds' relative attention to the speaker's mouth region was positively correlated with the accuracy of cortical tracking of VO speech. Second, adults' overall attention was negatively correlated with the accuracy of cortical tracking of VO speech, which was unexpected. We had hypothesised that individual differences in 5-month-olds' attentional preference for the speaker's mouth would be positively correlated to the accuracy of cortical tracking during VO and AV presentations, but the correlation analyses indicated that this was only true for VO speech. These correlational results must be taken with caution as prediction accuracy of VO TRFs for adults and 5-month-olds were not significantly greater than zero, suggesting that these results are not robust. We had also hypothesized that individual differences in children's and adults' attentional preference for the speaker's face (indexed by overall attention to the screen since the face took up approximately two-thirds of the screen) would be positively correlated to cortical tracking accuracy in VO and AV conditions, but this was not the case. These unexpected results are discussed further below.

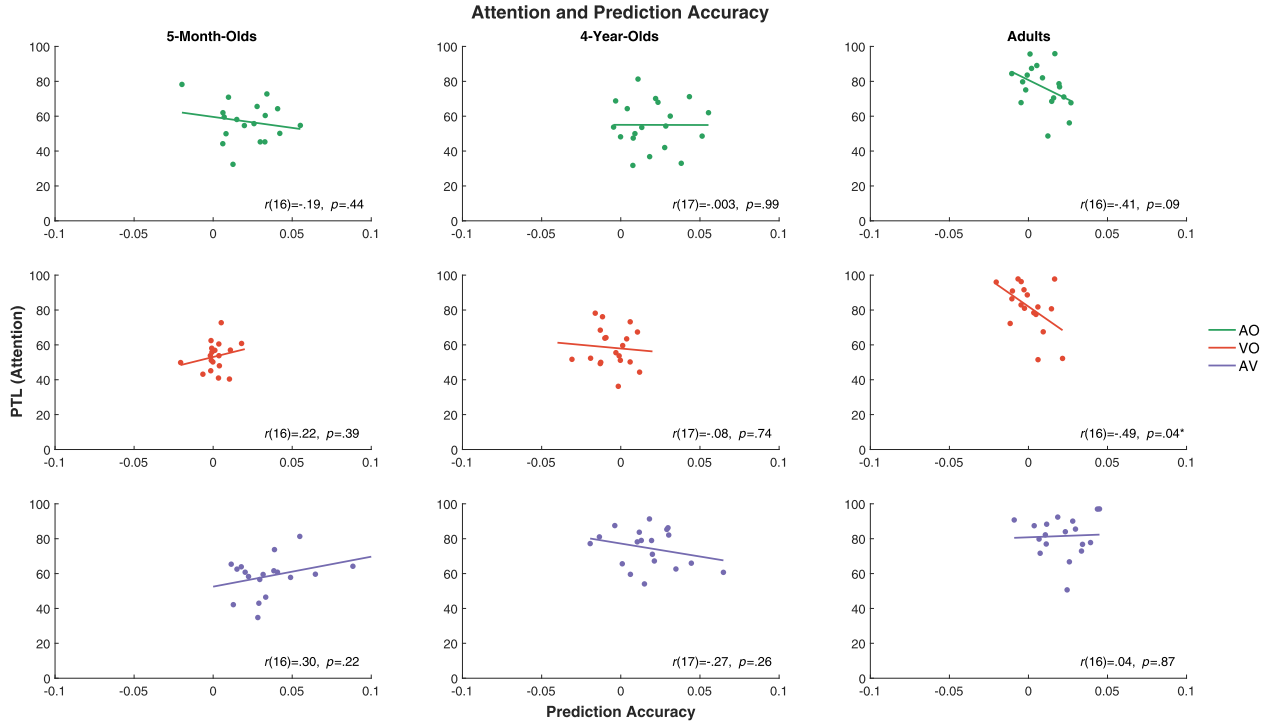

Fig. 8. Scatterplots and correlations between attention and prediction accuracy. As can be seen, only the negative correlation between adults' attention and prediction accuracy in VO condition was statistically significant.

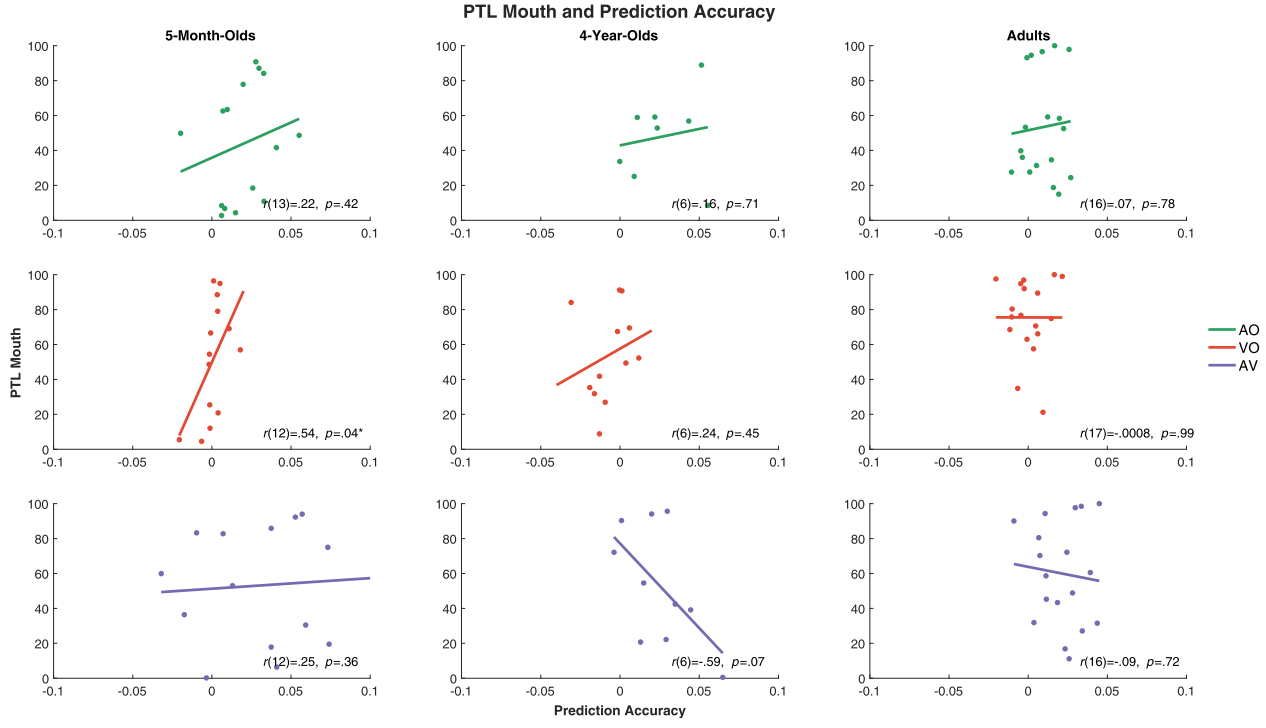

Fig. 9. Scatterplots and correlations between proportion of total looking times (PTL) mouth and prediction accuracy. As can be seen, only the positive correlation between 5-month-olds' relative attention to the speaker's mouth (vs. eyes) and prediction accuracy in VO condition was statistically significant.

Together, these results raise several noteworthy points. To begin, cortical tracking results of 5-month-olds and adults are in line with past studies. Behavioural studies with infants and adults have demonstrated that visual speech information improves performance on speech perception tasks. And more specifically, cortical tracking studies with adults have also shown that neural signals entrain to the speech envelope better during auditory-visual presentations compared to unimodal presentations (Crosse et al., 2016; Golumbic et al., 2013). Although previous neurophysiological studies with infants have shown that infants detect auditory-visual asynchrony (e.g., Hyde et al., 2011; Reynolds et al., 2013), none have directly examined the auditory-visual speech benefit. Results of 5-month-olds in this study is a first: not only are very

young infants sensitive to auditory-visual synchrony, but they are also processing visual speech information in such a way that enhances their perception of acoustic speech.

Current oscillation-based models of auditory-visual speech perception (Peelle & Sommers, 2015) posit that the onset of incoming visual speech cues resets the phase of ongoing oscillations in the auditory cortex (Mercier et al., 2015). This reset then allows for predictions of the upcoming auditory signal to be encoded (Arnal et al., 2011), and for the predicted input to be processed more easily (Friston, 2010; Henry and Obleser, 2012). The greater the amount of predictive information provided by visual speech cues, the higher the degree of auditory-visual speech benefit experienced (van Wassenhove et al., 2005). For example, visual speech cues related to the place of articulation will provide predictive information because these cues can readily be observed from the speaker's articulatory movements and are not affected by background noise (Grant and Bernstein, 2019). These oscillation-based models hold that phonemic level knowledge is necessary for this reset and subsequent reset-based predictions. However, whether 5-month-olds have acquired the phonemic repertoire required for such predictions is still unclear: phonemic acquisition was previously thought to occur during the second half of the first year (Friederici and Wessels, 1993; Jusczyk et al., 1994; Polka and Werker, 1994), but recent evidence suggests that infants at 3 months already show native-language phonological knowledge (Choi et al., 2017a, 2017b). Furthermore, even if the process of perceptual attunement is already in progress by 5 months, whether the knowledge accrued at 5 months is sufficient for phonemic-level predictions has yet to be determined.

It is possible that for 5-month-olds, the phase-reset of oscillatory activity by visual speech cues may instead serve to provide predictive information relating to the prosodic rhythm patterns of their native language since infants at this age are sensitive to these rhythmic properties (Nazzi et al., 2000). In this regard, cortical tracking in 8-month-old infants has been found to be better than in adults at the rhythmic and phonemic levels, whereas cortical tracking at the syllabic level did not differ between age groups (Leong et al., 2017). As rhythmic patterns have been associated with visual speech cues (Dohen et al., 2006), and as young infants already show a proclivity to associate temporally aligned auditory and visual information and perceive them as coming from a single source (e.g., Lewkowicz, 2003), it is theoretically reasonable that visual speech cues provide predictive information relating to the prosodic rhythm structure of the language. If this is indeed the case, then the phase-reset of oscillatory activity in the auditory cortex might augment speech perception differently for infants and adults: while the phase-reset may inform predictions at the phonemic level or prosodic level for infants, the phase-reset additionally serves to inform predictions at the syllabic level for adults. To verify this, in future studies phase-locking activity at the phonetic, syllabic and phrasal levels could be measured for unimodal and multimodal speech stimuli and compared between age groups (see Ding et al., 2015; Luo et al., 2010; Golumbic et al., 2013 for viable paradigms). Such studies will assist in elucidating the role that visual speech cues play and their interactions with age, and in addition, assist in the fine-tuning of models of auditory-visual speech perception. These studies would also further inform our unexpected results in the group of four-year-olds and provide replication for the results in the group of five-month-olds (we note that the effect size of the five-month-olds' visual speech benefit here is small, which calls for direct and indirect replications of this novel finding).

The second notable finding comes from gaze behaviour data: infants and adults direct their attention to the mouth when visual speech information is available (AV and VO conditions), hinting at the possibility that the looking patterns observed form part of an information-seeking strategy. The observed gaze patterns are in accordance with past studies (infants: Tenenbaum et al., 2013; adults: Birulés et al., 2020), and it has been postulated that young infants pay greater attention to a speaker's mouth than a speaker's eyes because they lack the language-specific phonological expertise in their native language to rely mainly or solely on the auditory signal (Lewkowicz and Hansen-Tift, 2012), whereas adults' relative attention to the speaker's mouth region increases in challenging listening conditions (Birulés et al., 2020). Thus here, infants and adults have the same overt behaviour directed at seeking linguistic information that may be due to different underlying causes; for infants they do not have sufficient native language experience to rely on the auditory signal, whereas for adults they seek visual speech information because they are in a difficult listening situation.

Next, the negative relationship between adults' attention and cortical tracking in the visual-only condition is, on first pass, somewhat surprising. However, this negative relationship may in part be explained by the type of speech (infant-directed speech) used for the stimuli. In infant-directed speech, facial expressions (Chong et al., 2003) and articulatory lip movements (Green et al., 2010) are exaggerated in addition to acoustic pitch (Kitamura et al., 2001) and prosody (Fernald and Mazzie, 1991). The exaggerated auditory and visual speech cues may come across as unnatural to adults and may interfere with adults' ability to effectively obtain additional acoustic and temporal information from visual speech cues. This line of reasoning is supported by previous findings that adults' ERP responses to IDS and ADS are different (Peter et al., 2016), indicating that their cortical processing of IDS differs from that of ADS. The lack of a significant relationship between relative attention to mouth and cortical tracking accuracy of visual-only and auditory-visual IDS lends support to the behavioural finding that increased attention deployed to the speaker's mouth in suboptimal listening conditions is not associated with better speech recognition (Lansing & McConkie, 2003). Although adults may direct their gaze to the speaker's mouth when confronted with challenging listening conditions as part of an information-seeking strategy, whether this strategy actually facilitates speech perception remains moot.

Results from four-year-olds were unexpected. Cortical tracking accuracy of 4-year-olds did not differ between AV and AO conditions nor was there an auditory-visual speech benefit. It is possible that these non-significant results stem mainly from 4-year-olds' attention, or rather their lack thereof. However, if 4-year-olds' attention is the main contributing factor of the differences observed, then one would expect 4-year-olds' cortical tracking (as indexed by prediction accuracy) to be strongest in the AV condition, and this was not the case. It can be argued that attention to the screen may not be a precise measure of general attention; however, the absence of any derived benefit from the addition of visual speech information for young children fits with the inconsistent findings of an AV>AO effect in the existing literature (e.g., Jerger et al. (2017a) found an AV>AO effect in 4- to 5-year-olds but Maidment et al. (2015) did not). To add on, 4-year-olds' attention during AV trials was greater than that of 5-month-olds and similar to that of adults and yet they did not show any auditory-visual benefit. It is possible that 4-year-olds processed the speech quite effectively based on auditory cues alone especially since the IDS stimuli used in this study were constructed to cater to infants, such that the addition of visual speech information did not augment their cortical tracking. The next unexpected result was the lack of significant relationships between 4-year-olds' cortical tracking accuracy and their looking behaviour. The greater movement and fidgeting in 4-year-olds than in infants or adults resulted in greater loss of eye-tracking data from 4-year-olds hence the smaller sample size ($n = 8$) for the analysis of their gaze data. With this in mind, it is difficult to draw any firm conclusion and any interpretation of the results from 4-year-olds must be made with caution. These limitations warrant further investigation.

4.1. Limitations and future directions

Results from 4-year-olds call for more investigations to be conducted. While inconsistent findings of the auditory-visual speech benefit in behavioural studies with 4- to 5-year-olds (e.g., Jerger et al., 2017b; Maidment et al., 2015) suggest that the absence of any facilitation by visual speech cues observed here in 4-year-olds should not come as a sur-

prise, that the present findings stem partly from their lack of engagement cannot be entirely ruled out. The IDS stimuli used in the present study were short and brief to accommodate for the short attentional spans of infants. To investigate the auditory-visual speech benefit in cortical tracking, it was necessary to repeat the stimuli in three conditions (AV, AO, and VO). So, for the 4-year-olds, even though the IDS stimuli may have been relatively easy for them to process, the experiment was relatively lengthy (~25 min in total), and so apparently failed to maintain their engagement. This was reflected by the restlessness most 4-year-old participants exhibited midway through the experimental session. If 4-year-olds were not engaged, then they may have been less motivated to listen to the speaker and this might have consequently affected their processing of the stimulus and as a result, the degree to which they integrated the auditory and visual information (Pichora-Fuller et al., 2016).

To address this issue, a viable modification of the current paradigm could be to use fewer but longer duration stimuli. For example, three 2-min videos of a speaker reciting children stories could be presented once per condition (a total of 18 min)—as opposed to the thirty 8–15-s short video clips used in this study. This was not done in the current study because, in addition to catering to infants' short attentional spans, there were concerns regarding whether the amount of EEG data recorded from each infant participant would be sufficient for the optimal implementation of the TRF approach, especially since adult studies had larger and longer datasets (e.g., Crosse et al. (2016) used 15×60-s passages per condition). However, these concerns are allayed by a recent demonstration that the TRF approach can be used effectively even with 7-month-olds' EEG data in a single 4-min cartoon video (Jessen et al., 2019), indicating that such a modification is feasible and could be applied in future studies. Until attention can be confidently ruled out as a confound, results from 4-year-olds must be interpreted with caution.

Second, it is interesting that the pattern of cortical tracking accuracy between conditions was similar for 5-month-olds and adults, yet cortical tracking accuracy was greater for 5-month-olds than for adults across all conditions. One possible explanation is the type of speech used in this study—infant-directed speech—which may be more familiar to infants than adults. To investigate this, infants' and adults' cortical tracking accuracy of adult-directed speech (ADS) could be analysed and compared. If speech type accounts for greater cortical tracking accuracy observed here, then cortical tracking accuracy of ADS is expected to be more accurate for adults than for infants. Another possible reason for increased cortical tracking accuracy in the 5-month-olds is that their auditory cortical responses to the speech envelope, as indexed by their TRFs, were on average larger than that of the adults (see GFP in Fig. 2). This phenomenon, which has been observed in numerous studies comparing auditory ERPs in children and adults, could be driven by disparate weightings of neural generators (e.g., P1, N1, P2 and N2) over the course of development, resulting in less phase cancellation and greater cortical response signal-to-noise ratio (Bishop et al., 2007).

Finally, while the positive correlation between 5-month-olds' prediction accuracy of VO TRFs and relative attention to the speaker's mouth is not robust and must be interpreted with caution, previous findings that articulatory information modulates infants' speech perception (e.g., Majorano et al., 2014) hint that the correlation observed here may not be insignificant. Infants' auditory-visual speech perception have been found to be influenced by the speaker's lip movements (Yeung and Werker, 2013) and infants' own articulatory movements (Bruderer et al., 2015). Furthermore, infants' relative attention to the speaker's mouth correlates with their neural responses in the left inferior frontal brain regions during auditory-visual speech processing (Altvater-Mackensen and Grossman, 2016). These findings point toward the inter-connectedness between looking behaviour, articulatory information, and auditory-visual speech processing. It is a preliminary yet tantalising possibility that the observed correlation in 5-month-old infants reflects their articulatory knowledge as a function of their looking behaviour especially since lip movements are highly correlated with speech (e.g., Chandrasekaran et al., 2009). Investigations on the relationship between infants' articulatory knowledge and their speech processing capacities are increasing, but there is still much more to be done. For future studies, it would be interesting to include a test of infants' recognition of target words in addition to examining infants' neural responses and looking behaviour to silent speech stimuli. This would allow a more direct investigation of the relationship between infant looking behaviour, articulatory knowledge, and their neural responses to speech.

5. Conclusion

To date, investigations of gaze behaviour and auditory-visual speech perception have largely been kept separate (cf. Rennig et al., 2020), even though studies of the McGurk effect demonstrate that individual differences in looking behaviour to the talker's facial regions influence perception (e.g., Gurler et al., 2015). This study sought to bridge this gap by simultaneously recording EEG and gaze behaviour as 5-month-old, 4-year-old and adult participants watched unimodal (AO and VO) and multimodal (AV) presentations of a speaker talking. Infants and adults, but not 4-year-olds, showed an auditory-visual speech benefit in cortical tracking of the speech envelope. Additionally, the influence of gaze behaviour is evident in infants' and adults' cortical tracking of silent speech. These findings have implications for populations who have greater reliance on visual cues (e.g., individuals with hearing loss, individuals learning a second language). While there is still much to learn, this study is an important first step toward teasing apart the interactions between a listener's looking behaviour and subsequent speech perception.

Data accessibility

Upon acceptance for publication, data will be uploaded to a publicly available repository.

Credit authorship contribution statement

S.H. Jessica Tan: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Marina Kalashnikova:** Conceptualization, Writing – original draft, Writing – review & editing. **Giovanni M. Di Liberto:** Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Michael J. Crosse:** Methodology, Formal analysis, Writing – original draft. **Denis Burnham:** Conceptualization, Supervision, Writing – original draft.

References

Altvater-Mackensen, N., Grossmann, T., 2016. The role of left inferior frontal cortex during audiovisual speech perception in infants. Neuroimage 133, 14–20. doi:10.1016/j.neuroimage.2016.02.061.

Altvater-Mackensen, N., Grossmann, T., 2018. Modality-independent recruitment of inferior frontal cortex during speech processing in human infants. Dev. Cognit. Neurosci. 34, 130–138. doi:10.1016/j.dcn.2018.10.002.

Birulés, J., Bosch, L., Pons, F., Lewkowicz, D.J., 2020. Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. Lang. Cognit. Neurosci. 92 (2), 1–12. doi:10.1080/23273798.2020.1762905.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., 2009. Hearing faces: how the infant brain matches the face it sees with the speech it hears. J. Cogn. Neurosci. 21 (5), 905–921. doi:10.1162/jocn.2009.21076.

Bruderer, A.G., Danielson, D.K., Kandhadai, P., Werker, J.F., 2015. Sensorimotor influences on speech perception in infancy. Proc. Natl. Acad. Sci. 112 (44), 13531–13536. doi:10.1073/pnas.1508631112.

Buchan, J.N., Paré, M., Munhall, K.G., 2008. The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. Brain Res. 1242, 162–171. doi:10.1016/j.brainres.2008.06.083.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol. 5 (7), e1000436. doi:10.1371/journal.pcbi.1000436.

Chang, C.Y., Hsu, S.H., Pion-Tonachini, L., Jung, T.P., 2019. Evaluation of Artifact Subspace Reconstruction for automatic artifact components removal in multi-channel EEG Recordings. IEEE Trans. Biomed. Eng. 67 (4), 1114–1121. doi:10.1109/TBME.2019.2930186.

Choi, J., Broersma, M., Cutler, A., 2017a. Early phonology revealed by international adoptees' birth language retention. Proc. Natl. Acad. Sci. 114 (28), 7307–7312. doi:10.1073/pnas.1706405114.
Choi, J., Cutler, A., Broersma, M., 2017b. Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory. R. Soc. Open Sci. 4 (1). doi:10.1098/rsos.160660, 160660–14.
Chong, S., Werker, J.F., Russell, J.A., 2003. Three facial expressions mothers direct to their infants. Infant Child Dev. 12 (3), 211–232. doi:10.1002/icd.286.
Corrigall, K.A., Trainor, L.J., 2014. Enculturation to musical pitch structure in young children: evidence from behavioral and electrophysiological methods. Dev. Sci. 17 (1), 142–158. doi:10.1111/desc.12100.
Crosse, M.J., Butler, J.S., Lalor, E.C., 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. J. Neurosci. 35 (42), 14195–14204. doi:10.1523/JNEUROSCI.1829-15.2015.
Crosse, M.J., Di Liberto, G.M., Lalor, E.C., 2016a. Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neurosci. 36 (38), 9888–9895. doi:10.1523/JNEUROSCI.1396-16.2016.
Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C., 2016b. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front. Hum. Neurosci. 10, 604. doi:10.3389/fnhum.2016.00604.
Crosse, M.J., Zuk, N.J., Di Liberto, G.M., Nidiffer, A., Molholm, S., & Lalor, E. (2021, May 11). Linear modeling of neurophysiological responses to naturalistic stimuli: methodological considerations for applied research. 10.31234/osf.io/jbz2w
deBoer, T., Scott, L.S., Nelson, C.A., 2007. Methods for acquiring and analyzing infant event-related potentials. In: de Haan, M. (Ed.), Studies in Developmental psychology. Infant EEG and Event-Related Potentials. Psychology Press, pp. 5–37.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21. doi:10.1016/j.jneumeth.2003.10.009.
Di Liberto, G.M., Lalor, E.C., 2017. Indexing cortical entrainment to natural speech at the phonemic level: methodological considerations for applied research. Hear. Res. 348, 70–77. doi:10.1016/j.heares.2017.02.015.
Di Liberto, G.M., Peter, V., Kalashnikova, M., Goswami, U., Burnham, D., Lalor, E.C., 2018. Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia. Neuroimage 175, 70–79. doi:10.1016/j.neuroimage.2018.03.072.
Ding, N., Simon, J.Z., 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107 (1), 78–89. doi:10.1152/jn.00297.2011.
Ding, N., Simon, J.Z., 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. 33 (13), 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013.
Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19 (1), 158–164. doi:10.1038/nn.4186.
Dohen, M., Loevenbruck, H., & Hill, H. (2006). Visual correlates of prosodic contrastive focus in French: description and inter-speaker variability. Speech Prosody 2006 Conference (pp.221–224). Dresden, Germany: TUD Press.
Fernald, A., Mazzie, C., 1991. Prosody and focus in speech to infants and adults. Dev. Psychol. 27 (2), 209–221. doi:10.1037/0012-1649.27.2.209.
Folland, N.A., Butler, B.E., Payne, J.E., Trainor, L.J., 2015. Cortical representations sensitive to the number of perceived auditory objects emerge between 2 and 4 months of age: electrophysiological evidence. J. Cognit. Neurosci. 27 (5), 1060–1067. doi:10.1162/jocn_a_00764.
Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., Spinelli, E., 2013. Seeing the initial articulatory gestures of a word triggers lexical access. Lang. Cognit. Process 28 (8), 1207–1223. doi:10.1080/01690965.2012.701758.
Foxe, J.J., Molholm, S., Del Bene, V.A., Frey, H.P., Russo, N.N., Blanco, D., …, Ross, L.A., 2015. Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. Cereb. Cortex 25 (2), 298–312.
Friederici, A.D., Wessels, J.M.I., 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. Percept. Psychophys. 54, 287–295. doi:10.3758/bf03205263.
Friston, K., 2010. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11 (2), 127–138. doi:10.1038/nrn2787.
Golumbic, E.Z., Cogan, G.B., Schroeder, C.E., Poeppel, D., 2013. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party. J. Neurosci. 33 (4), 1417–1426. doi:10.1523/jneurosci.3675-12.2013.
Grant, K.W., Bernstein, J.G.W., 2019. Toward a model of auditory-visual speech intelligibility. In: Lee, A.K.C., Wallace, M.T., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.), Multisensory Processes. Springer International Publishing.
Green, J.R., Nip, I.S.B., Wilson, E.M., Mefferd, A.S., Yunusova, Y., 2010. Lip movement exaggerations during infant-directed speech. J. Speech Lang. Hear. Res. 53 (6), 1529–1542. doi:10.1044/1092-4388(2010/09-0005).
Gurler, D., Doyle, N., Walker, E., Magnotti, J., Beauchamp, M., 2015. A link between individual differences in multisensory speech perception and eye movements. Atten. Percept. Psychophys. 77 (4), 1333–1341. doi:10.3758/s13414-014-0821-1.
Henry, M.J., Obleser, J., 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc. Natl. Acad. Sci. 109 (49), 20095–20100. doi:10.1073/pnas.1213390109.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87 (C), 96–110. doi:10.1016/j.neuroimage.2013.10.067.
Hollich, G., Newman, R.S., Jusczyk, P.W., 2005. Infants' use of synchronized visual information to separate streams of speech. Child Dev. 76 (3), 598–613. doi:10.1111/j.1467-8624.2005.00866.x.
Hyde, D.C., Jones, B.L., Flom, R., 2011. Neural signatures of face–voice synchrony in 5-month-old human infants. Dev. Psychobiol. 53 (4), 359–370. doi:10.1002/dev.20525.
Imafuku, M., Myowa, M., 2016. Developmental change in sensitivity to audiovisual speech congruency and its relation to language in infants. Psychologia 59, 163–172. doi:10.2117/psysoc.2016.163.
Imafuku, M., Kanakogi, Y., Butler, D., 2019. Demystifying infant vocal imitation: the roles of mouth looking and speaker's gaze. Dev. Sci. 55 (6), e12825. doi:10.1111/desc.12825.
Jerger, S., Damian, M.F., McAlpine, R.P., Abdi, H., 2017a. Visual speech alters the discrimination and identification of non-intact auditory speech in children with hearing loss. Int. J. Pediatr. Otorhinolaryngol. 94, 127–137. doi:10.1016/j.ijporl.2017.01.009.
Jerger, S., Damian, M.F., McAlpine, R.P., Abdi, H., 2017b. Visual speech fills in both discrimination and identification of non-intact auditory speech in children. J. Child Lang. 22, 1–23. doi:10.1017/s0305000917000265.
Jessen, S., Fiedler, L., Münte, T.F., Obleser, J., 2019. Quantifying the individual auditory and visual brain response in 7- month-old infants watching a brief cartoon movie. Neuroimage 202, 116060. doi:10.1016/j.neuroimage.2019.116060.
Jusczyk, P.W., Luce, P.A., Charles-Luce, J., 1994. Infants' sensitivity to phonotactic patterns in the native language. J. Mem. Lang. 33, 630–645. doi:10.1006/jmla.1994.1030.
Kaganovich, N., Schumaker, J., 2014. Audiovisual integration for speech during mid-childhood: electrophysiological evidence. Brain Lang. 139, 36–48. doi:10.1016/j.bandl.2014.09.011.
Kalashnikova, M., Peter, V., Liberto, G.M., Lalor, E.C., Burnham, D., 2018. Infant-directed speech facilitates seven-month-old infants' cortical tracking of speech. Sci. Rep. 8 (1), 1–8. doi:10.1038/s41598-018-32150-6.
Kitamura, C., Thanavishuth, C., Burnham, D., Luksaneeyanawin, S., 2001. Universality and specificity in infant-directed speech: pitch modifications as a function of infant age and sex in a tonal and non-tonal language. Infant Behav. Dev. 24 (4), 372–392. doi:10.1016/S0163-6383(02)00086-3.
Knowland, V., Mercure, E., Karmiloff-Smith, A., Dick, F., Thomas, M.S.C., 2014. Audiovisual speech perception: a developmental ERP investigation. Dev. Sci. 17 (1), 110–124. doi:10.1111/desc.12098.
Kothe, C.A.E., & Jung, T.P. (2014). U.S. Patent Application No. 14/895,440.
Kushnerenko, E., Teinonen, T., Volein, A., Csibra, G., 2008. Electrophysiological evidence of illusory audiovisual speech percept in human infants. Proc. Natl. Acad. Sci. 105 (32), 11442–11445. doi:10.1073/pnas.0804275105.
Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E.L., Murphy, E., Moore, D.G., 2013. Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. Eur. J. Neurosci. 38 (9), 3363–3369. doi:10.1111/ejn.12317.
Lalonde, K., Holt, R.F., 2015. Preschoolers benefit from visually salient speech cues. J. Speech Lang. Hear. Res. 58 (1), 135–150. doi:10.1044/2014_JSLHR-H-13-0343.
Lalonde, K., Holt, R.F., 2016. Audiovisual speech perception development at varying levels of perceptual processing. J. Acoust. Soc. Am. 139 (4), 1713–1723. doi:10.1121/1.4945590.
Lansing, C.R., McConkie, G.W., 1999. Attention to facial regions in segmental and prosodic visual speech perception tasks. J. Speech Lang. Hear. Res. 42 (3), 526–539. doi:10.1044/jslhr.4203.526.
Leong, V., Byrne, E., Clackson, K., Harte, N., Lam, S., de Barbaro, K., & Wass, S. (2017). Infants' neural oscillatory processing of theta-rate speech patterns exceeds adults'. bioRxiv, 108852. 10.1101/108852
Lewkowicz, D.J., 2003. Learning and discrimination of audiovisual events in human infants: the hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. Dev. Psychol. 39 (5), 795–804. doi:10.1037/0012-1649.39.5.795.
Lewkowicz, D.J., Hansen-Tift, A.M., 2012. Infants deploy selective attention to the mouth of a talking face when learning speech. Proc. Natl. Acad. Sci. 109 (5), 1431–1436. doi:10.1073/pnas.1114783109.
LoBue, V., Buss, K.A., Taber-Thomas, B.C., Pérez-Edgar, K., 2016. Developmental differences in infants' attention to social and nonsocial threats. Infancy 22 (3), 403–415. doi:10.1111/infa.12167.
Majorano, M., Vihman, M.M., DePaolis, R.A., 2014. The relationship between infants' production experience and their processing of speech. Lang. Learn. Dev. 10 (2), 179–204. doi:10.1080/15475441.2013.829740.
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. Res. 58 (1), 61–68. doi:10.1044/2014_JSLHR-S-14-0044.
Mehoudar, E., Arizpe, J., Baker, C.I., Yovel, G., 2014. Faces in the eye of the beholder: unique and stable eye scanning patterns of individual observers. J. Vis. 14 (7), 6. doi:10.1167/14.7.6.
Mercier, M.R., Molholm, S., Fiebelkorn, I.C., Butler, J.S., Schwartz, T.H., Foxe, J.J., 2015. Neuro-oscillatory phase alignment drives speeded multisensory response times: an electro-corticographic investigation. J. Neurosci. 35 (22), 8546–8557. doi:10.1523/JNEUROSCI.4527-14.2015.
Moradi, S., Lidestam, B., Rönnberg, J., 2013. Gated audiovisual speech identification in silence vs. noise: effects on time and accuracy. Front. Psychol. 4, 359. doi:10.3389/fpsyg.2013.00359.
Murray, M.M., Brunet, D., Michel, C.M., 2008. Topographic ERP Analyses: a step-by-step tutorial review. Brain Topogr. 20 (4), 249–264. doi:10.1007/s10548-008-0054-5.
Navarra, J., Soto-Faraco, S., 2007. Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol. Res. 71 (1), 4–12. doi:10.1007/s00426-005-0031-5.

Nazzi, T., Jusczyk, P.W., Johnson, E.K., 2000. Language discrimination by english-learning 5-month-olds: effects of rhythm and familiarity. J. Mem. Lang. 43 (1), 1–19. doi:10.1006/jmla.2000.2698.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. 2011 (1). doi:10.1155/2011/156869, 156869–9.

O'Sullivan, A.E., Lim, C.Y., Lalor, E.C., 2019. Look at me when I'm talking to you: selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. Eur. J. Neurosci. 50 (8), 3282–3295. doi:10.1111/ejn.14425.

Peter, V., Kalashnikova, M., Santos, A., Burnham, D., 2016. Mature neural responses to infant-directed speech but not adult-directed speech in pre-verbal infants. Sci. Rep. 6 (1), 34273. doi:10.1038/srep34273.

Peterson, M.F., Eckstein, M.P., 2012. Looking just below the eyes is optimal across face recognition tasks. Proc. Natl. Acad. Sci. 109 (48), E3314–E3323. doi:10.1073/pnas.1214269109.

Pichora-Fuller, M.K., Kramer, S.E., Eckert, M.A., Edwards, B., Hornsby, B.W.Y., Humes, L.E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C.L., Naylor, G., Phillips, N.A., Richter, M., Rudner, M., Sommers, M.S., Tremblay, K.L., Wingfield, A., 2016. Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). Ear Hear. 37, 5–27. doi:10.1097/AUD.0000000000000312.

Polka, L., Werker, J.F., 1994. Developmental changes in perception of nonnative vowel contrasts. J. Exp. Psychol. Hum. Percept. Perform. 20 (2), 421–435. doi:10.1037/0096-1523.20.2.421.

Pons, F., Bosch, L., Lewkowicz, D.J., 2015. Bilingualism modulates infants' selective attention to the mouth of a talking face. Psychol. Sci. 26 (4), 490–498. doi:10.1177/0956797614568320.

Pons, F., Bosch, L., Lewkowicz, D.J., 2019. Twelve-month-old infants' attention to the eyes of a talking face is associated with communication and social skills. Infant Behav. Dev. 54, 80–84. doi:10.1016/j.infbeh.2018.12.003.

Rennig, J., Wegner-Clemens, K., Beauchamp, M.S., 2020. Face viewing behavior predicts multisensory gain during speech perception. Psychon. Bull. Rev. 27 (1), 70–77. doi:10.1101/331306.

Reynolds, G.D., Bahrick, L.E., Lickliter, R., Guy, M.W., 2013. Neural correlates of intersensory processing in 5-month-old infants. Dev. Psychobiol. 56 (3), 355–372. doi:10.1002/dev.21104.

Richoz, A.-.R., Quinn, P.C., Hillairet de Boisferon, A., Berger, C., Loevenbruck, H., Lewkowicz, D.J., Kang, L., Dole, M., Caldara, R., Pascalis, O., 2017. Audio-visual perception of gender by infants emerges earlier for adult-directed speech. PLoS One 12 (1), e0169325. doi:10.1371/journal.pone.0169325.

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. Eur. J. Neurosci. 33 (12), 2329–2337. doi:10.1111/j.1460-9568.2011.07685.x.

Ru, P., 2001. Multiscale Multirate Spectro-Temporal Auditory Model. [Unpublished Doctoral Dissertation]. University of Maryland College Park.

Rudmann, D.S., McCarley, J.S., Kramer, A.F., 2003. Bimodal displays improve speech comprehension in environments with multiple speakers. Hum. Factors 45 (2), 329–336. doi:10.1518/hfes.45.2.329.27237.

Schwartz, J.L., Berthommier, F., Savariaux, C., 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93 (2), B69–B78. doi:10.1016/j.cognition.2004.01.006.

Stacey, J.E., Howard, C.J., Mitra, S., Stacey, P.C., 2020. Audio-visual integration in noise: influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. Attent. Percept. Psychophys. 82 (7), 3544–3557.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26 (2), 212–215. doi:10.1121/1.1907309.

Tenenbaum, E.J., Shah, R.J., Sobel, D.M., Malle, B.F., Morgan, J.L., 2013. Increased focus on the mouth among infants in the first year of life: a longitudinal eye-tracking study. Infancy 18 (4), 534–553. doi:10.1111/j.1532-7078.2012.00135.x.

Tsang, T., Atagi, N., Johnson, S.P., 2018. Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. J. Exp. Child Psychol. 169, 93–109. doi:10.1016/j.jecp.2018.01.002.

van Wassenhove, V., Grant, K.W., Poeppel, D., Halle, M., 2005. Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci. 102 (4), 1181–1186. doi:10.1073/pnas.0408949102.

Vander Ghinst, M., Bourguignon, M., Niesen, M., Wens, V., Hassid, S., Choufani, G., Jousmaki, V., Hari, R., Goldman, S., De Tiege, X., 2019. Cortical tracking of speech-in-noise develops from childhood to adulthood. J. Neurosci. 39 (15), 2938–2950. doi:10.1523/JNEUROSCI.1732-18.2019.

Vatikiotis-Bateson, E., Eigsti, I.M., Yano, S., Munhall, K.G., 1998. Eye movement of perceivers during audiovisual speech perception. Percept. Psychophys. 60 (6), 926–940. doi:10.3758/bf03211929.

Yehia, H., Rubin, P., Vatikiotis-Bateson, E., 1998. Quantitative association of vocal-tract and facial behavior. Speech Commun. 26 (1–2), 23–43. doi:10.1016/s0167-6393(98)00048-x.

Yeung, H.H., Werker, J.F., 2013. Lip movements affect infants' audiovisual speech perception. Psychol. Sci. 24 (5), 603–612. doi:10.1177/0956797612458802.

Young, G.S., Merin, N., Rogers, S.J., Ozonoff, S., 2009. Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. Dev. Sci. 12 (5), 798–814. doi:10.1111/j.1467-7687.2009.00833.x.

Further reading

Baart, M., Samuel, A.G., 2015. Early processing of auditory lexical predictions revealed by ERPs. Neurosci. Lett. 585, 98–102. doi:10.1016/j.neulet.2014.11.044.

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. doi:10.1016/j.cognition.2013.09.006.

Bernstein, L.E., Liebenthal, E., 2014. Neural pathways for visual speech perception. Front. Neurosci. 8, 386. doi:10.3389/fnins.2014.00386.

MathWorks, 2019. MATLAB: R2019a. Mathworks, Inc, Natick.

Taylor, G., Herbert, J.S., 2012. Eye tracking infants: investigating the role of attention during learning on recognition memory. Scand. J. Psychol. 54 (1), 14–19. doi:10.1111/sjop.12002.

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. doi:10.1016/j.cognition.2008.05.009.

Wunderlich, J.L., Cone-Wesson, B.K., Shepherd, R., 2006. Maturation of the cortical auditory evoked potential in infants and young children. Hear. Res. 212 (1–2), 185–202. doi:10.1016/j.heares.2005.11.010.